Ярославский Л.П.

ЦИФРОВАЯ ОБРАБОТКА СИГНАЛОВ В ОПТИКЕ И ГОЛОГРАФИИ:
ВВЕДЕНИЕ В ЦИФРОВУЮ ОПТИКУ

СОДЕРЖАНИЕ

Глава 1

ЭЛЕМЕНТЫ ТЕОРИИ СИГНАЛОВ

1.1. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ОПТИЧЕСКИХ СИГНАЛОВ

Для аналитического описания сигналов и процессов их преобразований используют математические модели. Прежде всего, сигналы рассматривают как функции, заданные в физических координатах. В этом смысле говорят об одномерных сигналах (например, зависящих от времени), двумерных, заданных на плоскости (например, изображениях), трехмерных (характеризующих, например, трехмерные пространственные объекты). Обычно в теории сигналов в качестве математических моделей сигналов используются скалярные функции. Но в оптике и голографии приходится прибегать к более сложным моделям – комплексным и векторным функциям. Например, для описания электромагнитного поля как сигнала удобно использовать комплексные функции, для описания цветных изображений – трехкомпонентные векторные функции, для описания данных многоспектральной съемки – 4–6-компонентные векторные функции.

Важными общими характеристиками сигналов как математических функций являются множества значений, которые могут принимать они сами и их аргументы. С этой точки зрения целесообразно различать финитные и инфинитные, ограниченные и неограниченные, непрерывные, дискретные, квантованные и цифровые сигналы.

Финитными называются сигналы, область определения которых .ограничена. Например, финитным является сигнал, характеризующий кадр фотоснимка, значение электромагнитного поля в раскрыве антенны и т.п. *Инфинитные* сигналы имеют неограниченную область определения. Финитный сигнал можно превратить в инфинитный, если доопределить его значения за пределами заданного интервала. Однако любое доопределение должно выполняться так, чтобы при обработке сигнала не изменялись его свойства внутри области определения.

Если ограничена область значений сигнала, то сигнал называется *ограниченным.* Так, например, ограниченным является сигнал, характеризующий степень почернения фотонегатива или распределение яркости свечения экрана телевизионного монитора. Фактически все встречающиеся в природе сигналы являются ограниченными. *Неограниченный* сигнал–идеализация, принимаемая в тех случаях, когда ограничение области значений сигнала несущественно для данной задачи или просто неизвестно.

Непрерывными называются сигналы, область определения и область значений которых непрерывны, т.е. для каждой точки области определения и области значения можно найти точку, удаленную от нее на бесконечно малое расстояние.

Если область определения сигнала состоит из отдельных точек, сигнал называют *дискретным.* Таким образом, дискретный сигнал– это последовательность чисел, называемых элементами дискретного сигнала.

Если область значений сигнала состоит из отдельных точек, т.е. сигнал может принимать только определенные, «квантованные», значения, сигнал называют *квантованным.* Квантованный дискретный сигнал называется *цифровым.*

Непрерывные сигналы называют также аналоговыми, подчеркивая, что они являются как бы аналогами порождающих их природных объектов, которые обычно непрерывны (если отвлечься от квантовых явлений). Дискретные, квантованные и цифровые сигналы – это, как правило, искусственные и в определенном смысле абстрактные объекты.

Мы будем обозначать непрерывные сигналы первыми буквами латинского алфавита **a**, **b**, **c**, а их аргументы буквой x, подразумевая, что для многомерных непрерывных сигналов x – векторная переменная с компонентами $x_1, \ldots, x_n$, где n – размерность сигнала. Дискретные и цифровые сигналы будем обозначать векторами, компонентами которых являются элементы сигнала, например $\mathbf{a}=\{a_0, a_1, \ldots, a_k, \ldots, a_{N-1}\}$, где N – количество элементов сигнала.

Любой оптический или электрический сигнал – это непрерывный сигнал. Примером цифрового сигнала может служить последовательность чисел, записанных в памяти цифровой вычислительной машины. Если бы эти числа можно было записывать с неограниченным количеством цифр, т.е. с бесконечно высокой точностью, то это был бы дискретный сигнал.

Непрерывным аналогом дискретного сигнала является сигнал, который в отдельных точках области определения может принимать произвольные значения, а в остальных точках равен какой-нибудь константе (например, нулю).

С точки зрения математического описания сигналов различают также детерминированное и вероятностное описания. При детерминированном описании сигналы рассматриваются индивидуально, независимо друг от друга, и считается, что значение сигнала может быть задано в каждой точке, где он определен. Однако иногда индивидуальное рассмотрение характеристик физических объектов невозможно, а можно измерить и учесть только некоторое число «макропараметров», характеризующих объекты в среднем. В этих случаях используется вероятностное описание, т.е. сигналы рассматриваются как выборочные функции, или реализации из некоторого ансамбля сигналов, и строится математическое описание не каждого отдельного сигнала, а ансамбля в целом.

Давая математическое описание сигналов, удобно рассматривать их как точки или векторы в некотором функциональном пространстве (пространстве сигналов), а преобразования сигналов – как отображения в этом пространстве. При этом свойства сигналов трактуются как свойства пространства. Слово «пространство» используется, чтобы придать понятию множества сигналов геометрический смысл и, тем самым, наглядность.

Для того чтобы математически описать различия между сигналами, вводится понятие метрики пространства, т.е. способа, в соответствии с которым каждой паре точек пространства, скажем a_1 и a_2, может быть поставлено в соответствие некоторое вещественное неотрицательное число $d(a_1, a_2)$, имеющее смысл расстояния между ними. Обычно этот способ удовлетворяет следующим правилам:

$d(a_1, a_2)=0$, если $a_1 = a_2$

$d(a_1, a_2)= d(a_2, a_1)$

$d(a_1, a_2) <= d(a_1, a_2)+ d(a_2, a_3)$

Смысл первых двух условий очевиден. Смысл введения третьего условия, *которое* называется «правилом треугольника», в том, что оно является формальным выражением следующего естественного требования к метрике: если дзе точки близки к третьей, то они должны быть близки и между собой.

Примеры наиболее часто используемых в теории сигналов метрик приведены в табл. 1.1. Для компактности и удобства для метрик дискретных сигналов в таблице использованы матричные обозначения сигналов как вектор-столбцов, индекс «т» означает транспонирование.

Таблица 1.1. *Примеры метрик*

Тип сигнала	Обозначение	Определение
Дискретный	l_N	$\sum_{k=0}^{N-1} \lvert a_{1k} - a_{2k} \rvert$
	l_N^2	$\sqrt{\sum_{k=0}^{N-1} \lvert a_{1k} - a_{2k} \rvert} = \sqrt{(a_1 - a_2)^T (a_1 - a_2)}$
	m_N	$\max_k \lvert a_{1k} - a_{2k} \rvert$
	mh_N	$\sqrt{(a_1 - a_2)^T (a_1 - a_2)}; \sum$ - *весовая матрица*
Непрерывный	L_x	$\int_X \lvert a_1(x) - a_2(x) \rvert dx$
	L_x^2	$\sqrt{\int_X \lvert a_1(x) - a_2(x) \rvert^2 dx}$
	M_x	$\sup \lvert a_1(x) - a_2(x) \rvert$

Метрика l_N^2 и ее непрерывный аналог – метрика L_x^2 а также их обобщение на случай $N \to \infty$ и $X \to \infty$, называются эвклидовыми, так как l_3^2 совпадает с эвклидовой метрикой реального физического пространства.

Поскольку в теории сигналов понятие расстояния используется для трактовки отличия одного сигнала от другого или ошибки представления одного сигнала другим, для характеристики пространства сигналов должна выбираться такая метрика, которая наиболее полно может описать это отличие одним числом.

Пусть, например, отличия одного сигнала от другого возникают в результате действия на сигналы аддитивного некоррелированного гауссовского шума. Рассмотрим для простоты случай различения двух дискретных сигналов $\{a_{1k}\}$ и $\{a_{2k}\}$, таких, что

$$a_{2k} = a_{1k} + n_k, k = 0,1,..., N-1 \qquad (1.1)$$

где n_k– случайные величины с нормальной плотностью вероятностей

$$p(n) = (1/\sqrt{2\pi\sigma})\exp(-n^2/2\sigma^2)$$

и дисперсией σ^2

Очевидно, все различия между сигналами a_1 и a_2 заключены в сигнале $n = \{n_k\}$ а он может быть полностью статистически описан многомерной плотностью вероятностей

$$p(n_0, n_1,..., n_{N-1}) = (2\pi\sigma^2)^{-N/2}\exp\left[-(1/2\sigma^2)\sum_{k=0}^{N-1}(a_{2k} - a_{1k})^2\right] \quad ,$$

которая в свою очередь полностью определяется величиной

$$d(a_1, a_2) = \sqrt{\sum_{k=0}^{N-1} | a_{2k} - a_{1k} |^2}$$

– эвклидовым расстоянием между a_1 и a_2. Так порождается *эвклидова метрика.*

Эвклидова метрика очень популярна в теории сигналов по двум причинам. Во-первых, она удобна в расчетах и имеет определенный физический смысл: это мера энергии разности двух сигналов, измерение которой легко воплотить в физическом приборе. Во-вторых, эта метрика в точности адекватна задачам, где отличия между сигналами порождаются суммарным действием большого числа помех или ошибок измерения.

Эвклидову метрику часто называют также среднеквадратической, ибо она дает квадрат разности сигналов, усредненный по области их определения. В этом смысле ее обобщением является взвешенная среднеквадратическая метрика, определяемая для дискретного случая как

$$d(a_1, a_2) = \sqrt{\sum_{k=0}^{N-1} \omega_k (a_{2k} - a_{1k})^2}$$

где $\{\omega_k\}$–набор весовых констант. Такая метрика потребовалась бы, например, если бы в (1.1) мы предположили, что $\{n_k\}$ имеют разные значения дисперсии σ_k^2. Если бы, далее, мы предположили, что величины $\{n_k\}$ не были независимыми, мы получили бы метрику mh_N, в которой Σ–матрица, обратная ковариационной: матрице величин $\{n_k\}$. Такая метрика называется *махаланобисовой.*

Обычно для сигналов как физических объектов выполняется принцип суперпозиции. Математически он формулируется как свойство линейности пространства сигналов, т.е. считается, что пространство сигналов обладает следующими свойствами.

1. Для любых двух его элементов $\mathbf{a}_1$ и $\mathbf{a}_2$ однозначно определен принадлежащий ему третий элемент $\mathbf{a}_3$, называемый их суммой и обозначаемый $\mathbf{a}_1 + \mathbf{a}_2$, причем операция суммирования подчиняется законам коммутативности:

   $$\mathbf{a}_1 + \mathbf{a}_2 = \mathbf{a}_2 + \mathbf{a}_1$$

   и ассоциативности:

   $$\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}.$$

2. Существует такой элемент $\varnothing$, что $\mathbf{a} + \varnothing = \mathbf{a}$ для всех элементов пространства.
3. Каждому элементу a пространства можно поставить в соответствие противоположный ему элемент $-\mathbf{a}$, такой, что $\mathbf{a} + (-\mathbf{a}) = \varnothing$
4. Для любого числа α и любого элемента пространства a определен принадлежащий этому пространству элемент $\alpha\mathbf{a}$, причем так, что

   $\alpha_1(\alpha_2\mathbf{a}) = (\alpha_1\alpha_2)\mathbf{a}$; $\quad 1\mathbf{a} = \mathbf{a}$;

   $(\alpha_1 + \alpha_2)\mathbf{a} = \alpha_1\mathbf{a} + \alpha_2\mathbf{a}$; $\quad \alpha_1(\mathbf{a}_1 + \mathbf{a}_1) = \alpha\mathbf{a}_1 + \alpha\mathbf{a}_2$.

Элементы линейного пространства будем называть векторами.

Вектор, образованный суммированием нескольких векторов со скалярными коэффициентами, называется линейной комбинацией:

$$a = \sum_{k=0}^{N-1} \alpha_k \varphi_k \qquad (1.2)$$

Множество векторов φ_k называется линейно-независимым, если при любых α_k отличных от нуля, $\sum_{k=0}^{N-1} \alpha_k \varphi_k \neq 0$. Следовательно, линейно-независимое множество таково, что ни один из его векторов не может быть представлен в виде линейной комбинации других.

Пространство A_N составленное из линейных комбинаций N линейно-независимых векторов φ_k, называется N-мерным линейным пространством. Множество линейно-независимых векторов φ_k называется *базисом* этого пространства. Любое множество N линейно-независимых векторов в A_N может служить его базисом.

Каждый вектор в A_N – соответствует единственной линейной комбинации векторов φ_k – единственному множеству скалярных коэффициентов α_k. Набор (упорядоченный) скалярных коэффициентов α_k разложения данного вектора по данному базису является *представлением* вектора по отношению к данному базису.

Обычно операции сложения сигналов и умножения их на скаляры – это обычные операции поточечного (поэлементного) арифметического сложения и умножения значений сигнала, нулевой вектор – это сигнал, все значения которого равны нулю.

Нулевой вектор можно рассматривать как некоторый стандартный элемент пространства сигналов. Отличие данного сигнала a от нулевого является индивидуальной характеристикой сигнала. Математически эта характеристика описывается *нормой вектора* $//a//$ – числом, способ определения которого для каждого вектора удовлетворяет условиям, аналогичным условиям на метрику:

а) $\| a \| \geq 0; \| a \| = 0,$ если только $\mathbf{a} \neq \varnothing$;

б) $\| a_1 + a_2 \| \leq \| a_1 \| + \| a_2 \|$;

в) $\| \alpha a \| = | \alpha | \| a \|$.

Геометрический аналог нормы вектора – длина вектора.

В линейном пространстве, в котором определена норма, естественно и метрику определять через норму:$d(a_1, a_2)$, т.е. $\|a\|=d(a, \varnothing)$.

В соответствии с (1.2), зная набор скалярных величин $\{\alpha_k\}$ и базисных функций $\{\varphi_k\}$, можно найти сигнал **a**, который этому набору скаляров соответствует. Но такое представление имеет смысл только тогда, когда существует способ решить обратную задачу–для каждого вектора **a** найти его представление $\{\alpha_k\}$ по данному базису $\{\boldsymbol{\varphi}_k\}$, Для этого вводится понятие *скалярного произведения двух векторов.* Это число (вообще говоря, комплексное), способ вычисления которого характеризуется следующими свойствами:

а) $(\mathbf{a}_1, \mathbf{a}_2) = (\mathbf{a}_2, \mathbf{a}_1)^*$,

где * означает комплексно-сопряженную величину;

б) $(\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2, \mathbf{a}_3) = \alpha_1(\mathbf{a}_1, \mathbf{a}_3) + \alpha_2(\mathbf{a}_2, \mathbf{a}_3)$;

в) $(\mathbf{a}, \mathbf{a}) \geq 0$; $(\mathbf{a}, \mathbf{a}) = 0$, только если $\mathbf{a} = \varnothing$

Обычно пользуются следующим способом вычисления скалярного произведения:

$$(a_1, a_2) = \int_X a_1(x_1) a_2^*(x) dx .$$

Понятия скалярного произведения и нормы векторов можно связать, определив норму через скалярное произведение:

$$\| a \| = (a, a)^{1/2}$$

Векторы, скалярное произведение которых равно нулю, называются *ортогональными.* Если векторы a_k взаимно ортогональны, то они линейно независимы. Поэтому ортогональные векторы можно использовать как базисы линейных пространств.

Если в пространстве определено скалярное произведение, то, пользуясь им, можно установить простое соотношение между сигналом и его представлением.

Пусть A_N–N-мерное пространство с базисом $\{\varphi_k\}$, k=0, 1, ..., N–1, т.е. состоящее из векторов вида:

$$a = \sum_{k=0}^{N-1} \alpha_k \varphi_k$$

а $\{\psi_k\}$– векторы, которые попарно ортогональны к $\{\varphi_k\}$ и нормированы так, что

$$(\varphi_k, \psi_k) = \delta_{k,l} = \begin{cases} 1, k = l; \\ 0, k \neq l. \end{cases} \qquad (1.3)$$

Символ $\delta_{k,l}$ называется символом (дельта-функцией) Кронекера. Базис $\{\psi_l\}$, удовлетворяющий этому соотношению, называется *взаимным* к $\{\varphi_k\}$. Его можно использовать для вычисления коэффициентов представления $\{\alpha_k\}$:

$$(a, \psi_l) = \sum_{k=0}^{N-1} \alpha_k (\varphi_k, \psi_l) = \sum_{k=0}^{N-1} \alpha_k \delta_{k,l} = \alpha_l \qquad (1.4)$$

Если базис $\{\varphi_k\}$ содержит нормированные попарно ортогональные векторы, т.е. взаимен самому себе:

$$(\varphi_k, \varphi_l) = \delta_{k,l}$$

его называют *ортонормальным.* В этом базисе

$$a = \sum_{k=0}^{N-1} (a, \varphi_k) \varphi_k$$

Зная представление векторов по ортонормальному базису, можно вычислить их норму

$$\| a \|^2 = (a, a) = \sum_{k=0}^{N-1} | \alpha_k |^2$$

и скалярное произведение

$$(a, b) = \left(\sum_{k=0}^{N-1} \alpha_k \varphi_k, \sum_{k=0}^{N-1} \beta_k \varphi_k \right) = \sum_{k=0}^{N-1} \alpha_k \beta_k^*.$$

Представления $\{\alpha_{1k}\}$, $\{\beta_{1k}\}$, $\{\alpha_{2k}\}$, $\{\beta_{2k}\}$ любой пары сигналов a, b по двум ортонормальным базисам $\{\varphi_{1k}\}$ и $\{\varphi_{2k}\}$ связаны между собой соотношением:

$$\sum_{k=0}^{N-1} \alpha_{1k} \beta_{1k}^* = \sum_{l=0}^{N-1} \alpha_{2l} \beta_{2l}^*, \qquad (1.5)$$

называемым *соотношением Парсе валя.*

Представление сигналов как элементов линейного конечномерного пространства удобно потому, что позволяет описать любой сигнал набором некоторых стандартных базисных функций и набором чисел. Выбор базиса определяется удобством нахождения представления сигналов и, конечно, существом задачи, т.е. особенностями сигналов.

Наиболее употребительны два класса базисных функций: сдвиговые базисные функции и мультипликативные.

Сдвиговые базисные функции строятся из одной функции путем сдвига по ее аргументу. Наиболее употребительными сдвиговыми базисными функциями являются импульсные базисные функции и функции отсчетов.

Импульсные базисные функции в одномерном случае определяются как

$$\varphi_k(x) = \mathrm{rect}\,(x - k\Delta x)/\Delta x,$$

где $\mathrm{rect}\,(x) = \begin{cases} 1, & 0 \leqslant x < 1, \\ 0, & \text{в противном случае, } \Delta x \text{ — стандартный сдвиг.} \end{cases}$ (1.6а)

Функции $\varphi_k(x)$ (1.6а) ортогональны на всей оси x. Пространство с этим базисом составляют ступенчатые функции. Взаимный базис для этой системы образуют функции

$$\psi_k(x) = (1/\Delta x)\,\mathrm{rect}\,(x - k\Delta x)/\Delta x. \quad (1.6б)$$

Представлением сигналов по базису (1.6а) являются их средние значения на соответствующих интервалах:

$$a_k = \frac{1}{\Delta x}\int_{k\Delta x}^{(k+1)\Delta x} a(x)\,dx.$$

Функциями отсчетов называют функции, определяемые в одномерном случае как

$$\varphi_k(x) = \mathrm{sinc}\,\pi(x - k\Delta x)/\Delta x = \frac{\sin[\pi(x - k\Delta x)/\Delta x]}{\pi(x - k\Delta x)/\Delta x}. \quad (1.7)$$

Эти функции ортогональны на $(-\infty, \infty)$:

$$\int_{-\infty}^{\infty} \mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x]\,\mathrm{sinc}\,[\pi(x - l\Delta x)/\Delta x]\,dx =$$

$$= \Delta x\,\mathrm{sinc}\,\pi(k - l) = \begin{cases} \Delta x, & k = l; \\ 0, & k \neq l. \end{cases}$$

Отсюда видно, что взаимный базис к ним образуют функции

$$\{(1/\Delta x)\,\mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x]\}$$

Функции отсчетов обычно используются для дискретного представления сигналов по теореме отсчетов (см. § 2.2). Свое название они получили потому, что для сигналов с ограниченным спектром Фурье (см. ниже § 1.3 и § 2.2) коэффициенты {at} представления по этим базисным функциям являются просто значениями, или отсчетами, сигналов при $x = k\Delta x$:

$$a_k = (1/\Delta x)\int_{-\infty}^{\infty} a(x)\,\mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x]\,dx = a(k\Delta x)$$

Мультипликативными называются базисные функции, обладающие тем свойством, что произведение двух функций дает также базисную функцию из той же системы. Это свойство используется для построения системы базисных функций путем многократного перемножения одной и той же функции. Наиболее известными и употребительными семействами мультипликативных функций являются экспоненциальные функции и функции Уолша.

Экспоненциальные базисные функции определяются как:

$$\varphi_k(x) = \exp(i\,2\pi kx/X) \quad (1.8а)$$

Разложение сигналов по этим базисным функциям называется разложением в ряд Фурье.

Взаимный базис образуют функции

$$\psi_k(x) = (1/X)\exp(-i\,2\pi kx/X) \quad (1,8б)$$

Коэффициенты ряда Фурье вычисляются по формуле

$$a_k = (1/X)\int_{-X/2}^{X/2} a(x)\exp(-i\,2\pi kx/X)\,dx$$

Функции Уолша замечательны тем, что принимают всего два значения: 1 и -1. Они порождаются *функциями Радемахера*

$$\mathrm{rad}_k(x) = \mathrm{sign}\,(\sin 2^k\pi x/X) \quad (1.9)$$

Любые две функции Радемахера ортогональны между собой. Но система функций $\{\mathrm{rad}_k(x)\}$ не является полной: на отрезке (0, *X)* существуют и другие функция, ортогональные функциям Радемахера (1.9). Функции Уолша являются расширением системы функций Радемахера до полной системы. Они определяются так:

$$\mathrm{wal}_k(x) = (1/\sqrt{X}) \prod_{m=0}^{\infty} (\mathrm{rad}_{m+1}(x))^{k'_m}, \tag{1.10}$$

где k'_m– m-й разряд кода Грэя номера *k (* Код Грэя образуется из двоичного номера *k* по следующему правилу. $k'_m = k_m \oplus k_{m+1}$, где m – номер двоичного разряда (читается справа налево), k_m–*m-й* двоичный разряд в двоичной записи номера *k:* $k = \sum_{m=1}^{\infty} k_m 2^m$, а $\oplus$ означает сложение по модулю 2 *.).*

Формула (1.10) удобна для понимания природы функций Уолша. Для вычислений их значений удобнее иная форма представления функций Уолша – через значения разрядов $\{\xi_i\}$ двоичного кода нормированного значения аргумента $\xi = x/X$:

$$\mathrm{wal}_k(\xi) = \frac{1}{X} \prod_{m=0}^{\infty} [(-1)^{\xi_{m+1}}]^{k'_m} = \frac{1}{\sqrt{X}} (-1)^{\sum_{m=0}^{\infty} k'_m \xi_{m+1}}. \tag{1.11}$$

где

$$\xi = \sum_{m=0}^{\infty} \xi_{m+1} 2^{-(m+1)}. \tag{1.12}$$

Функции Уолша ортонормальны на отрезке (0, *X).* При перемножении двух функций Уолша происходит сдвиг по индексу или по аргументу, называемый в отличие от арифметического сдвига *диадическим* и определяемый через поразрядное сложение по модулю 2:

$$\mathrm{wal}_k(x)\,\mathrm{wal}_l(x) = \mathrm{wal}_{k \oplus l}(x);$$
$$\mathrm{wal}_k(\xi X)\,\mathrm{wal}_k(\zeta X) = \mathrm{wal}_k((\xi \oplus \zeta) X)$$

Сочетание импульсных функций rect(*x*) и функций Радемахера порождает еще одну интересную с точки зрения цифровой обработки систему ортогональ-' них функций – *функции Хаара.* Одномерные функции Хаара определяются на интервале (0, *X)* следующим образом:

$$\mathrm{har}_k(x) = \frac{2^{m/2}}{\sqrt{X}} \mathrm{rad}_{m+1}(x)\ \mathrm{rect}\left[\frac{x}{X} 2^m - (k) \bmod 2^m\right] \tag{1.13}$$

где $\mathrm{rad}_{m+1}(x)$–функция Радемахера (1.41); *m*– номер самого старшего ненулевого разряда в двоичном представлении *k; (k)mod2*^{m}– величина *k* по модулю 2^m. Функции Хаара ортонормальны на интервале (0, *X)* .

Численное значение функции Хаара в каждой точке можно найти, выразив его, как и в случае функций Уолша, через представление аргумента в двоичном коде (1.12):

$$\mathrm{har}_k(x) = \frac{2^{m/2}}{\sqrt{X}} (-1)^{\xi_{m+1}} \prod_{n=0}^{m-1} \delta(\xi_{m-n}, k_n) =$$
$$= \frac{2^{m/2}}{\sqrt{X}} (-1)^{\xi_{m+1}} \delta([\xi]_m, (k) \bmod 2^m), \tag{1.14}$$

где $\delta(\cdot)$–символ Кронекера; $[\xi]_m$–двоичное число, составленное из *m* старших двоичных разрядов числа ξ.

Двумерные базисные функции обычно получают как произведение одномерных. Эта делается для того, чтобы упростить вычисление коэффициентов представления сигналов по таким функциям: в случае разделимых базисов, являющихся произведением функций одной переменной, вычисление двумерного интеграла скалярного произведения сводится к вычислению двух одномерных интегралов.

1.2. ИНТЕГРАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ СИГНАЛОВ

Представление сигналов как элементов конечномерного линейного пространства со скалярным произведением, ставящее сигналу в соответствие (при заданном базисе) набор чисел, можно назвать *дискретным.* Оно является основой цифрового описания непрерывных сигналов. Для того чтобы понять, как получается такое описание, удобно рассматривать дискретное представление сигналов

$$a(x) = \sum_{k=0}^{N-1} \alpha_k \varphi_k(x) \qquad (1\text{-}15)$$

как частный случай непрерывного представления, которое получается, если заменить номер базисной функции k непрерывкой переменной f∈F, где F – конечный ила бесконечный интервал [37]. Тогда аналогом (1.15) будет формула

$$a(x) = \int_F \alpha(f)\, \varphi(x, f)\, df; \quad x \in X \qquad (1.16)$$

Естественно распространить такой подход и на способ определения $\alpha(f)$ по $a(x)$, введя взаимные функции $\psi(f, x)$ или *сопряженное базисное ядро:*

$$\alpha(f) = \int_X a(x)\, \psi(f, x)\, dx \qquad (1.17)$$

Функцию $\alpha(f)$ называют интегральным преобразованием сигнала *a(x),* или его *спектром по непрерывному базису.*

Условие взаимности функций $\varphi(x, f)$ и $\psi(f, x)$ можно получить, подставив (1.17) в (1.16):

$$a(x) = \int_F \left[\int_X a(\xi)\, \psi(f, \xi)\, d\xi \right] \varphi(x, f)\, df = \int_X a(\xi)\, \delta(x, \xi)\, d\xi \qquad (1\text{-}18)$$

где

$$\delta(x, \xi) = \int_F \varphi(x, f)\, \psi(f, \xi)\, df \qquad (1.19)$$

Таким образам, функции $\varphi(x, f)$ и $\psi(f, x)$ взаимны, если интеграл (1.19) от их произведения удовлетворяет условию (1.18). Функция $\delta(x, \xi)$, определяемая соотношением (1.19), называется *дельта-функцией.* Она может рассматриваться как обобщение символа Кронекера (1.3) взаимности базисных функций конечномерного пространства сигналов.

Каждой паре взаимных базисных функций соответствует, вообще говоря, своя дельта-функция. Так, например, базисным функциям, полученным с помощью предельного перехода при $\Delta x \to 0$ из импульсных базисных функций (1.6а), соответствует дельта-функция вида

$$\delta(x, \xi) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \operatorname{rect}(x - \xi)/\Delta x = \begin{cases} \infty, & x = \xi, \\ 0, & x \neq \xi, \end{cases} \qquad (1.20)$$

а формулу (1.18) можно рассматривать как вариант интегрального представления сигнала по базису из дельта-функций. Другие примеры будут приведены ниже при рассмотрении других интегральных преобразований.

Для базисов типа сдвиговых дельта-функция оказывается функцией одного аргумента: $\delta(x, \xi) = \delta(x - \xi)$. При этом (1.18) переходит в

$$a(x) = \int_X a(\xi)\, \delta(x - \xi)\, d\xi \qquad (1.21)$$

Пользуясь понятиями непрерывного представления, дискретное представление сигналов (1.15) можно записать так:

$$a(x) = \sum_k \alpha_k \varphi(x, f_k) \qquad (1\text{-}22)$$

где $\{f_k\}$ – дискретные значения непрерывного аргумента f базисной функции $\varphi(x, f)$, соответствующие базисным функциям $\varphi_k(x)$ в (1.15). Подставив (1.22) в (1.17), найдем связь дискретного представления сигнала {a*} и его непрерывного представления – спектра $\alpha(f)$:

$$\alpha(f) = \sum_k a_k \int_X \varphi(x, f_k)\, \psi(f, x)\, dx = \sum_k a_k \delta(f, f_k) \tag{1.23}$$

Это выражение можно рассматривать как непрерывный способ записи дискретного спектра.

Пользуясь (1.18) как аналогом (1.3). можно обобщить и понятие ортонормального базиса:

$$\int_F \varphi(x, f)\, \varphi^*(\xi, f)\, df = \delta(x, \xi)$$

Базис $\varphi(x, f)$, удовлетворяющий этим соотношениям, называется *самосопряженным.* Для самосопряженного базиса справедливо следующее соотношение:

$$\int_X a_1(x)\, a_2^*(x)\, dx = \int_F \alpha_1(f)\, \alpha_2^*(f)\, df \tag{1.24}$$

являющееся непрерывным аналогом соотношения Парсеваля (1.5). В частности, когда $a_1(x) = a_2(x)$, (1.24) переходит в

$$\int_X |a(x)|^2\, dx = \int_F |\alpha(f)|^2\, df$$

Все базисные функции, описанные в предыдущем параграфе, могут использоваться также в качестве непрерывных базисов, порождая соответствующее интегральное преобразование. Из них важнейшими в цифровой оптике являются преобразования, выполняемые над сигналами в оптических системах: преобразование Фурье, а также тесно связанные с ним преобразования Ганкеля, Френеля, Меллина и Гильберта.

Преобразование Фурье определяется в одномерном случае как

$$\alpha(f) = \int_{-\infty}^{\infty} a(x) \exp(i 2\pi f x)\, dx \tag{1.25а}$$

Согласно интегральной теореме Фурье $a(x)$ можно получить из $\alpha(f)$ в результате обратного преобразования Фурье:

$$a(x) = \lim_{F \to \infty} \int_{-F}^{F} \alpha(f) \exp(-i 2\pi f x)\, df \tag{1.25б}$$

откуда следует, что ядро $\varphi(f, x)$ является самосопряженным, причем дельта-функция, соответствующая этому ядру, понимается в смысле:

$$\delta(x - \xi) = \lim_{F \to \infty} \int_{-F}^{F} \exp(i 2\pi f x) \exp(-i 2\pi f \xi)\, df =$$

$$= \lim_{F \to \infty} 2F \frac{\sin 2\pi F (x - \xi)}{2\pi F (x - \xi)} = \lim_{F \to \infty} 2F \operatorname{sinc} 2\pi F (x - \xi) \tag{1.26}$$

Имея это в виду, *обратное преобразование Фурье* записывают так:

$$a(x) = \int_{-\infty}^{\infty} \alpha(f) \exp(-i 2\pi f x)\, df \tag{1.25в}$$

Преобразование Фурье двумерных сигналов обычно определяют в прямоугольной системе координат, как

$$\alpha(f_1, f_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(x_1, x_2) \exp[i 2\pi (f_1 x_1 + f_2 x_2)]\, dx_1 dx_2 \tag{1.27}$$

Благодаря использованию прямоугольной системы координат *двумерное преобразование Фурье* распадается на два одномерных:

$$\alpha(f_1, f_2) = \int_{-\infty}^{\infty} \exp(i 2\pi f_2 x_2)\, dx_2 \int_{-\infty}^{\infty} a(x_1, x_2) \exp(i 2\pi f_1 x_1)\, dx_1$$

Значение преобразования Фурье в оптике определяется тем, что оно связывает комплексную амплитуду волнового поля на объекте и в дальней зоне дифракции [27]. Важнейшими свойствами преобразования Фурье, которые находят многочисленные применения в системах

оптической обработки информации (см., например, [13]), являются *теорема о свертке,* согласно которой спектр Фурье свертки двух сигналов $\int_{-\infty}^{\infty} a(\xi) b(x-\xi)\, d\xi$ равен произведению спектров этих сигналов $\alpha(f)\beta(f)$, и *теорема о сдвиге,* согласно которой модуль спектра сигнала инвариантен к его сдвигу по координатам.

Если двумерный сигнал обладает круговой симметрией, т.е. является, по существу, функцией одной переменной – радиуса в полярной системе координат, то и его двумерное преобразование Фурье является функцией одной переменной:

$$\alpha_H(f) = \int_0^{\infty} r a(r)\, J_0(fr)\, dr, \qquad (1.28)$$

где $J_0(x)$ – функция Бесселя первого рода нулевого порядка:

$$J_0(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\, x \cos(\theta - \theta_0))\, d\theta$$

Это преобразование называется *преобразованием Ганкеля.* Поскольку оно является двумерным преобразованием Фурье функций, обладающих круговой симметрией, его свойства выводятся из соответствующих свойств преобразования Фурье. Спектр Ганкеля $\alpha_H(f)$ и двумерный спектр Фурье $\alpha(f_1, f_2)$ сигнала *c* круговой симметрией связаны соотношением

$$\alpha(f_1, f_2) = \alpha_H\left(\sqrt{f_1^2 + f_2^2}\right)$$

Обратное преобразование Ганкеля совпадает с прямым.

Свойство инвариантности к сдвигу преобразования Фурье можно превратить в свойство инвариантности к масштабу, если ввести логарифмическую шкалу по аргументу сигнала. Действительно, заменив в (1.25) x=ln z, получим

$$\alpha(f) = \int_0^{\infty} a(\ln z) \exp(i\, 2\pi f \ln z)\, dz/z = \int_0^{\infty} a(\ln z)\, z^{i2\pi f - 1}\, dz$$

Так мы приходим к преобразованию, которое называется *преобразованием Меллина:*

$$\mu(f) = \int_0^{\infty} a(z)\, z^{i2\pi f - 1}\, dz. \qquad (1.29)$$

Простой подстановкой $\bar{z} = kz$ нетрудно получить, что преобразование Меллина $\bar{\mu}(f)$ сигнала $a(kz)$ равно $(k)^{i2\pi f}\mu(f)$, где $\mu(f)$–преобразование Меллина сигнала a(z), т.е.

$|\bar{\mu}(f)| = |\mu(f)|$

Двумерное преобразование Меллина порождается двумерным преобразованием Фурье и определяется как разделимое по двум переменным:

$$\mu(f_1, f_2) = \int_0^{\infty}\int_0^{\infty} a(z_1, z_2)\, z_1^{i2\pi f_1 - 1} z_2^{i2\pi f_2 - 1}\, dz_1 dz_2 \qquad (1.30)$$

Преобразование Френеля, играющее большую роль в обработке волновых полей как приближение к дифракционному интегралу Кирхгофа ([27]), определяется в одномерном случае как

$$\alpha_D(f) = \int_{-\infty}^{\infty} a(x) \exp[-i\pi (x - f)^2/D^2]\, dx \qquad (1.31a)$$

где D – некоторый параметр преобразования (иногда (см., например, [27]), встречается несколько другое определение преобразования Френеля:

$$\alpha_\sigma(f) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} a(x) \exp(ix^2/2\sigma^2) \exp(-i2\pi f x)\, dx$$

которое, как нетрудно видеть, тесно связано с данным определением). При трактовке (1.31a) как соотношения, связывающего комплексную амплитуду волнового поля на объекте и амплитуду волнового поля в зоне Френеля объекта, D^2 есть произведение длины волны поля и расстояния между объектом и местом наблюдения поля.

Преобразование Френеля тесно связано с преобразованием Фурье:

$$\alpha_D(f) = \exp(-i\pi f^2/D^2) \int_{-\infty}^{\infty} [a(x) \exp(-i\pi x^2/D^2)] \exp(i\, 2\pi x f/D^2)\, dx$$

Отсюда нетрудно получить, что обратное преобразование Френеля определяется выражением

$$a(x) = \int_{-\infty}^{\infty} a_D(f) \exp[i\pi(x-f)^2/D^2]\, df \qquad (1.31б)$$

а дельта-функция, соответствующая ядру преобразования Френеля, – выражением

$$\delta(x, \xi) = \exp[-i\pi(x^2-\xi^2)/D^2] \lim_{F\to\infty} 2F \operatorname{sinc} 2\pi F(x-\xi) \qquad (1.32)$$

Результат преобразования Френеля произвольной функции – это, вообще говоря, комплексная функция. Ее модуль обычно описывает амплитуду волнового поля, а фаза – фазу этого поля. Иногда нет необходимости определять *фазу* поля (как, например, в случае, когда интересуются только яркостью изображения, восстанавливаемого с помощью голограммы, или пространственным распределением только интенсивности излучения системы, и в других подобных случаях). Для этих случаев можно ввести упрощенное, или *частичное преобразование* Френеля, записываемое как:

$$\tilde{a}_D(f) = \int_{-\infty}^{\infty} a(x) \exp(-i\pi x^2/D^2) \exp(i2\pi f x/D^2)\, dx \qquad (1.33а)$$

Нетрудно видеть, что обратным этому преобразованию будет несколько модифицированное обратное преобразование Фурье:

$$a(x) = \exp(i\pi x^2/D^2) \int_{-\infty}^{\infty} \tilde{a}_D(f) \exp(-i2\pi f x/D^2)\, df \qquad (1.33б)$$

Двумерные преобразования Френеля и частичное преобразование Френеля определяются как разделимые по двум переменным:

$$a_D(f_1, f_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(x_1, x_2) \exp\left\{-i\frac{\pi}{D^2}[(x_1-f_1)^2 + (x_2-f_2)^2]\right\} dx_1 dx_2 \qquad (1.34)$$

$$\tilde{a}_D(f_1, f_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(x_1, x_2) \exp\left[-i\frac{\pi}{D^2}(x_1^2+x_2^2)\right] \times \exp\left[i\frac{2\pi}{D^2}(f_1x_1 + f_2x_2)\right] dx_1 dx_2 \qquad (1.35)$$

Преобразование Гильберта является, как и преобразование Френеля, примером преобразования с ядром, зависящим от разности аргументов:

$$\hat{a}(f) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{a(x)}{f-x}\, dx \qquad (1.36а)$$

где интеграл понимается в смысле

$$\int_{-\infty}^{\infty} = \lim_{\varepsilon\to 0}\left(\int_{-\infty}^{f-\varepsilon} + \int_{f+\varepsilon}^{\infty}\right), \quad \varepsilon > 0$$

Обратное преобразование Гильберта имеет вид:

$$a(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\hat{a}(f)}{f-x}\, df \qquad (1.36б)$$

Прямое и обратное преобразования Гильберта имеют вид свертки. Поэтому преобразования Фурье $\alpha(f)$ и $\hat{\alpha}(f)$ сигналов *a(x)* и $\hat{a}(\xi)$, связанных преобразованием Гильберта, выражаются друг через друга с помощью очень простого соотношения:

$$\hat{\alpha}(f) = -i\alpha(f)\operatorname{sign} f \qquad (1\text{-}37)$$

так как преобразованием Фурье от ядра $1/\pi(f-x)$ является функция $i \operatorname{sign} f$ ([27]). В теории сигналов эта связь чаще всего используется для получения так называемого *комплексного аналитического сигнала c(x),* преобразование Фурье которого $\zeta(f)$ отлично от нуля только при f>0. Этот сигнал определяется как:

$$c(x) = a(x) + i\hat{a}(x) \qquad (1.38)$$

Беря преобразование Фурье от обеих частей и учитывая (1.37), получаем:

$$\zeta(f) = \begin{cases} 2\alpha(f), & f > 0; \\ 0, & f < 0. \end{cases}$$

Аналитический сигнал строится обычно только для вещественных сигналов. В этом случае из (1.38) вытекает:

$$a(x) = \mathrm{Re}\{c(x)\}$$

Замена сигнала аналитическим сигналом производится тогда, когда возникает необходимость аналитического описания так называемых узкополосных высокочастотных сигналов, т.е. сигналов, чей спектр Фурье отличен от нуля в небольшой окрестности, удаленной от начала координат.

Модуль аналитического сигнала $c(x) = A(x)\exp[i\Theta(x)]$ называется *огибающей* породившего его сигнала $a(x)$:

$$A(x) = \sqrt{a^2(x) + \hat{a}^2(x)}. \qquad (1.39)$$

Такое название объясняется тем, что функции $A(x)$ и $a(x)$ не пересекаются и имеют общие касательные в общих точках:

$$|A(x)| \geqslant |a(x)|; \quad \left.\frac{dA(x)}{dx}\right|_{A(x)=a(x)} = \frac{da(x)}{dx}$$

Фаза комплексной функции

$$\Theta(x) = \mathrm{Im}[\ln c(x)] = \mathrm{arctg}(\hat{a}(x)/a(x)) \qquad (1.40)$$

называется *фазой комплексной огибающей.*

Физический смысл введенных определений можно проследить, рассмотрев их для синусоидальных сигналов с изменяющейся амплитудой $u_0(x)$ и фазой $\Theta_0(x)$:

$$a(x) = u_0(x)\cos(2\pi f_0 x + \Theta_0(x))$$

Такими сигналами, например, описываются голограммы и интерферограммы. Представим $a(x)$ в виде:

$$a(x) = a_1(x) - a_2(x) = u_0(x)\cos\Theta_0(x)\cos 2\pi f_0 x -$$
$$- u_0(x)\sin\Theta_0(x)\sin 2\pi f_0 x.$$

Пусть спектр Фурье функций $u_0(x)\cos\Theta_0(x)$ и $u_0(x)\sin\Theta_0(x)$ не выходит за пределы интервала $(-f_0, f_0)$. Тогда преобразования Гильберта функций $a_1(x)$ и $a_2(x)$ будут равны

$$\hat{a}_1(x) = u_0(x)\cos\Theta_0(x)\sin 2\pi f_0 x,$$
$$\hat{a}_2(x) = -u_0(x)\sin\Theta_0(x)\cos 2\pi f_0 x,$$

а аналитический сигнал $c(x)$, соответствующий $a(x)$:

$$c(x) = u_0(x)\cos(2\pi f_0 x + \Theta_0(x)) + i\,u_0(x)\sin(2\pi f_0 x + \Theta_0(x))$$

Отсюда $A(x) = u_0(x)$; $\Theta(x) = \Theta_0(x)$

Таким образом, для сигнала с амплитудной и угловой модуляцией при условии, что f_0 достаточно велика, определение огибающей и фазы по Гильберту совпадает с естественной физической трактовкой огибающей и фазы таких сигналов.

1.3. ПРЕОБРАЗОВАНИЯ СИГНАЛОВ И МОДЕЛИ ОПТИЧЕСКИХ СИСТЕМ

Оптические сигналы в оптических системах претерпевают разнообразные преобразования. Для математического описания этих преобразований в общем случае нужно задать все возможные пары сигналов до и после преобразования (входных и выходных сигналов). Но это – не конструктивный способ, ибо объем такого описания даже для цифровых сигналов настолько велик, что исключает практическую возможность составления (и реализации в устройствах обработки сигналов) подобных списков. Поэтому описание преобразований сигналов и модели систем преобразований сигналов строятся по иерархическому принципу, т.е. преобразования сигналов представляются как совокупность некоторых «элементарных» преобразований, каждое из которых может быть описано с помощью небольшого подмножества из всех возможных пар «вход – выход». Важнейшими из таких преобразований являются так называемые поэлементные преобразования и линейные преобразования.

Понятие поэлементных преобразований проще всего пояснить для случая дискретных сигналов. Преобразование T дискретного сигнала $a=\{a_k\}$ называется *поэлементным,* если в результате него получается дискретный сигнал вида:

$$b = Ta = \{T_k(a_k)\}, \quad k = 0, 1, \ldots, N-1 \qquad (1.41)$$

т.е. если преобразование сводится к поэлементному функциональному преобразованию отсчетов элементов сигнала. Если функции не зависят от $T_k(\cdot)$, преобразование называется *однородным.*

Очевидно, для задания поэлементного преобразования цифрового сигнала, содержащего *N* отсчетов, принимающих *M* значений, достаточно задать *N* таблиц по *M* чисел, а если преобразование является однородным, то только одну таблицу из *M* чисел.

Поэлементные преобразования непрерывных сигналов полностью описываются функциональной зависимостью значений выходного сигнала в точке от значений входного в той же точке:

$$b(x) = T_x\{a(x)\} \qquad (1.42)$$

В случае однородного преобразования функция $T_x(\cdot)$ не зависит от *x*.

Характерным примером поэлементного преобразования является преобразование энергии светового излучения в почернение фотографической пластинки, описываемое так называемой характеристической кривой фотоматериала. Блоки, осуществляющие поэлементные преобразования сигналов, широко используются в моделях оптических и фотографических систем.

Линейные преобразования – это преобразования, для которых выполняется принцип суперпозиции. Математически он записывается следующим образом. Преобразование **L** является линейным, если для любых сигналов $\mathbf{a}_1$ и $\mathbf{a}_2$, заданных в линейном пространстве, и скаляров α_1 и α_2:

$$\mathbf{L}(\alpha_1 a_1 + \alpha_2 a_2) = \alpha_1 \mathbf{L}a_1 + \alpha_2 \mathbf{L}a_2 \qquad (1.43)$$

Очевидно, $\mathbf{L}\varnothing = \varnothing$; $\mathbf{L}(-\mathbf{a}) = -\mathbf{L}(\mathbf{a})$, т.е. множество линейно-преобразованных сигналов также образует линейное пространство. Для линейных преобразований поэтому можно ввести операцию произведения $\mathbf{L} = \mathbf{L}_1\mathbf{L}_2$, определяемую выражением $\mathbf{L}_1(\mathbf{L}_2\mathbf{a}) = \mathbf{L}_1\mathbf{L}_2\mathbf{a}$ и суммы $\mathbf{L}_1+\mathbf{L}_2$.

Физическим эквивалентом произведения является *последовательное (каскадное) соединение* блоков, реализующих преобразования-множители. Физическим эквивалентом суммы линейных преобразований является *параллельное соединение блоков,* реализующих преобразования-слагаемые.

Благодаря линейности преобразований умножение дистрибутивно по отношению к сложению:

$$\mathbf{L}_1(\mathbf{L}_2+\mathbf{L}_3) = \mathbf{L}_1\mathbf{L}_2 + \mathbf{L}_1\mathbf{L}_3; \quad (\mathbf{L}_1+\mathbf{L}_2)\,\mathbf{L}_3 = \mathbf{L}_1\mathbf{L}_3 + \mathbf{L}_2\mathbf{L}_3.$$

Если линейное преобразование L осуществляет взаимно однозначное отображение сигнала, то существует обратное преобразование L^{-1}, такое, что

$$\mathbf{L}\mathbf{L}^{-1}\mathbf{a} = \mathbf{L}^{-1}\mathbf{L}\mathbf{a} = \mathbf{a} \qquad (1.44)$$

Линейное преобразование дискретных сигналов полностью характеризуется матрицей $\underline{\Lambda} = \{\lambda_{k,n}\}$, связывающей входные и выходные последовательности $\mathbf{A} = \{\alpha_k\}$ и $\mathbf{B} = \{\beta_n\}$ линейного преобразования:

$$\mathbf{B} = \Lambda\mathbf{A} = \left\{\beta_n = \sum_k \alpha_k \lambda_{k,n}\right\} \qquad (1.45)$$

Таким образом, для описания линейных преобразований последовательностей объемом в N чисел достаточно задать матрицу из N^2 чисел (а не M^n, как в общем случае).

Формула (1.45) показывает, что в общем случае линейные преобразования не являются поэлементными и становятся таковыми только при $\lambda_{k,n} = c_k\delta_{k-n}$, где δ_{k-n}–символ Кронекера.

Интегральные представления, рассмотренные в предыдущем параграфе, являются инструментом конструктивного описания линейных преобразований непрерывных сигналов.

Пусть $\alpha(f)$ –спектр входного сигнала *a(x)* линейного преобразования L по базису $\varphi(x, f)$, $\beta(p)$– спектр его выходного сигнала $b(\xi) = \mathbf{L}a(x)$ по базису $\eta(\xi, p)$. Спектры $\alpha(f)$ и $\beta(p)$ связаны между собой выражением:

$$\beta(p) = \int_F \alpha(f)\, H(f, p)\, df, \qquad (1.46)$$

где

$$H(f, p) = \int_T [L\varphi(x, f)]\, \eta(\xi, p)\, d\xi \qquad (1.47)$$

Таким образом, линейное преобразование непрерывных сигналов описывается формулой интегрального преобразования (1.46) и полностью определяется ядром преобразований *(1.47)* при заданных базисах интегральных преобразований входных и выходных сигналов. Ядро (1.47) линейного преобразования имеет простой физический смысл – это семейство спектров по базису выходных сигналов функций, являющихся откликами линейного преобразования на базисные функции входных сигналов. Очевидно также, что интегральные преобразования сигналов, описанные в предыдущем параграфе, можно рассматривать как варианты линейных преобразований с ядрами, определяемыми базисами преобразований.

Наиболее распространенным способом описания линейных систем (систем, осуществляющих линейные преобразования сигналов) является их описание с помощью *импульсной реакции,* (отклика), определяемой как отклик оператора на дельта-функцию:

$$h(x, \xi) = L\delta(x, \xi) \qquad (1.48)$$

Такое описание соответствует базису из дельта-функций. Нетрудно видеть, что для этого базиса ядро линейного преобразования совпадает с импульсной реакцией:

$$b(x) = \int_{\xi} a(\xi)\, h(x, \xi)\, d\xi \qquad (1\text{-}49)$$

В теории оптических изображающих систем импульсная реакция (1.48) называется *функцией рассеивания точки.*

Другим часто используемым базисом является базис интегрального преобразования Фурье $\varphi(f, x) = \exp(-i2\pi fx)$. Этот базис приводит к частотному представлению сигналов и линейных систем. В этом случае $\alpha(p)$ и $\beta(f)$ –преобразования (спектры) Фурье сигналов $a(x)$ и $b(\xi)$. Ядром линейного преобразования по отношению к этому базису является двумерное преобразование Фурье импульсной реакции:

$$H(f, p) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, \xi)\, \exp[i2\pi(fx - p\xi)]\, dx d\xi \qquad (1.50)$$

– так называемая *частотная характеристика* (в теории оптических изображающих и фотографических систем иногда используется термин *«частотно-контрастная характеристика»).*

Ядро линейного преобразования, описывающее линейные системы для одномерных сигналов – функция двух переменных; в случае двумерных сигналов – это функция четырех переменных. Спектры входных и выходных сигналов связаны между собой интегральным соотношением (1.46). Это описание и соотношение существенно упрощаются, если в качестве базисов входных сигналов системы выбирать так называемые *собственные базисы системы,* т.е. функции, вид которых не меняется под действием линейного преобразования в системе:

$$Le(f, x) = E(f)\, e(f, x) \qquad (1.51)$$

а в качестве базиса выходных сигналов – взаимные к ним в смысле (1.18), (1,19) функции. В этом случае (1.46) переходит в

$$H(f, p) = E(f)\, \delta(f, p) \qquad (1.52а)$$

а *(1.47)*– в

$$\beta(p) = E(p)\, \alpha(p) \qquad (1.52б)$$

т.е. ядро преобразования является, по существу, функцией только одной переменной (в двумерном случае – двух переменных), а интегральное соотношение между спектрами входных и выходных сигналов превращается в простое произведение спектра входного сигнала на спектр оператора.

Важным частным случаем линейных систем являются системы, импульсная реакция которых зависит только от разности координат входного и выходного сигнала:

$$h(x, \xi) = h(x - \xi) \qquad (1.53)$$

так что сдвиг входного сигнала преобразования приводит к такому же сдвигу выходного сигнала без изменения его формы.

Такие системы называются *инвариантными к сдвигу* или *пространственно-\ инвариантными.* Понятие пространственной инвариантности тождественно понятию

однородности, введенному для поэлементных преобразований. Для пространственно-инвариантных систем формула (1.49) переходит в уже упоминавшийся в § 1.2 интеграл свертки

$$b(x) = \int_{-\infty}^{\infty} a(\xi)\, h(x-\xi)\, d\xi. \qquad (1.54)$$

Очевидно, что понятие пространственно-инвариантной системы по необходимости предполагает, что область определения сигнала не ограничена, иначе на краях этой области происходили бы нарушения инвариантности к сдвигу.

Собственным базисом пространственно-инвариантных систем является базис преобразования Фурье, что находит свое подтверждение в теореме о свертке для преобразования Фурье, которая в данном случае записывается так:

$$\beta(f) = H(f)\, \alpha(f) \qquad (1.55)$$

где $H(f)$ – частотная характеристика системы:

$$H(f) = \int_{-\infty}^{\infty} h(x) \exp(i 2\pi f x)\, dx \qquad (1.56)$$

Пространственная инвариантность систем существенно зависит от системы координат, в которых рассматриваются действующие в системах сигналы. Некоторые системы становятся пространственно-инвариантными, если их входные и выходные сигналы подвергнуть преобразованиям координат, или *геометрическим преобразованиям.* Например, пусть

$$h(x, \xi) = h(x/\xi) \qquad (1.57)$$

Тогда, перейдя к координатам:

$$x = \exp s; \quad \xi = \exp \sigma; \quad x, \xi \in [0, \infty] \qquad \{1\text{-}58\}$$

$$b(x = \int_{-\infty}^{\infty} [\exp \sigma a(\exp \sigma)]\, h(\exp(s-\sigma))\, d\sigma$$

получим вместо (1.49) ,т.е. в новых координатах (1.58) система стала пространственно-инвариантной для сигналов вида $\exp \sigma a(\exp \sigma)$

Примерами оптических систем, которые после преобразования координат входных и выходных сигналов становятся пространственно-анвариантными, могут служить изображающие системы с аберрациями типа комы и некоторые аэрофотографические системы съемки Земли с самолетов или космических аппаратов.

Блоки, осуществляющие линейные преобразования (их обычно называют фильтрами), наряду с блоками поэлементных преобразований составляют тот арсенал средств, из которых строятся математические модели оптических и голографических систем при их детерминистическом описании.

При статистическом описании сигналов и их преобразований параметры этих блоков– функции $T_x(\cdot)$ поэлементных преобразований и ядро линейных преобразований считают случайными и, кроме того, из поэлементных преобразовавши выделяют особый класс случайных преобразований:

$$T_x\{a(x)\} = n_m(x)\, a(x) + n_a(x) \qquad (1.59)$$

где $n_m(x)$ и $n_a(x)$ –случайные процессы, которые описывают простейшие случайные искажения сигнала: так называемые *мультипликативный* $n_к(x)$ и *аддитивный* $n_з$ *(к)* шум. Случаю $n_m(x)=const$ соответствует модель аддитивного шума, случаю $n_a(x)=0$ – модель мультипликативного шума. В общем случае формула (1.59) описывает совместное действие на сигнал аддитивного и мультипликативного шума.

Важным частным случаем, описываемым моделью (1.59), является также так называемый *импульсный шум:*

$$T_x\{a(x)\} = e(x)\, a(x) + [1 - e(x)]\, n(x) \qquad (1.60)$$

где $e(x)$ – бинарный случайный процесс, принимающий значения 0 и 1. При $e(x)=1$ не происходит искажения сигнала; при $e(x)=0$ сигнал $a(x)$ заменяется случайным процессом $n(x)$. Вероятность того, что $e(x)=0$, называется *вероятностью ошибки.*

В общем случае поэлементные и линейные преобразования не являются однородными или пространственно-инвариантными,' и вид их может изменяться от точки к точке в области определения сигнала. Если этот вид не зависит от сигнала, и его изменения определяются

внешними по отношению к сигналу причинами, будем называть эти преобразования *управляемыми.* Но вид преобразований может зависеть от самого сигнала. Такие преобразования будем называть *адаптивными.* Если преобразование зависит от сигнала во всей области определения и остается одним и тем же для всей области определения сигнала, будем называть его *глобально-адаптивным.* Если вид преобразования зависит от значений сигнала в пределах некоторого фрагмента области определения, и при этом преобразование изменяет сигнал только внутри этого фрагмента, будем называть его *локально-адаптивным.*

Линейные преобразования являются адаптивными, если коэффициенты $\{\lambda_{k,n}\}$ матрицы линейного преобразования (1.45) зависят от элементов преобразуемого сигнала $\{a_k\}$. Адаптивные поэлементные преобразования, по существу, уже перестают быть поэлементными, так как функция преобразования $T_k(a_k)$ зависит от значений элементов сигнала, как правило, составляющих некоторую окрестность вокруг преобразуемого элемента.

В настоящее время в цифровой обработке сигналов появился новый класс адаптивных преобразований – так называемые *ранговые алгоритмы или фильтры* [23, 51]. Это алгоритмы обработки дискретных и цифровых сигналов, которые осуществляют преобразования вида

$$\mathbf{b} = T(\mathbf{a} = \{a_k\}) = \{b_k = T_k(a_k)\} \tag{1.61}$$

где $T_k(a_k)$ – нелинейная в общем случае функция, вид которой определяется некоторым подмножеством *ранговых или порядковых статистик* выборки, образованной элементами сигнала в некоторой окрестности данного элемента в последовательности упорядоченных элементов сигнала, *r-й порядковой статистикой m(r)* выборки элементов сигнала по окрестности S, содержащей N_S элементов, называется величина, имеющая r-й ранг, т.е. находящаяся на r-м месте в списке элементов выборки, упорядоченных по возрастанию, .или в *вариационном ряду* из N_S элементов.

Подробнее этот класс алгоритмов и его использование для обработки рассмотрен в гл. 9.

Глава 2
ЦИФРОВОЕ ПРЕДСТАВЛЕНИЕ ОПТИЧЕСКИХ СИГНАЛОВ

2.1 ПРИНЦИПЫ ДИСКРЕТИЗАЦИИ И КВАНТОВАНИЯ СИГНАЛОВ

Для обработки непрерывных сигналов в цифровых процессорах необходимо прежде всего преобразовать их в цифровые сигналы. Это преобразование представляет собой первый этап обработки, который выполняется с помощью специальных устройств – преобразователей аналог-код.

С общей точки зрения цифровой сигнал представляет собой некоторое число, записанное в системе исчисления значений сигнала и имеющее количество разрядов (цифр), равное количеству элементов цифрового сигнала. Преобразование непрерывного сигнала в цифровой можно трактовать как отображение пространственных непрерывных сигналов в конечное множество сигналов, составляющих так называемую ε-сеть этого пространства {42]: пространство сигналов разбивается на конечное множество подпространств (ε-областей), и в каждом подпространстве выбирается один принадлежащий ему сигнал-представитель так, чтобы все остальные сигналы в данном подпространстве можно было считать неотличимыми от сигнала-представителя в пределах заданной степени точности. Подпространства пронумеровываются и, таким образом, каждому аналоговому сигналу в пространстве сигналов, разбитом на области эквивалентности, может быть поставлен в соответствие номер сигнала-представителя. Этот номер и есть цифровой сигнал, соответствующий данному непрерывному сигналу. Определение этого номера (т.е. определение области эквивалентности, в которую попадает данный непрерывный сигнал) осуществляется специальным устройством – *преобразователем аналог-код.* После обработки цифрового сигнала требуется получить непрерывный сигнал, соответствующий результирующему цифровому сигналу, т.е. по номеру области эквивалентности сформировать ее сигнал-представитель (например, требуется сформировать обработанное изображение, голограмму или интерферограмму). Это осуществляется устройствами, получившими название *преобразователей код-аналог.* В цифровой оптике – это преобразователи код–оптический сигнал. В этой главе будут рассмотрены основные принципы, на которых основано построение преобразователей оптический сигнал – цифровой сигнал и цифровой сигнал – оптический сигнал, являющихся существенной составной частью цифровых систем обработки оптических сигналов.

Принципиальным вопросом является объем цифрового представления непрерывных сигналов, или количество сигналов-представителей. От него непосредственно зависит сложность (а значит, реализуемость и стоимость) цифровой системы обработки. Эта связь особенно наглядно проявляется при кодировании сигналов для их запоминания в цифровых запоминающих устройствах, где количество различных сигналов-представителей определяет требуемую емкость запоминающих устройств. Ясно, что при прочих равных условиях всегда необходимо стремиться к тому, чтобы минимизировать объем цифрового представления.

Объем цифрового представления сигнала удобно оценивать количеством двоичных разрядов (двоичных единиц, или бит), достаточных для перенумерации всех сигналов-представителей. Минимальное количество двоичных единиц, требуемое для цифрового описания непрерывных сигналов из данного класса, называется ε-энтропией этого класса. В теории информации показано, как определять ε-энтропию, исходя из статистического описания класса (ансамбля) сигналов для заданного критерия точности воспроизведения сигналов их квантованными представителями.

Численная оценка ε-энтропии для оптических сигналов, в частности таких сигналов, как изображения, является нерешенной задачей, так как вопрос о количественном критерии отличия одного изображения от другого в конкретных практических задачах остается до настоящего времени открытым. Оценкой сверху значения ε-энтропии для изображений может служить (как это следует из достигнутых на сегодняшний день результатов по кодированию

изображений с сохранением высокого визуального качества воспроизведения) произведение площади изображения на площадь его пространственного спектра Фурье.

Оценка объема сигнала, необходимого для воспроизведения объемных стереоскопических изображений (см. § 8.4), показывает,, что он лишь незначительно превышает объем сигнала для плоского изображения. Данных об ε-энтропии других оптических сигналов в настоящее время не имеется, ε-энтропия – потенциальный нижний предел объема сигнала, который никогда не достигается как из-за сложности процедуры отображения пространства непрерывных сигналов на его ε-сеть (ситуация здесь такая же, как и при описании преобразований сигналов в § 1.3), так и ввиду других практически важных обстоятельств, связанных с необходимостью получения цифрового описания, сохраняющего определенную аналогию с непрерывными сигналами (без нее трудно построить представление непрерывных преобразований сигналов).

Так же как для описания преобразований сигнала используют иерархические структуры, построенные из линейных и поэлементных преобразований, так и преобразование непрерывных сигналов в цифровые осуществляют чаще всего в виде последовательности элементарных процедур – *дискретизации* и *поэлементного квантования.* Дискретизация относится к классу линейных преобразований сигнала, поэлементное квантование – к классу поэлементных нелинейных преобразований.

Дискретизация – это замена непрерывного сигнала последовательностью чисел – представлением этого сигнала по какому-либо конечномерному базису. Это представление состоит в проектировании сигнала на данный базис, т.е. коэффициенты представления находятся как скалярные произведения сигнала, на соответствующие базисные функции:

$$\alpha_k = (\mathbf{a}, \varphi_k) \qquad (2.1)$$

Размерность базиса (количество коэффициентов $\{\alpha_k\}$ при дискретизации) ограничивают, основываясь на требуемой точности аппроксимации сигналов *a(x),* конечной суммой

$$a(x) \approx \tilde{a}(x) = \sum_{k=0}^{N-1} \alpha_k \varphi_k(x) \qquad (2.2)$$

В качестве базисов дискретизации чаще всего используются сдвиговые базисы (прямоугольные и отсчетные базисные функции, см. § 2.2). При кодировании изображений для запоминания в цифровых запоминающих устройствах и передачи по цифровым системам связи используются базисы Уолша, Хаара, экспоненциальные базисные функции, а также ряд специальных базисов, введенных с учетом специфики видеосигнала для получения наилучшего приближения по (2.2) [64]. Некоторые из них, а также некоторые принципы синтеза новых базисов с заданными свойствами рассмотрены в гл. 5.

В теории дискретизации важным является вопрос об объеме дискретного описания сигнала, т.е. о количестве *N* базисных функций, используемых для представления (2.2), или о размерности конечномерного пространства, на которое проектируется сигнал при дискретизации. Естественно считать оптимальным такой способ дискретизации, при котором размерность базиса минимальна при заданной точности восстановления сигнала. Чтобы найти оптимальный базис, нужно прежде всего определить класс сигналов, для которых он отыскивается, а также определить меру точности восстановления для этого класса.

При детерминистическом задании класса сигналов они рассматриваются как результат преобразования произвольных сигналов некоторым линейным оператором (фильтром), так что разные классы различаются между собой и описываются характеристиками этого оператора. При таком подходе вопрос о выборе оптимального базиса сводится к нахождению связи между характеристиками линейного оператора, определяющего класс сигналов, и базисными функциями. Примером может служить оценка размерности пространства сигналов почти ограниченной протяженности с почти ограниченным спектром [37].

Пусть *a(x)*–сигнал, удовлетворяющий следующим условиям:

$$(\mathbf{P}_x\mathbf{a}, \mathbf{a}) = 1 - \varepsilon_x^2;$$
$$(\mathbf{P}_f\mathbf{a}, \mathbf{a}) = 1 - \varepsilon_p^2;$$
$$(\mathbf{a}, \mathbf{a}) = \|\mathbf{a}\|^2 = 1 \qquad (2.3)$$

где P*x*.– оператор стробирования, выделяющий из сигнала участок протяженностью *X; P_f* – идеальный полосовой фильтр, пропускающий только частоты спектра сигнала в интервале (*–F, F)*; ε_x^2, ε_F^2 – ошибки такого усечения сигнала по протяженности и по спектру.

Тогда наилучшим является представление сигнала по функциям, являющимся решением уравнения

$$2F\int_{-X/2}^{X/2}\varphi_k(\xi)\,\mathrm{sinc}\,2\pi F(x-\xi)\,d\xi=\lambda_k\varphi_k(x) \tag{2.4}$$

и называемым сфероидальными волновыми функциями (СВФ), причем

$$\|a-\sum_{k=0}^{N-1}\alpha_k\varphi_k\|^2\leqslant 12(\varepsilon_x+\varepsilon_F)^2+\varepsilon_F^2 \tag{2.5}$$

если *N*– наименьшее целое число, превышающее *2XF* ([37]). При $X\to\infty$ – сфероидальные волновые функции приближаются к отсчетным функциям $\mathrm{sinc}\,2\pi Fx$, и разложение по ним переходит в разложение по теореме отсчетов (см. § 2.2). При конечном *X* представление (2.2) по СВФ сигналов, заданных (2.3), лучше их разложения по отсчетным функциям при том же *N*.

При статистическом описании сигналов оптимальный *N-мер*ный базис для представления отдельных реализаций сигналов обычно определяют как базис, при котором норма ошибки, усредненная по ансамблю реализаций, минимальна.

В этом случае получается результат, известный как *теорема Карунена– Лоэва* [37]: минимальное значение нормы ошибки при представлении сигналов на интервале протяженностью *X* достигается при использовании в качестве базиса *N* собственных функций, соответствующих *N* наибольшим собственным значениям оператора, ядром которого является корреляционная функция сигналов $R_a(x,\xi)$:

$$\int_{-X/2}^{X/2}R_a(x,\xi)\varphi_k(\xi)\,d\xi=\lambda_k\varphi_k(x) \tag{2.6}$$

Это минимальное значение нормы ошибки равно

$$\|\varepsilon\|^2_{\min}=\|a(x)-\sum_{k=0}^{N-1}\alpha_k\varphi_k(x)\|^2_{\min}=\sum_{k=N}^{\infty}\lambda_k \tag{2.7}$$

Такое представление называется *разложением Карунена – Лоэва.* Коэффициенты разложения Карунена – Лоэва являются некоррелированными (ввиду ортогональности $\varphi_k(x)$) случайными величинами [37]. Отметим, что для стационарных процессов, когда корреляционная функция зависит только от разности аргументов ($R_a(x,\xi)= R_a(x-\xi)$) , при $X\to\infty$ *(X* становится достаточно большим по сравнению с протяженностью $R_a(x)$*)* собственные функции $\varphi_k(x)$ приближаются к комплексным экспонентам с частотами *k/X.*

Поэлементное квантование – это квантование каждого по отдельности из чисел $\{\alpha_k\}$, представляющих данный сигнал по заданному базису, т.е. замена непрерывной, и вообще говоря, бесконечной шкалы значений $\{\alpha_k\}$ дискретной и конечной (см. § 2.3).

Поэлементное квантование может быть *однородным,* когда шкала квантования одинакова для всех $\{\alpha_k\}$, и *неоднородным.* В последнем случае квантование делают *зональным,* т.е. разбивают все множество значений *k* на зоны, и в пределах каждой зоны осуществляют квантование $\{\alpha_k\}$ по своей шкале, или *адаптивным,* когда разбиение на зоны и квантование производятся адаптивно в зависимости от поведения сигнала. Например, отбрасываются (обнуляются) все коэффициенты преобразования (отсчеты), сумма квадратов которых не превышает заданной доли от суммы квадратов всех коэффициентов, или коэффициенты, не превышающие по абсолютной величине заданный порог, или разбиение на группы производится по сумме квадратов коэффициентов и т.д.

Использование для преобразования непрерывного сигнала в цифровой двухступенчатой процедуры из последовательно осуществляемых дискретизации и поэлементного квантования означает, что в пространстве сигналов строятся области эквивалентности, представляющие собой гиперкубы с гранями, параллельными координатным осям. Размеры граней определяются *шагом квантования*– интервалом квантования в данной точке шкалы чисел, соответствующих данной базисной функции. При этом задача оптимального квантования сводится к выбору

такой системы базисных функций и такого расположения шкалы квантования по координатным осям, соответствующим этим базисным функциям, при которых минимальное количество получающихся гиперкубов наилучшим образом упаковывается в ε-областях, соответствующих заданному критерию точности представления. Чтобы решить эту задачу, критерий точности воспроизведения непрерывных сигналов сигналами-представителями нужно разбить на два частных критерия: точности дискретизации и точности квантования дискретного представления сигнала.

Простейшая двухступенчатая процедура «дискретизация – поэлементное квантование» дает завышенные значения объема сигнала по сравнению с потенциальным нижним пределом. Чтобы уменьшить эту избыточность, используются усложненные схемы дискретизации, квантования и их сочетаний, из которых наиболее известны и апробированы схемы многоканальной дискретизации, дискретизации и квантования по фрагментам сигнала, адаптивной дискретизации, дискретизации и квантования с предсказанием и обратной связью, блочного квантования.

При многоканальной дискретизации сигнал представляется в виде двух или нескольких компонент, которые подвергаются дискретизации и квантованию раздельно, причем базис дискретизации и шкала поэлементного квантования выбираются для каждой компоненты независимо. Примерами могут служить известные методы кодирования изображений с раздельной дискретизацией и квантованием низкочастотной и высокочастотной компонент сигнала или низкочастотной компоненты и градиента видеосигнала.

Для того чтобы учесть пространственную неоднородность изображения, дискретизацию и квантование осуществляют по фрагментам. Обычно способ дискретизации (базисные функции и их число) одинаков для всех фрагментов, а изменяется только шкала квантования. В этом случае используется адаптивное квантование.

Другим способом учета пространственной неоднородности сигнала является *адаптивная дискретизация.* Сущность адаптивной дискретизации состоит в том, что при выбранном базисе дискретизации оставляются только те из возможных коэффициентов представления сигнала по данному базису, по которым остальные коэффициенты можно восстановить с заданной точностью. Примером является представление двумерного сигнала в виде его контурной карты (линий равных значений при заданном шаге квантования этих значений).

Идея восстановления сигнала по части его цифрового представления используется и при *дискретизации и квантовании с предсказанием и обратной связью.* Этот способ кодирования проще всего объяснить для одномерных сигналов. Для дискретизации используются сдвиговые базисы (прямоугольные импульсные функции, отсчетные функции), так что дискретизация и квантование происходит последовательно при движении вдоль сигнала, причем для каждого следующего отсчета сигнала квантуется значение разности между этим значением и значением, предсказанным по предыдущим отсчетам, восстановленным из прежних квантованных разностей. В простейшем случае предсказанное значение принимается равным значению предыдущего отсчета, восстановленного из цифрового разностного сигнала.

В заключение упомянем блочное квантование, которое является редуцированным вариантом общей процедуры, описанной в начале параграфа. В этом случае вместо разбиения всего пространства сигналов на области эквивалентности, что практически невозможно, разбиению подвергаются подпространства небольшой размерности, образуемые небольшими группами элементов дискретного сигнала, полученного после дискретизации. Практически это осуществляется так, что сначала производится поэлементное квантование, а потом блоки квантованных значений отсчетов дискретного сигнала записываются в виде одного числа, которое подвергается повторному квантованию.

2.2. ДИСКРЕТИЗАЦИЯ РАСТРИРОВАНИЕМ И ТЕОРЕМА ОТСЧЕТОВ

Наиболее удобным с точки зрения организации обработки и естественным способом дискретизации является представление сигналов в виде выборок их значений (отсчетов) в отдельных, регулярно расположенных точках. Чтобы выделить этот способ дискретизации среди других, будем называть его *растрированием.* Последовательность точек, в которых берутся отсчеты, называется растром.

Практически операция растрирования осуществляется путем измерения значений сигнала с помощью датчика, действие которого можно описать как свертку сигнала с некоторым ядром $\lambda_д(x)$:

$$\tilde{a}(k\Delta x) = \int_{-\infty}^{\infty} a(x)\lambda_д(x - k\Delta x)\,dx \qquad (2.8)$$

Набор значений $\{\tilde{a}(k\Delta x)\}$ составляет дискретное представление сигнала. Ядро $\lambda_д(x)$ называется *апертурой дискретизации.* Например, в устройствах дискретизации изображений $\lambda_д(x)$ описывает значения чувствительности датчика видеосигнала как функции координат в подвижной системе координат с началом в точке $k\Delta x$ При растрировании апертура дискретизации, как правило, такова, что величины $\{\tilde{a}(k\Delta x)\}$ близки к значениям сигнала $a(k\Delta x)$ в точках растра дискретизации.

Восстановление непрерывного сигнала из полученной последовательности его приближенных значений $\{\tilde{a}(k\Delta x)\}$ выполняется путем интерполяции его по этим значениям:

$$\tilde{\tilde{a}}(x) = \sum_k \tilde{a}(k\Delta x)\lambda_в(x - k\Delta x) \approx a(x) \qquad (2.9)$$

с помощью интерполирующей функции $\lambda_в(x)$, которая называется *апертурой восстановления.* Например, в электромеханических устройствах воспроизведения изображений $\lambda_в(x)$–функция, которая описывает распределение интенсивности светового пятна, осуществляющего экспонирование фотоматериала, как функцию координат в подвижной системе координат с началом в точке $k\Delta x$; в электронно-лучевых устройствах записи и телевизионных дисплеях– распределение яркости свечения пятна люминофора и т.п.

Таким образом, растрирование и восстановление непрерывного сигнала из растрированного можно трактовать с точки зрения теории дискретизации как представление сигнала по сдвиговому базису, причем набор апертур дискретизации $\{\lambda_д(x-k\Delta x)\}$ образует базис дискретизации, набор апертур восстановления $\{\lambda_в(x-k\Delta x)\}$ – базис восстановления. Выбор апертуры дискретизации, апертуры восстановления и шага растра дискретизации Δx определяется возможностями их реализации и требуемой точностью аппроксимации сигнала $a(x)$ восстановленным сигналом $\tilde{\tilde{a}}(x)$

Если исходить только из точности аппроксимации, то существует важный класс сигналов и соответствующие ему базисные функции, для которых представления (2.8) и (2.9) являются абсолютно точными. Это сигналы, спектр Фурье которых отличен от нуля только в пределах ограниченного участка области определения *(сигналы с ограниченным спектром).* К ним можно отнести и оптические сигналы – изображения и голограммы. Спектр Фурье изображений, получаемых в оптических изображающих системах, ограничен из-за ограниченности размеров линз, объективов и т.п. Для голограмм ограничение протяженности спектра связано с ограниченными размерами голографируемых объектов. Пусть спектр сигналов отличен от нуля на интервале $(-1/2\Delta x, 1/2\Delta x)$, т.е.

$$a(f) = a(f)\,\mathrm{rect}(f\Delta x + 1/2) \qquad (2.10)$$

Для таких сигналов базисы дискретизации и восстановления образуются из функций отсчетов:

$$\lambda_д(x) = (1/\Delta x)\,\mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x]$$
$$\lambda_в(x) = \mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x], \qquad (2.11)$$

а (2.8) и (2.9) переходят в точные равенства:

$$a(k\Delta x) = \frac{1}{\Delta x}\int_{-\infty}^{\infty} a(x)\,\mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x]; \qquad (2.12)$$

$$a(x) = \sum_{k=-\infty}^{\infty} a(k\Delta x)\,\mathrm{sinc}\,[\pi(x - k\Delta x)/\Delta x] \qquad (2.13)$$

Эти соотношения, означающие возможность точного восстановления сигналов с ограниченным спектром по последовательности их отсчетов, взятых на растре с шагом Δx, называются *теоремой отсчетов.* Равенство (2.12) означает, что отсчетами сигнала являются его значения в точках $\{k\Delta x\}$, полученные после пропускания сигнала через инвариантный к сдвигу фильтр с импульсной реакцией:

$$\lambda_{\text{д}}(x) = (1/\Delta x)\,\mathrm{sinc}\,(\pi x/\Delta x) \qquad (2\text{-}14)$$

и частотной характеристикой

$$H_{\text{д}}(f) = \mathrm{rect}\,(f\Delta x + 1/2) \qquad (2.15)$$

называемый идеальным фильтром нижних частот.

Равенство (2.13) означает, что процедуру восстановления непрерывного сигнала *a(x)* из его отсчетов $\{a(k\Delta x)\}$ можно представить как результат пропускания через идеальный фильтр нижних частот (2.14), (2.15) непрерывного сигнала вид

$$\tilde{a}(x) = \sum_{k=-\infty}^{\infty} a(k\Delta x)\,\delta(x - k\Delta x), \qquad (2.16)$$

спектр которого $\tilde{\alpha}(f)$, представляет собой периодически продолженный с периодом $(1/\Delta x)$ спектр $\alpha(f)$ сигнала *a(x):*

$$\tilde{\alpha}(f) = \sum_{m=-\infty}^{\infty} \alpha(f - m/\Delta x) \qquad (2.17)$$

Действительно, при такой фильтрации спектр $\tilde{\alpha}(f)$ (2.17) умножается на частотную характеристику фильтра (2.15), выделяющую только один период спектра, соответствующий *m* = 0 и равный спектру сигнала *a(x).* Периодическое продолжение (2.17.)

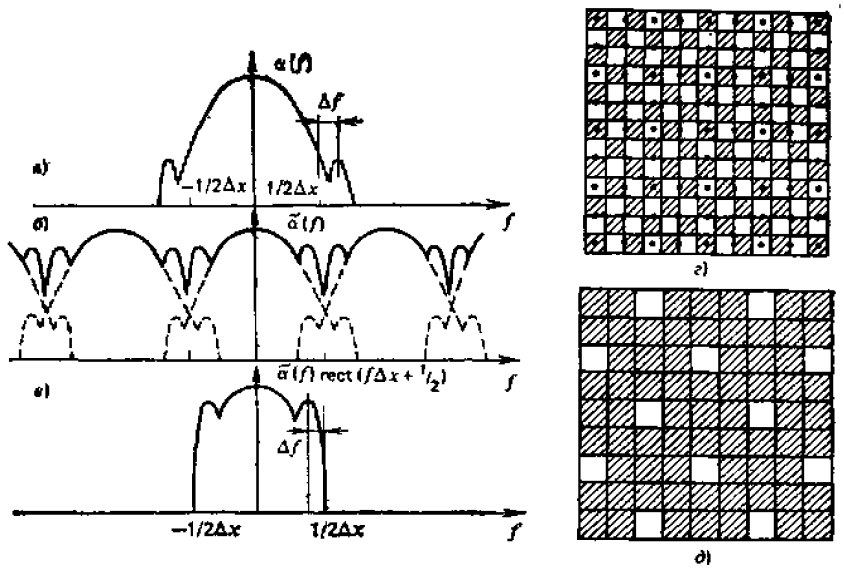


Рис. 2.1. Строб-эффект при растрировании сигналов:
а – спектр исходного непрерывного сигнала; б – спектр непрерывного сигнала *а* после растрирования: *в* – спектр сигнала, восстановленного после растрирования; г–изображение шахматной клетки (точки указывают расположение узлов растра); *д*– искаженное изображение рис. *г* вследствие строб-эффекта

спектра без перекрытия возможно, если шаг растрирования Δx меньше или равен величине, обратной протяженности спектра. В противном случае происходит перекрытие (перехлестывание) соседних периодов спектра сигнала *a(x),* и идеальным фильтром нижних частот уже невозможно выделить спектр сигнала в чистом виде. В восстановленном сигнале появляются излишние компоненты за счет наложившихся слева и справа на основной (нулевой) период спектра фрагментов спектра плюс первого, минус первого и следующих порядков (рис. 2.1). При этом, если в исходном сигнале они имели частоту, скажем, (l/2Δx)+ Δf, то в восстановленном сигнале их частота оказывается равной (1/2Δf)– Δf, то более низкой (см. рис. 2.1,о–в). Это явление снижения частоты периодических составляющих в сигнале при растрировании с шагом, не соответствующим максимальной частоте сигнала, называется *строб-эффектом.* Например, этим объясняется часто наблюдаемое в телевидении и кинематографе явление обратного вращения вращающихся объектов. Для двумерных сигналов строб-эффект проявляется в уменьшении пространственной частоты периодических структур, истинная пространственная частота которых близка к частоте дискретизации (рис. 2.1,г–*д).* Для того чтобы этих искажений не было, очевидно, необходимо перед растрированием с шагом Δx пропустить сигнал через идеальный фильтр нижних частот с полосой пропускания ($\pm 1/2\Delta x$).

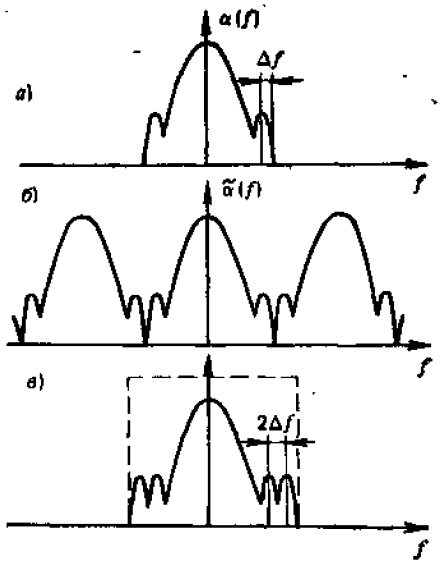


Рис. 2.2. Муар-эффект при растрировании сигналов:
а – спектр исходного непрерывного сигнала; б – спектр непрерывного сигнала после растрирования; в – спектр восстановленного сигнала (– – – – частотная характеристика восстанавливающего фильтра)

Сходные по своей природе искажения возникают, если восстанавливающий фильтр имеет полосу пропускания шире, чем $(\pm 1/2\Delta x)$. В этом случае в восстановленный после растрирования с шагом Δx; непрерывный сигнал попадают компоненты исходного сигнала из плюс-минус первого и более высоких порядков периодически продолженного спектра, так что если в исходном сигнале была составляющая с частотой $(1/2\Delta x)-\Delta f$, то в восстановленный сигнал из плюс первого порядка продолженного спектра попадет составляющая, имеющая частоту $(1/2\Delta x)+\Delta f$, и между этими двумя составляющими возникают биения с частотой $2\Delta f$. Это явление, называемое *муар-эффектом,* иллюстрируется рис. 2.2.

Теорема отсчетов может быть обобщена на сигналы, содержащие так называемую *несущую частоту.* Это сигналы, спектр которых отличен от нуля на ограниченных интервалах, смещенных относительно нулевой частоты. Таковы сигналы голограмм и интерферограмм.. Это обобщение можно выполнить двумя способами.

Во-первых, вместо данного вещественного сигнала *a(x)* можно рассмотреть аналитический сигнал $c(x)=a(x)+i\hat{a}(x)$, где $\hat{a}(x)$ – преобразование Гильберта *a(x).* Аналитический сигнал имеет односторонний спектр (см. § 1.3), и к нему теорема отсчетов применима уже в своем обычном виде:

$$c(x)=\sum_{k=-\infty}^{\infty} c\left(\frac{k}{f_п-f_л}\right)\lambda_a\left(x-\frac{k}{f_п-f_л}\right) \qquad (2.18)$$

где $c\left(\frac{k}{f_п-f_л}\right)$ – отсчеты аналитического сигнала,

$$\lambda_a(x)=\frac{1}{f_п-f_л}\int_{f_л}^{f_п}\exp(i2\pi fx)\,dx=$$
$$=\operatorname{sinc}[\pi(f_п-f_л)x]\exp[i\pi(f_п+f_л)x] \qquad (2.19)$$

а $f_л$ и $f_п$ – границы частотного интервала на положительных частотах, где спектр *c(x)* отличен от нуля.

Подставив (2.19) в (2.18), получим

$$c(x)=\sum_{k=-\infty}^{\infty} c\left(\frac{k}{f_п-f_л}\right)\operatorname{sinc}\left[\pi(f_п-f_л)\left(x-\frac{k}{f_п-f_л}\right)\right]\times$$
$$\times\exp\left[i\pi(f_л+f_п)\left(x-\frac{k}{f_п-f_л}\right)\right].$$

Так как $a(x)=\operatorname{Re}\{c(x)\}$, то

$$a(x)=\sum_{k=-\infty}^{\infty}\sqrt{a^2\left(\frac{k}{f_п-f_л}\right)+\hat{a}^2\left(\frac{k}{f_п-f_л}\right)}\,\mathrm{sinc}\,[\pi(f_п-f_л)x]\cos\left[\pi(f_п-f_л)x+\operatorname{arctg}\frac{\hat{a}(k/(f_п-f_л))}{a(k/(f_п-f_л))}\right] \quad (2.20)$$

Физический смысл этой формулы в том, что сигнал на несущей определяется отсчетами своей огибающей $A(x)=\sqrt{a^2(x)+\hat{a}^2(x)}$ и фазы $\Theta(x)=\operatorname{arctg}\hat{a}(x)/a(x)$. Количество его отсчетов, требуемых на единицу протяженности сигнала, определяется только шириной его полосы частот.

Другой способ представления состоит в следующем. Пусть спектр сигнала *a(x)* отличен от нуля на частотах в интервалах $(-f_п,-f_л]$ и $[f_л, f_п)$ (**эта** запись обозначает, что интервалы включают своя крайние точки: первый – правую, второй – левую). Подберем на оси частот такую ближайшую слева к *h* точку, чтобы ее координата $f'_л$ была кратна расстоянию $f_п-f'_л$: $f'_л=m(f_п-f_л')$, *m* – целое число. Обозначим $F=2(f_п-f_л')>2(f_п-f_л)$ и рассмотрим спектр

$$\tilde{a}(f)=\sum_{l=-\infty}^{\infty}a(f+lF), \quad (2\text{-}21)$$

который представляет собой периодически продолженный спектр α(f), причем его составляющие не перекрываются на интервалах $(f_л, f_п)$, $(-f_п, -f_л)$ (рис. 2.3). Действительно, составляющие левого спектра на частотах $-f_л$ через *m* периодов повторения попадут в точку $2f_л'-f_л<f_л'$, левее $f_л'$, а составляющие на часто-

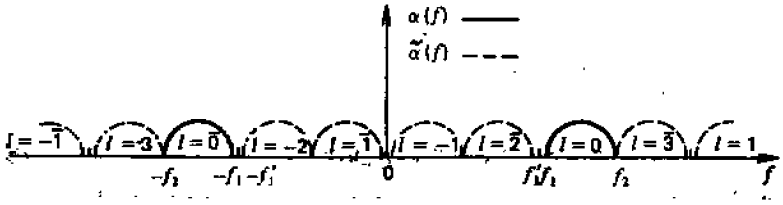


Рис. 2.3. К растрированию сигнала на пространственной несущей

тах–$f_п+\varepsilon$ через m+1 период – в точку $f_п+\varepsilon$, правее точки $f_п$. Спектру $\tilde{a}(f)$ (2.21) соответствует сигнал

$$\tilde{a}(x)=\frac{1}{F}\sum_{k=-\infty}^{\infty}a\left(\frac{k}{F}\right)\delta\left(x-\frac{k}{F}\right) \quad (2.22)$$

Поскольку

$$a(f)=\tilde{a}(f)\left[\mathrm{rect}\frac{f+f_п}{f_п-f_л}+\mathrm{rect}\frac{f-f_л}{f_п-f_л}\right],$$

то по теореме о свертке для преобразования Фурье

$$a(x)=\frac{1}{F}\sum_{k=-\infty}^{\infty}a\left(\frac{k}{F}\right)\lambda\left(x-\frac{k}{F}\right)$$

где $\lambda(x)=(f_п-f_л)\,\mathrm{sinc}\,[f_п-f_л)x]\cos(2\pi f_0 x)$, а $f_0=(f_п+f_л)/2$.

Итак

$$a(x)=\frac{f_п-f_л}{F}\cdot\sum_{k=-\infty}^{\infty}a\frac{k}{2F}\,\mathrm{sinc}\left[\pi(f_п-f_л)\left(x-\frac{k}{F}\right)\right]\times$$
$$\times\cos\left[2\pi f_0\left(x-\frac{k}{F}\right)\right] \quad (2.23)$$

Сравнивая представления (2.20) и (2.23), можно отметить, что в последнем приходится брать отсчеты сигнала примерно вдвое чаще, чем в первом, зато в первом нужно брать отсчеты и огибающей, и фазы, так что общее число отсчетов сигнала на единицу его протяженности остается равным *2FX.*

Формулы (2.13), (2.16), (2.20) и (2.23) записаны в одномерном представлении. При переходе к двумерному случаю необходимо учитывать, что если одномерный интервал – это отрезок, и граница его – две точки, то двумерный интервал – фигура, ограниченная, вообще

говоря, произвольной замкнутой линией. Как следует из предыдущего, замена непрерывного сигнала совокупностью его отсчетов, взятых на некотором растре, соответствует периодическому продолжению картины его спектра, причем для уменьшения количества отсчетов сигнала на единицу его площади необходимо стремиться к максимально плотной упаковке размноженных компонент спектра в плоскости пространственных частот.

Простейший способ периодического продолжения спектра – продолжение его в прямоугольной системе координат (рис. 2.4)

$$\tilde{\alpha}(f_1, f_2) = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} \alpha(f_1 + n_1 F_1, f_2 + n_2 F_2) \tag{2.24}$$

Для продолжения без перекрытия периоды повторения по координатам F_1 и F_2 должны превышать продольные (по направлению f_1) и поперечные (по направлению f_2) размеры фигуры S, ограничивающей спектр сигнала (рис. 2.4,а).

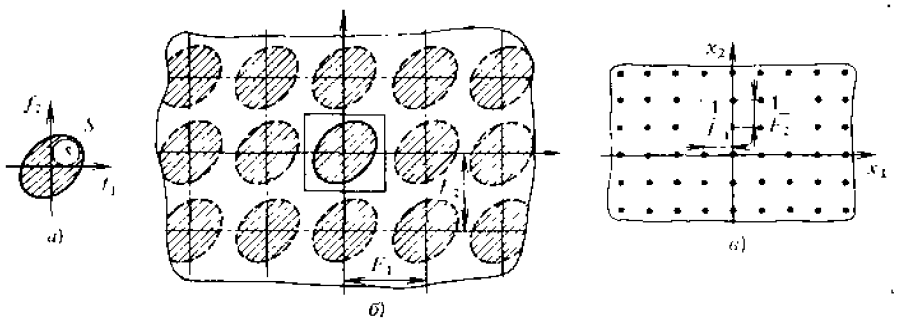


Рис. 2.4. Растрирование двумерного сигнала по прямоугольному растру:
а – фигура, ограничивающая спектр сигнала; б – периодическое продолжение спектра *а*
в прямоугольной системе координат; а – прямоугольный растр дискретизации

Спектру (2.24) соответствует растрирований непрерывный сигнал вида:

$$\tilde{a}(x_1, x_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} a\left(\frac{k_1}{F_1}, \frac{k_2}{F_2}\right) \delta\left(x_1 - \frac{k_1}{F_1}, x_2 - \frac{k_2}{F_2}\right), \tag{2.25}$$

т.е. отсчеты исходного непрерывного сигнала $a(x_1, x_2)$ берутся на прямоугольном растре (рис. 2.4, в) с шагом $\Delta x_1 = 1/F_1$ по координате x_1 и $\Delta x_2 = 1/F_2$ *по* координате x_2.

Очевидно, что восстановить исходный сигнал $a(x_1, x_2)$ из сигнала $\tilde{a}(x_1, x_2)$ можно, если пропустить последний через двумерный фильтр с частотной характеристикой

$$H_в(f_1, f_2) = \mathrm{rect}(f_1/F_1 + 1/2)\,\mathrm{rect}(f_2/F_2 + 1/2) \tag{2.26}$$

т.е. через фильтр, пропускающий только те частотные компоненты пространственного спектра, которые находятся внутри прямоугольника, показанного на рис. 2.4, *б* сплошной линией. По теореме о свертке для двумерного преобразования Фурье получим:

$$a(x_1, x_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} a(k_1/F_1, k_2/F_2)\,\mathrm{sinc}\,[2\pi F_1(x_1 - k_1/F_1)]\,\mathrm{sinc}\,[\pi F_2(x_2 - k_2/F_2)]. \tag{2.27}$$

Формула (2.27) выражает двумерную теорему отсчетов для случая растрирования изображений по прямоугольному растру. Количество отсчетов изображения на единицу его площади при такой дискретизации равно, очевидно, величине F_1F_2– площади прямоугольника, ограничивающего пространственный спектр изображения.

Более плотной, чем на рис. 2.4,6, упаковки спектров и соответственно менее плотного расположения отсчетов при дискретизации изображения со спектром, показанным на рис. 2.4, *а,* можно

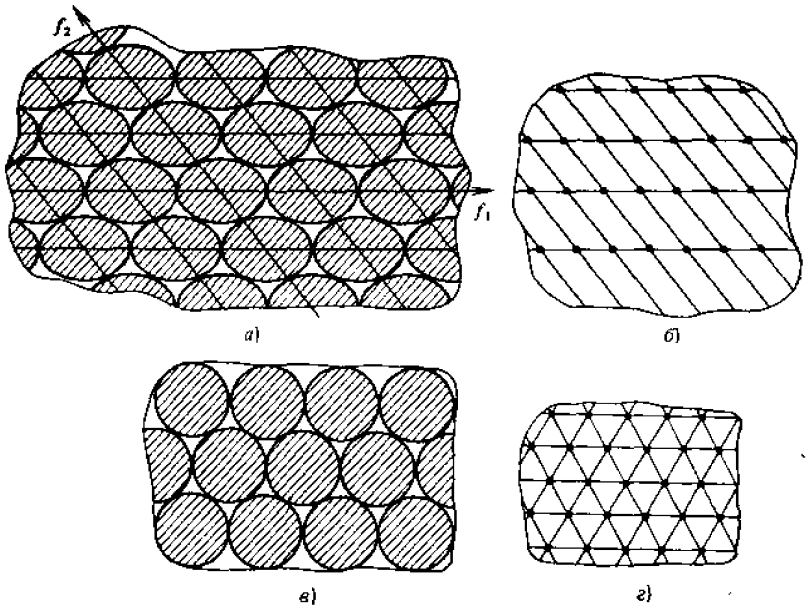


Рис. 2.5. Растрирование двумерного сигнала по косоугольному растру: *а* – периодическое продолжение спектра рис. *2.4,а в* косоугольной системе координат; *б* – форма растра дискретизации; в – периодическое продолжение изотропного спектра В гексагональной системе координат; *г* – гексагональный растр дискретизации

добиться, если периодическое продолжение спектра и соответственно дискретизацию производить в косоугольной системе координат, согласованной с формой кривой, ограничивающей спектр изображения (рис. 2.5,*а, б).* Если пограничной линией спектра является окружность, оптимальным будет расположение отсчетов изображения в узлах правильной шестиугольной решетки (рис. 2.5,8, *г).* При этом достигается экономия числа отсчетов по сравнению *с* дискретизацией по квадратному растру, равная $(2/\sqrt{3})$ – 1, или приблизительно 15%.

Выбор оптимальной формы растра – не единственный способ добиться плотной упаковки спектра при дискретизации. Другая возможность, которая также не имеет одномерного аналога – использование поворота системы координат. При этом мы приходим к своеобразной форме дискретизации, которую можно назвать *«дискретизацией с пропусками»* ([46]).

Рассмотрим спектр на рис. 2.6, а, отличный от нуля только на заштрихованном участке. Этот участок вписывается в квадрат. На рис. 2.6, б показана плотная упаковка спектральной плоскости, полученная наложением двух периодически повторенных в декартовой системе координат и повернутых на 90° друг относительно друга исходных спектров рис. 2.6, *а.* Каждому из периодически повторенных спектров соответствует дискретное изображение с расположением отсчетов в узлах квадратной решетки, как показано на рис. 2.6, в. Эти два изображения, повернутые, как и их спектры, на 90° друг относительно друга (допустим, вокруг точки 0, как на рис. 2.6, в) при наложении складываются.

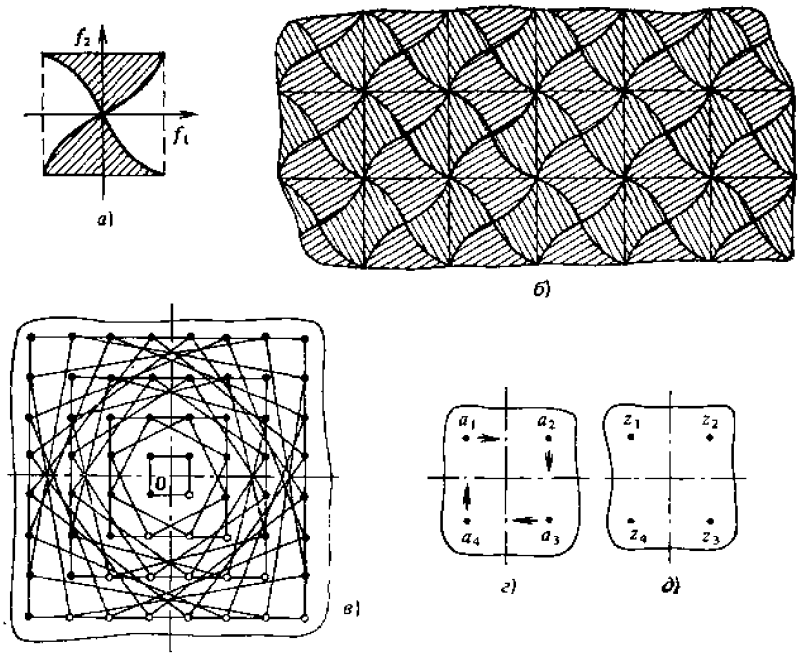


Рис. 2.6. Дискретизация «с пропусками»:
а– фигура, ограничивающая спектр сигнала (заштрихована); б –картина спектра, полученного после наложения повернутого на 90° сигнала и периодического продолжения в прямоугольной системе координат; в – растр дискретизации, соответствующий *б;* г –схема вращения отсчетов при повороте на 90°; *д* – результирующие отсчеты сигнала после сложения исходного сигнала с повернутым на 90°

В результате получается новый дискретный сигнал, отсчеты которого равны, очевидно, попарным суммам отсчетов исходного изображения. Способ образования этих пар при повороте растра относительно точки 0 (рис. 2.6, *в)* иллюстрируется рис. 2.6, *в–д.* На рис. 2.6, *в* линиями соединены узлы растра, которые при повороте на 90° накладываются друг на друга. На рисунке видно, что эти узлы образуют группы по четыре, в которых происходит циклическая перестановка узлов при повороте, как показано на рис. 2.6, *г* стрелками. Пусть z_1, z_2, z_3, z_4 – отсчеты в каждой группе, образованные в результате наложения отсчетов a_1, a_2, a_3, a_4 исходного изображения. Нетрудно заметить, что из четырех чисел (z_1, z_2, z_3, z_4) только три являются линейно-независимыми, четвертое же может быть получено из их линейной комбинации. Так, $z_4=z_1- z_2+z_3$.

Отсюда следует, что в каждой группе из четырех отсчетов, показанной на рис. 2.6, *в,* один отсчет является излишним: его можно вычислить, зная три других отсчета. Таким образом, за счет плотной упаковки поворотом экономится 25% отсчетов. Эти излишние отсчеты показаны на рис. 2.6, в незачерненными кружками, которые, как это видно из рисунка, заполняют собой нижнюю ¼ часть плоскости, как бы выпадающую из дискретизации. Процедура восстановления непрерывного изображения при такой «дискретизации с пропусками:», очевидно, должна состоять из двух этапов: восстановления пропущенных отсчетов и восстановления непрерывного сигнала пропусканием дискретного сигнала

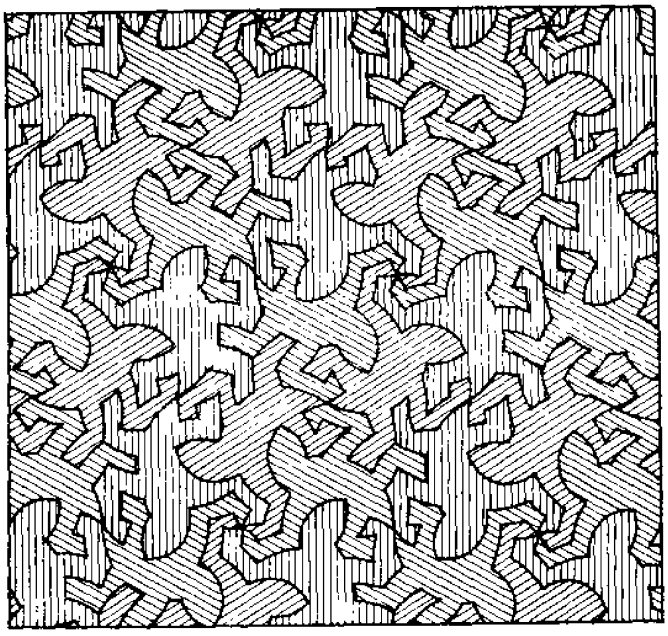


Рис. 2.7. Пример плотной упаковки сложного узора (картина М. Эшера)
через фильтр, подавляющий частотные компоненты спектра вне области, очерченной на рис. 2.6, *б* жирной линией.

Любопытный пример плотной упаковки на плоскости дает узор голландского художника Эшера, показанный на рис. 2.7. Он состоит из трех узоров, построенных периодическим продолжением фигуры ящерицы в гексагональной системе координат. Два узора повернуты на 60° по и против часовой стрелки относительно третьего и сдвинуты на полпериода по одной из координатных осей. В результате достигается плотная упаковка плоскости. Каждому из узоров, если рассматривать его как периодическое продолжение спектра некоторого сигнала, соответствует дискретный сигнал с расположением отсчетов в узлах гексагональной решетки, причем повернутым на 60° узорам соответствуют сигналы, отсчеты которых умножаются на комплексную экспоненту, зависящую от их координат (за счет сдвига в соответствии с теоремой сдвига). Можно показать, что при наложении двух повернутых на 60° гексагональных растров отсчеты распадаются на группы по шесть отсчетов, суммируемых попарно. В результате из каждых шести сумм одна оказывается излишней: ее можно вычислить по остальным пяти (аналогично предыдущему случаю квадратного растра). Таким образом, при плотной упаковке спектров экономится 1/3= (1/6+1/6) отсчетов исходного гексагонального растра. Из этих рассуждений видно, что сокращение числа отсчетов при плотной упаковке в спектральной области за счет поворотов не столь значительно, как при оптимальном выборе системы координат.

Можно указать еще более общую процедуру дискретизации, предполагающую нарушение топологии в спектральной области, при которой можно, в принципе, довести число отсчетов на единицу площади изображения до минимальной величины, равной площади пространственного спектра изображения. Эта процедура состоит в том, что фигура пространственного спектра изображения разбивается на фрагменты, из которых с помощью поворотов, сдвигов и зеркального отображения укладывается квадрат той же площади (рис,2.8).

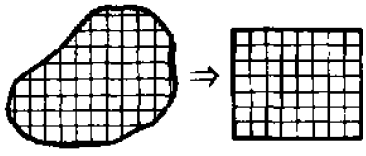


Рис. 2.8. Дискретизация «расфильтровкой»

Это соответствует выделению из изображения отдельных составляющих фильтрами, частотные характеристики которых постоянны в пределах площади каждого фрагмента и равны нулю вне ее, повороту этих составляющих, умножению на комплексную экспоненту сообразно сдвигу фрагмента и последующему •сложению во вспомогательное изображение, которое

может быть уже без потерь подвергнуто дискретизации на квадратном растре. Исходное изображение восстанавливается в обратной последовательности из восстановленного обычным образом по своим отсчетам вспомогательного изображения.

Такой способ дискретизация можно назвать *«дискретизацией с расфильтровкой».* Заметим, что как при «дискретизации с пропусками», так и при «дискретизации с расфильтровкой» получаемая дискретная последовательность величин не является уже последовательностью отсчетов сигнала, как при простом растрировании.

2.3. ОПТИМАЛЬНОЕ ПОЭЛЕМЕНТНОЕ КВАНТОВАНИЕ

Поэлементное квантование состоит в том, что в области значений сигнала выбирается отрезок конечной длины, который разбивается на *интервалы квантования,* и значения, попадающие в каждый интервал, обозначают одним числом – номером интервала. При восстановлении сигнала номер заменяется значением, являющимся представителем данного интервала. Способ разбиения на интервалы и значения-представители интервалов выбираются так, чтобы удовлетворялись требования к точности представления непрерывного сигнала цифровым.

Пусть α–коэффициент дискретного представления сигнала, $\hat{\alpha}^{(r)}$–значение-предста-

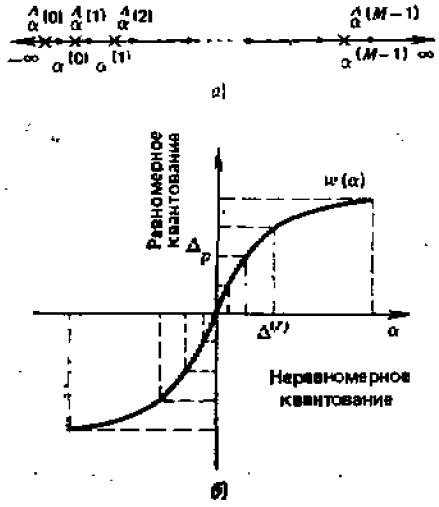


Рис. 2.9. Неравномерное квантование:
а – схема расположения границ н представителей интервалов квантования; *б* – неравномерное квантование с помощью нелинейного предыскажения

витель r-го интервала квантования области значений α (рис. 2.9). Ошибка квантования на r-м интервале может характеризоваться величиной

$$\varepsilon^{(r)} = \alpha - \hat{\alpha}^{(r)} \qquad (2.28)$$

Требования к точности квантования обычно формулируют в терминах ограничений, накладываемых на $\varepsilon^{(r)}$. Наиболее общий подход к формулировке этих ограничений состоит в том, что величину α, а значит и $\varepsilon^{(r)}$, считают случайной и вводят некоторую функцию потерь $D(\varepsilon^{(r)})$, связанных с отличием а от его квантованного представления $\alpha^{(r)}$. При таком подходе точность представления характеризуется средним по распределению вероятностей p(a) значением функции потерь $D(\varepsilon^{(r)})$:

$$Q = \sum_{r=0}^{M-1} \int_{\alpha^{(r)}}^{\alpha^{(r+1)}} p(\alpha)\, D(\varepsilon^{(r)})\, d\alpha \qquad (2.29)$$

где $\{\alpha^{(r)},\ \alpha^{(r+1)}\}$ - границы r-го интервала квантования; M – число интервалов, или уровней квантования.

Оптимальным считается такой выбор интервалов квантования и их представителей, при котором Q минимально и не превышает заданного граничного значения.

Здесь, однако, не учитывается, что на практике целесообразно различать два рода ошибок квантования: ошибки ограничения вследствие ограничения области значений квантуемой величины и ошибки квантования внутри выбранного ограниченного интервала

значений. Действительно, если плотность распределения $p(\alpha)$ имеет длинные «хвосты», ошибки ограничения могут достигать больших значений, тогда как ошибки квантования внутри ограниченного отрезка существенно ограничены. Функции распределения этих ошибок также сильно разнятся. Поэтому требования к точности квантованного представления непрерывной величины следует формулировать отдельно для

$$Q_{1\,гр} = \int_{-\infty}^{\alpha^1} p(\alpha)\, D_{1\,гр}(\varepsilon^{(0)})\, d\alpha$$
$$Q_{2\,гр} = \int_{\alpha^{(M-1)}}^{\infty} p(\alpha)\, D_{2\,гр}(\varepsilon^{(M-1)})\, d\alpha \qquad (2.30)$$

и для квантования внутри ограниченного отрезка

$$Q_0 = \sum_{r=1}^{M-2} \int_{\alpha^{(r)}}^{\alpha^{(r+1)}} p(\alpha)\, D_0(\varepsilon^{(r)})\, d\alpha \qquad (2.31)$$

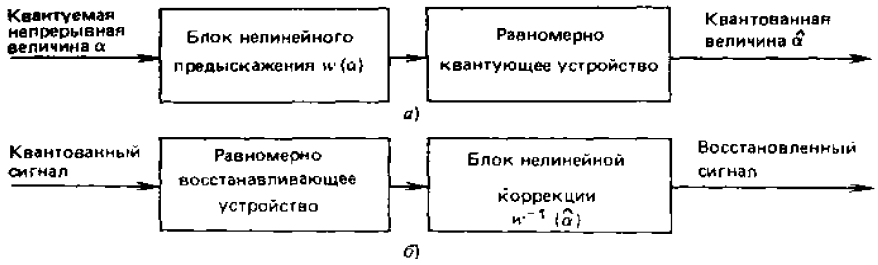
где $D_{1гр}(\varepsilon)$, $D_{2гр}(\varepsilon)$, $D_0(\varepsilon)$ – искажения, связанные с ошибками ограничения и ошибками квантования на ограниченном отрезке.

В такой постановке оптимальным является квантование, обеспечивающее минимум каждого из $Q_{1гр}(\varepsilon)$, $Q_{2гр}(\varepsilon)$, $Q_0(\varepsilon)$

Оптимальные значения границ $\alpha^{(0)}$, $\alpha^{(M-1)}$ отрезка, внутри которого производится квантование, границ интервалов квантования $\{\alpha^{(r)}\}$, r–*1, 2, ...,* M–2, а также значений представителей $\{\hat{\alpha}^{(r)}\}$ определяется, очевидно, системами уравнений, полученными из (2.30) и (2.31) дифференцированием по искомым величинам и приравниванием производных нулю. Решение этих равнений для нахождения $\{\alpha^{(r)}\}$ и $\{\hat{a}^{(r)}\}$ и минимального значения M может быть выполнено численными методами. В [61] приведены некоторые результаты таких расчетов для квадратичной функции потерь $D_0(\varepsilon)$ и гауссовского распределения $p(a)$. В [62] приведены результаты расчетов для критериев минимума четвертой и шестой степени ошибки квантования и функций распределения Гаусса, Рэлея и Лапласа.

Следует отметить, что задача поэлементного квантования возникает в цифровой обработке сигналов не только при преобразовании непрерывных сигналов в цифровые для ввода их в цифровой процессор, но также и на разных стадиях вычислений в цифровых процессорах при переходе от одной формы представления чисел к другой, например от формата чисел с плавающей запятой к байтовому формату для хранения их в архивных запоминающих устройствах. Описанный подход к оптимальному квантованию, который позволяет получить численное решение, с точки зрения его реализации лучше всего согласован именно с задачей квантования в цифровом процессоре. При выборе же оптимального квантования непрерывного сигнала для ввода в цифровой процессор удобнее несколько видоизменить постановку задачи, чтобы получить ее аналитическое решение, которое можно было бы воплотить в устройстве преобразования аналогового сигнала в цифровой.

Дело в том, что существующие устройства квантования обычно осуществляют равномерное квантование сигналов, при котором границы интервалов квантования размещаются равномерно в заданном диапазоне значений сигнала, а представители уровней квантования располагаются посередине между этими границами. При восстановлении непрерывного сигнала из квантованного также обычно используют цифроаналоговые преобразователи с равномерно квантованным входным сигналом. Используя такие устройства, оптимальное (неравномерное в общем случае) квантование можно обеспечить, если перед квантованием сигнал подвергнуть соответствующим образом выбранному нелинейному преобразованию (предыскажению) и соответственно при восстановлении непрерывного сигнала сигнал на выходе восстанавливающего устройства с равномерным квантованием подвергнуть соответствующей нелинейной коррекции (рис. 2.10). При квантовании с высокой точностью функция, описывающая вид коррекции, близка к обратной функции нелинейного предыскажения при квантовании ([46]). Для того чтобы найти оптимальную предыскажающую



Рис, 2.10: Схемы оптимального квантования (а) и восстановления (б) с нелинейным предыскажением и коррекцией

функцию, ошибку квантования удобно оценивать по средней величине функции потерь $D(\Delta_r)$ в пределах интервала квантования $\Delta_r = \alpha^{(r+1)} - \alpha^{(r)}$

$$Q_0 = \int_{\alpha^{(1)}}^{\alpha^{(M-1)}} p(\alpha) D(\Delta_r)\, d\alpha \qquad (2.32)$$

Пусть $w(\alpha)$ – функция, описывающая предыскажающее нелинейное преобразование. Это обычно гладкая монотонная функция, так что можно приближенно считать, что $\Delta_r \approx w'(\alpha)\Delta_p$, где Δ_p– ширина интервала равномерного квантования (см. рис. 2.9,6), $w'(\alpha)$ – производная $w(\alpha)$ по α. Тогда (2.32) можно переписать в виде:

$$Q_0 = \int_{\alpha^1}^{\alpha^{M-1}} p(\alpha) D(\Delta_p / w'(\alpha))\, d\alpha \qquad (2.33)$$

Тем самым задача оптимального квантования свелась к стандартной вариационной задаче минимизации интеграла (2.33) по функции $w(\alpha)$. Функция $w(\alpha)$, обеспечивающая минимум (2.33), определяется известным уравнением Эйлера – Лагранжа, которое в данном случае записывается как

$$\frac{\partial}{\partial w'}\{p(\alpha) D(\Delta_p / w'(\alpha))\} = \text{const} \qquad (2.34)$$

Разберем на нескольких примерах, как выбирается оптимальное расположение и число интервалов квантования.

Пример 1. Пороговый критерий. Пусть

$$D(\Delta_r) = \begin{cases} 0, & |\Delta_r| \leqslant \Delta_{\text{пор}}; \\ 1, & |\Delta_r| > \Delta_{\text{пор}}. \end{cases}$$

Такой критерий можно назвать пороговым, или критерием незаметности ошибки квантования. Для этого критерия решение задачи оптимального расположения интервалов квантования и выбор числа интервалов тривиальны: ширина r-го интервала квантования должна выбираться равной $2\Delta_{\text{пор}}$, а представителем r-го уровня – значение α в центре r-го интервала.

Если $\Delta_{\text{пор}}$ не зависит от α, мы получаем равномерную шкалу квантования. При этом число уровней (интервалов) квантования определяется соотношением

$$M = (\alpha_{\max} - \alpha_{\min})/2\Delta_{\text{пор}}$$

Важным и часто встречающимся на практике является случай, когда порог чувствительности к ошибкам квантования пропорционален значению квантуемой величины:

$$\Delta_{\text{пор}} = \delta_0 \alpha$$

т.е. когда ограничено значение не абсолютной, а относительной ошибки квантования. Так, в соответствии с психофизическим законом Вебера – Фехнера можно в первом приближении описать требования к точности квантования значений яркости изображения, предъявляемые зрительной системой человека. Нетрудно показать, что в этом случае необходимо подвергать равномерному квантованию не величину α, а ее нормированный логарифм:

$$\frac{w(\alpha) - w(\alpha_{\min})}{w(\alpha_{\max}) - w(\alpha_{\min})} = \frac{\ln(\alpha/\alpha_{\min})}{\ln q} \qquad (2\text{-}35)$$

где $q = \alpha_{\max}/\alpha_{\min}$

Число уровней квантования по такой логарифмической шкале должно быть при этом равно

$$M = (\ln q)/\delta_0 \qquad (2.36)$$

Так на практике и поступают: перед равномерным квантованием сигнал подвергают логарифмическому предыскажению (компрессии), значения представителей интервалов квантования при восстановлении выбирают по равномерной шкале, а затем синтезированный сигнал подвергают потенцированию (экспандированию).

Найдем количество уровней, требуемое при квантовании яркости изображений. Данные психофизиологических измерений показывают, что в обычных условиях освещения порог относительной контрастной чувствительности зренияδ_0 оценивается величиной 1,5–2% для тестового пятна большой площади при длительной адаптации к фону. Существующие устройства воспроизведения изображений могут обеспечить динамический диапазон яркостей ?«ЮО. Подставив эти величины в (2.36), получим М≈230. Прямые эксперименты по квантованию яркости с использованием телевизионных устройств воспроизведения изображений показали, что достаточно 64–128 уровней. Это снижение требований отчасти связано с наличием собственных шумов датчиков видеосигнала и устройств восстановления изображений. В настоящее время в устройствах квантования видеосигнала и восстановления изображений принято 64–256 уровней при логарифмическом предыскажении квантуемого видеосигнала.

Интересно оценить выигрыш в числе уровней квантования, даваемый логарифмическим предыскажением, по сравнению с равномерным квантованием при той же точности. Он, очевидно, равен

$$g = (q - 1)/\ln q$$

При больших q выигрыш может быть достаточно велик. Так, при *g-100* g≈20. Однако практически выигрыш обычно не столь значителен, поскольку оценка погрешности по «наихудшему» случаю, как это делается в пороговом критерии, излишне строга.

Пример 2. Степенной критерий абсолютной ошибки квантования. Пусть

$$D(\Delta_r) = (\Delta_r)^{2n} \qquad (2.37)$$

Подставив (2.37) в (2.34) и решив получившееся дифференциальное уравнение, найдем:

$$\frac{w(\alpha) - w(\alpha_{\min})}{w(\alpha_{\max}) - w(\alpha_{\min})} = \frac{\int_{\alpha_{\min}}^{\alpha} (p(\alpha))^{1/(2n+1)} d\alpha}{\int_{\alpha_{\min}}^{\alpha_{\max}} (p(\alpha))^{1/(2n+1)} d\alpha}$$

Таким образом, требуемое нелинейное предыскажение зависит только от распределения вероятностей квантуемой величины. Смысл этой зависимости очевиден из соотношения

$$\Delta_r \approx \Delta_p / w'(\alpha) \sim (p(\alpha))^{-1/(2n+1)}$$

т.е. интервалы квантования значений α обратно пропорциональны плотности вероятностей этих значений в соответствующей степени.

Для распространенного случая оценки по среднеквадратическому значению ошибки квантования *(n=1):*

$$\frac{w(\alpha) - w(\alpha_{\min})}{w(\alpha_{\max}) - w(\alpha_{\min})} = \frac{\int_{\alpha_{\min}}^{\alpha} \sqrt[3]{p(\alpha)}\, d\alpha}{\int_{\alpha_{\min}}^{\alpha_{\max}} \sqrt[3]{p(\alpha)}\, d\alpha}$$

В некоторых случаях, например при квантовании спектральных коэффициентов сигналов в базисах Фурье, Уолша и др., можно приближенно считать, что квантуемые коэффициенты дискретного представления сигнала имеют усеченную гауссовскую плотность распределения вероятностей в интервале ($\alpha_{\min}$, $\alpha_{\max}$):

$$p(\alpha) = c \exp\left[-(\alpha - \alpha_0)^2 / 2\sigma_\alpha^2\right]$$

где c – константа нормировки. Тогда при n=1

$$\frac{w(\alpha) - w(\alpha_{\min})}{w(\alpha_{\max}) - w(\alpha_{\min})} = \frac{\Phi\left(\frac{\alpha - \alpha_0}{\sqrt{3}\,\sigma_\alpha}\right) - \Phi\left(\frac{\alpha_{\min} - \alpha_0}{\sqrt{3}\,\sigma_\alpha}\right)}{\Phi\left(\frac{\alpha_{\max} - \alpha_0}{\sqrt{3}\,\sigma_\alpha}\right) - \Phi\left(\frac{\alpha_{\min} - \alpha_0}{\sqrt{3}\,\sigma_\alpha}\right)}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt$$

где

Выигрыш в числе уровней квантования при заданной точности по сравнению с равномерным квантованием зависит в этих случаях от параметра

$$q = (\alpha_{max} - \alpha_{min})/\sigma_\alpha$$

и равен

$$g = \frac{q}{\sqrt{2\pi}\sqrt[4]{27}} \frac{(\Phi(q/2) - \Phi(-q/2))^{1/2}}{(\Phi(q/2\sqrt{3}) - \Phi(-q/2\sqrt{3}))^{3/2}}$$

При достаточно больших q эта величина примерно равна $q/5,7$. Таким образом, выигрыш становится ощутимым только при очень больших q, когда начинают сказываться «хвосты» распределения $p(\alpha)$. Эксперименты с квантованием ортогональных компонент ноля голограмм диффузных объектов (см. § 2.4} показывают, что для высокого качества восстановления таких голограмм ограничение динамического диапазона значений ортогональных компонент поля голограммы должно производиться на уровне не менее $\pm 4\sigma$, что соответствует $q=8$. При этом выигрыш в числе уровней квантования за счет оптимального квантования будет составлять приблизительно 40%.

Пример 3. Степенной критерий относительной ошибки квантования. Пусть

$$D(\Delta_r) = (\Delta_r/\alpha)^{2n}$$

Из уравнения Эйлера – Лагранжа (2.33) для этого случая нетрудно получить:

$$\frac{w(\alpha) - w(\alpha_{min})}{w(\alpha_{max}) - w(\alpha_{min})} = \frac{\int_{\alpha_{min}}^{\alpha} [p(\alpha)/\alpha]^{1/(2n+1)} d\alpha}{\int_{\alpha_{min}}^{\alpha_{max}} [p(\alpha)/\alpha]^{1/(2n+1)} d\alpha}$$

если величина а имеет равномерное распределение, то

$$\frac{w(\alpha) - w(\alpha_{min})}{w(\alpha_{max}) - w(\alpha_{min})} = \frac{\alpha^{2n/(2n+1)} - \alpha_{min}^{2n/(2n+1)}}{\alpha_{max}^{2n/(2n+1)} - \alpha_{min}^{2n/(2n+1)}}$$

При оценке точности квантования по среднему значению квадрата модуля относительной ошибки (n=½)

$$\frac{w(\alpha) - w(\alpha_{min})}{w(\alpha_{max}) - w(\alpha_{min})} = \frac{\sqrt{\alpha} - \sqrt{\alpha_{min}}}{\sqrt{\alpha_{max}} - \sqrt{\alpha_{min}}}. \qquad (2.38)$$

Выигрыш в числе уровней квантования, обеспечиваемый нелинейным предыскажением по (2.38) по сравнению с равномерным квантованием, равен

$$g = (\sqrt{q} - 1)(\ln q)/4(\sqrt{q} - 1)$$

Таким образом, при равномерном распределении $p(\alpha)$ выигрыш невелик. При q=100 он равен 1,4. Сравнивая этот результат с величиной g=20, полученной для порогового критерия относительной ошибки, мы видим, что здесь выигрыш за счет нелинейного расположения уровней квантования значительно скромнее, чем при оценке по наихудшему случаю.

2.4. ПРАКТИЧЕСКИЕ ВОПРОСЫ РАСТРИРОВАНИЯ И КВАНТОВАНИЯ ИЗОБРАЖЕНИЙ, ГОЛОГРАММ И ИНТЕРФЕРОГРАММ

Разновидности методов дискретизации. Теоремы отсчетов определяют правило растрирования сигналов с ограниченным спектром. Для того чтобы использовать их для дискретизации оптических сигналов {например, изображений, голограмм, интерферо-грамм) для выбора растра дискретизации и оценки искажений вследствие дискретизации, необходимо знать форму и размеры фигуры, ограничивающей фурье-спектр сигналов. Эти данные иногда непосредственно связаны с известными конструктивными характеристиками соответствующих систем (скажем, с характеристиками объективов и фотопленки для фотографических систем), и их можно оценить, зная эти характеристики. Но во многих случаях эти данные не известны.

Кроме того, не всегда удобно описывать искажения таких сигналов искажением их спектров, как это удобно делать, если пользоваться -теоремами отсчетов.

В этих случаях параметры растрирования – форму и шаг растра – выбирают, исходя из других, более простых критериев оценки ошибки растрирования и задаваясь конкретным способом интерполяции отсчетов при восстановлении непрерывного сигнала'. Одним из наиболее простых критериев оценки ошибки является максимальное значение отличия сигнала от результата его восстановления по дискретному представлению. Например, учитывая реальные характеристики устройств воспроизведения изображений, считают, что при восстановлении происходит ступенчатая интерполяция его отсчетов (при квадратной апертуре). Поэтому задаются квадратным растром, а расстояние между отсчетами выбирают так, чтобы на этом расстоянии сигнал не мог измениться больше, чем на заданную величину ошибки. Для этого нужно, конечно, задаться ограничениями на возможную скорость изменения сигнала изображения по координатам, скажем на максимальное значение его производных.

Наконец, в некоторых случаях решение о выборе параметров растрирования изображений приходится принимать, основываясь на данных о требуемом количестве отсчетов на объект минимальной площади или других подобных показателях, полученных эмпирическим путем.

В описанных способах растрирования оптических сигналов дискретным представлением служат отсчеты сигнала, что, как уже отмечалось, соответствует представлению сигнала по отсчетным и дельта-функциям, В последнее время при преобразовании излучения в цифровой сигнал находят применение методы представления сигналов по иным базисным функциям. При использовании этих методов измеряемое поле излучения пропускают через сменные кодирующие-маски, функция пропускания которых соответствует каждой базисной функции, и затем измеряют энергию излучения за маской, т.е. коэффициент представления поля излучения по данной базисной функции. Удобнее всего использовать бинарные маски (прозрачные или непрозрачные), соответствующие базисным функциям с двоичными значениями (например, функциям Уолша). Такой способ дискретизации получил название *мультиплексного кодирования,* или метода кодированных апертур. Он находит применение в устройствах измерения и дискретизации слабых излучений (радиоактивного, рентгеновского, инфракрасного и т.п.) для увеличения чувствительности датчиков. Заметим, что дискретизация измерением отсчетов (растрированием) может рассматриваться как частный случай такого метода, когда каждая маска – это непрозрачная для излучения пластина с малым отверстием, координаты которого меняются от маски к маске в соответствии с выбранным растром.

Квантование с учетом шума датчика сигнала**.** При определении требований к способу квантования непрерывных оптических сигналов необходимо учитывать, что реальные датчики (фотопленка, фотоумножители, передающие телевизионные трубки, кристаллические свето-, тепло-, рентгеночувствительные датчики и т.п.) непрерывного сигнала и устройства воспроизведения оптических сигналов (люминофор электронно-лучевых трубок, модулируемые источники света, фоточувствительные материалы) обладают собственными шумами, в результате чего отсутствует абсолютно точное соответствие между сигналом и объектом изучения. При расчете оптимального нелинейного предыскажения при квантовании можно считать, что этот шум, пересчитанный на вход равномерно квантующего устройства, аддитивно складывается с шумом квантования. Таким образом можно учесть возможную зависимость интенсивности шума датчиков и синтезаторов от уровня сигнала (так, дисперсия шума фотоэлектронных умножителей пропорциональна величине сигнала). Критерий точности квантования при этом должен быть сформулирован с учетом совместного действия шума датчика (синтезатора) и шума квантования. Важной особенностью взаимодействия этих двух видов искажений является рандомизация шума квантования, разрушение его корреляционных связей с квантуемым сигналом. В результате требования к допустимой величине шума квантования могут быть несколько снижены. Так, случайный шум датчика видеосигнала разрушает ложные контуры при грубом квантовании яркости изображения, уменьшая тем самым их заметность. На этом основан даже один из способов кодирования изображений [68], заключающийся в том, что к видеосигналу перед грубым квантованием добавляется псев-дослучайный шум с независимыми отсчетами, а при восстановлении изображений этот шум вычитается из квантованного видеосигнала. Очевидно, что благотворное действие случайного

шума датчика сигнала и синтезатора на шумы квантования сказывается только до определенной степени. Существует некоторое оптимальное соотношение между этими видами шумов, зависящее от свойств квантуемого сигнала и содержания решаемой задачи. В большинстве случаев случайный шум датчика и синтезатора должен иметь примерно ту же интенсивность (дисперсию), что и шум квантования. В работе [9] на основании расчетов по трем разным критериям рекомендуется, чтобы среднеквадратичное значение случайного шума было примерно в 3 раза меньше ширины интервала квантования (т.е. отношение дисперсий равно 4/3, если считать шум квантования равномерно распределенным в интервале квантования).

Квантование при обработке сигналов в цифровом процессоре.

Для повышения точности вычислений в цифровых процессорах, как правило, на представление чисел отводится гораздо больше двоичных разрядов, чем то, которое используется для ввода аналогового сигнала. Например, для ввода в ЦВМ отсчетов видеосигнала отводится 8 разрядов (один байт), что соответствует 256 уровням квантования, а при обработке в цифровой ЭВМ числа представляются в виде «целых» чисел (обычно 16 разрядов) или чисел с фиксированной или плавающей точкой (обычно 32 разряда). Задача перевода чисел из форматов целых чисел, чисел, с фиксированной и плавающей точкой в байтовый формат для вывода их на устройства визуализации и обратного перевода из байтового формата в форматы *с* большей разрядностью – типичная задача квантования. Поэтому в математическом обеспечении цифровых систем обработки аналоговых оптических сигналов необходимо иметь стандартные программные блоки, реализующие процедуры оптимального квантования – восстановления, показанные на рис. 2.10: определение максимального и минимального значений массива чисел, нелинейного предыскажения со «срезкой» чисел по минимуму и максимуму (заменой чисел, меньших минимума, минимальным и, больших максимума, максимальным), равномерного квантования, нелинейной коррекции. При выборе вида нелинейного предыскажения и коррекции обычно целесообразно основываться на пороговых критериях точности квантования.

Квантование и задача цифровой коррекции нелинейных искажений сигналов. Одна из типичных задач цифровой обработки аналоговых и, в частности, оптических сигналов – это задача коррекции нелинейных искажений сигналов. Например, это искажения, вызванные нелинейностью характеристической кривой фотографических материалов, зависимости свет – сигнал фотоумножителей и телевизионных передающих трубок и т.д. Эти искажения можно, как правило, рассматривать как поэлементное нелинейное преобразование отсчетов сигнала. Задача коррекции таких искажений состоит в том, чтобы найти и выполнить в цифровом процессоре корректирующее преобразование квантованных искаженных отсчетов сигнала. Поскольку корректируемые значения сигнала квантованы, корректирующее преобразование может быть описано таблицей, в которой по значению входной квантованной величины можно найти ее скорректированное квантованное значение. При построении этой таблицы необходимо учитывать, что корректируемый сигнал перед коррекцией в цифровой системе и после нее подвергается дополнительным нелинейным преобразованиям: нелинейному предыскажению при вводе сигнала в цифровой процессор, квантованию и нелинейной коррекции перед восстановлением сигнала на выходе цифрового процессора.

Пусть a – значения неискаженного сигнала; $w_н(a)$ –функция, описывающая нелинейное искажение; $b = w_н(a)$– искаженные значения сигнала; $\tilde{b}$–значения сигнала после его нелинейного преобразования на входе квантователя; r – номер интервала равно-

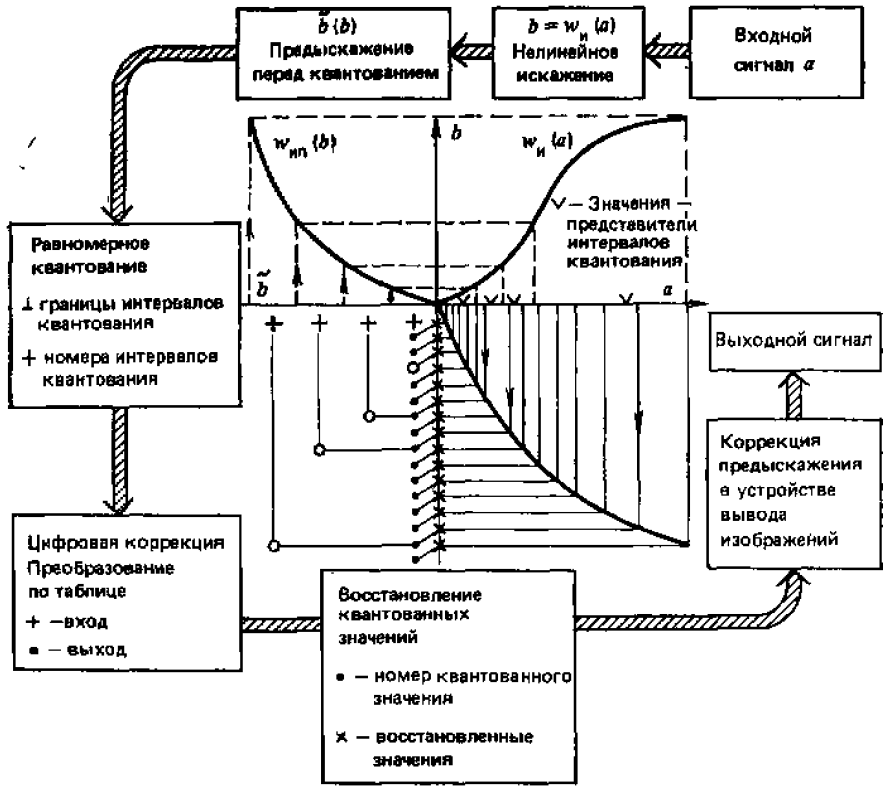


Рис. 2.11. Последовательность операций при цифровой коррекции нелинейных искажений сигнала

мерного квантования величины $\tilde{b}$; $\{b_r\}$–границы интервалов; p – номер интервала квантования выходного сигнала; $\hat{b}_p$– соответствующие значения-представители интервалов квантования выходного сигнала.

Задача оптимальной цифровой коррекции – свести к минимуму отличие скорректированного сигнала $\{\hat{a}_p\}$ от неискаженного. Эта задача родственна задаче оптимального квантования, рассмотренной в § 2.3. В соответствии с результатами этого параграфа решением является следующий метод коррекции нелинейности, схематически показанный на рис. 2.11.

1. По заданной устройством квантования шкале квантования {&,} и заданным предыскажающей $w_{нп}(b)$ и искажающей функциям $w_u(a)$ определяются границы $\{a_r\}$ интервалов квантования сигнала до искажения:

$$a_r = w_н^{-1}(w_{нп}^{-1}(b_r))$$

2. Для каждого r-го интервала квантования находится опти-

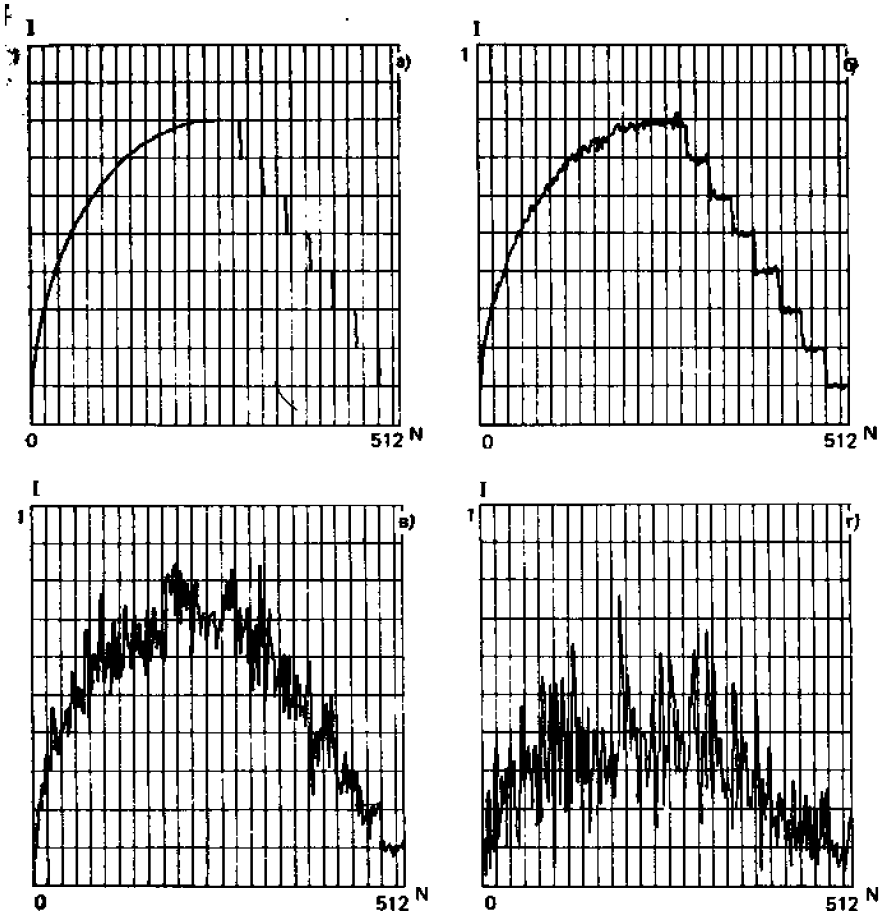
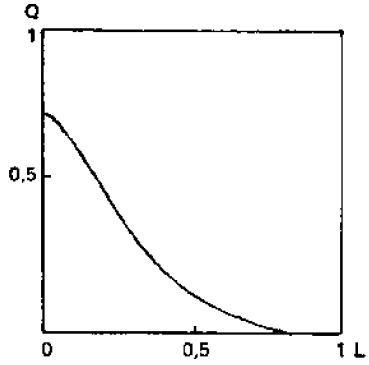


Рис. 2.12. Влияние ограничения динамического диапазона ортогональных компонент поля голограммы диффузного объекта:
а – исходное распределение амплитуды поля *I* как функции пространственной координаты *N; б* – распределение при ограничении ортогональных компонент сигнала голограммы на уровне ±3σ; *в* – на уровне ±2σ; *г* – на уровне ±σ (σ – среднеквадратические значения компонент сигнала).

мальное значение $\hat{a}_r$ представителя этого интервала, обеспечивающего минимальную ошибку квантования.

3. По заданной устройством воспроизведения функции $w_k(\tilde{b})$ коррекции предыскажения для каждого r-го представителя определяется номер *p* представителя $\hat{b}_p$ интервала квантования непрерывной величины, восстанавливаемой из квантованных значений $\{\tilde{b}\}$. Полученная таблица *p(r)* является требуемой таблицей коррекции.

Квантование голограмм. Рассмотренный в § 2.3 способ оценки погрешности поэлементного квантования и связанный с ним метод выбора оптимального квантования удовлетворительно соответству-



Рис, 2.13. Зависимость спекл-контраста Q (отношения стандартного отклонения шума к интенсивности сигнала) от глубины ограничения *L* значений ортогональных компонент сигнала голограммы

ет задаче квантования отсчетов сигнала изображений, но не может быть непосредственно использован для оценки погрешности квантования таких оптических сигналов, как голограммы или интерферограммы.

Погрешность квантования сигнала голограммы естественно оценивать не по искажениям голограммы, а по искажениям восстанавливаемого с такой искаженной голограммы изображения.

Изучение эффектов квантования голограмм показывает [49], что квантование по-разному сказывается на голограммах «зеркальных» и «диффузных» объектов («Зеркальными» будем называть объекты с сильно коррелированной шероховатостью поверхности; если поверхность такого объекта плоская, то он рассеивает падающий на него свет преимущественно в одном направлении, подобно зеркалу. «Диффузными» будем называть объекты, рассеивающие свет в достаточно широком телесном угле. Фурье-спектр «зеркальных» объектов имеет ярко выраженную пространственную неоднородность. Фурье-спектр «диффузных» объектов пространственно практически однороден.). Ограничение диапазона значений и квантование голограмм «зеркальных» объектов приводит к разрушению макроформы объекта, в частности восстановленные изображения становятся контурными. Правильным выбором неравномерной шкалы квантования эти искажения удается существенно уменьшить [49].

Голограммы «диффузных» объектов более устойчивы к ограничению диапазона значений и квантованию. Эти искажения не приводят к полному разрушению восстановленного изображения, а сказываются в появлении на изображении случайного шума, называемого шумом диффузности или «спекл-шумом».

На рис. 2.12 представлены результаты моделирования влияния ограничения максимального и минимального значений при записи ортогональных компонент голограммы диффузного объекта. На этих рисунках хорошо заметно появление и увеличение шума диффузности с увеличением глубины ограничения и то, что макроструктура объекта сохраняется.

Для количественной оценки шума на рис. 2.13 приведен график зависимости интенсивности шума диффузности от глубины ограничения ортогональных компонент поля голограммы.

Аналогичная закономерность получается при квантовании ортогональных компонент поля голограммы: уменьшение числа уровней квантования голограммы диффузного объекта приводит к увеличению шума диффузности, но макроструктура объекта сохраняется (рис. 2.14). Отметим характер зависимости спекл-контраста

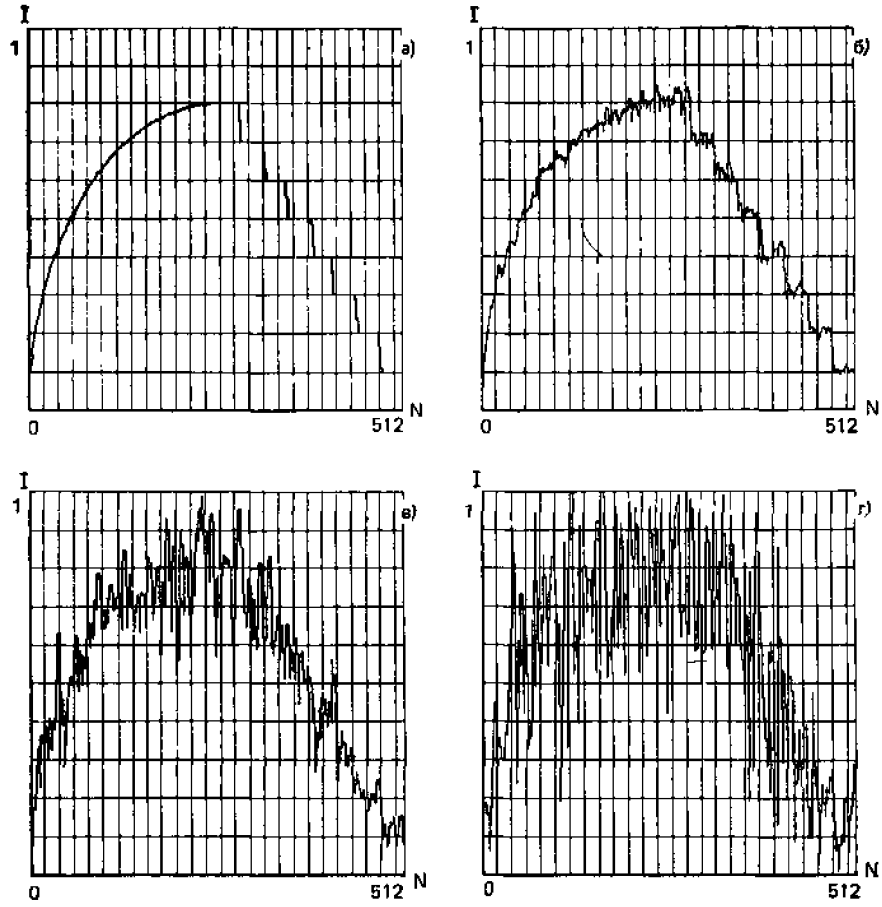


Рис. 2.14. Влияние квантования значений ортогональных компонент поля голограммы диффузного объекта:
а – исходное распределение амплитуды поля объекта / как функции пространственной координаты *N; б* – восстановленное распределение при равномерном квантовании на 128 уровней; а – на 64 уровня; *г* – на 32 уровня

от числа уровней квантования голограммы (рис. 2.15), показывающий, что при уменьшении числа уровней квантования относительная интенсивность шума сначала растет сравнительно медленно, а примерно после 32 уровней – быстро. Это говорит о том, что для высокой точности воспроизведения количество уровней квантования голограмм должно быть порядка 128–256 при ограничении диапазона значений на уровне (3,5–4)σ. Как показывают результаты предыдущего параграфа, использование нелинейной шкалы квантования позволяет в этом случае уменьшить число уровней квантования на 20–40%.

Устойчивость голограмм диффузных объектов к ограничению и квантованию значений используют при синтезе голограмм, применяя имитацию диффузной подсветки объектов.

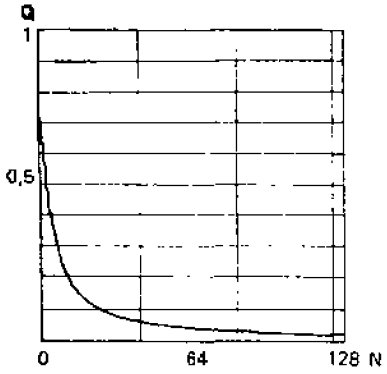


Рис. 2.15. Зависимость спекл-контраста Q от числа уровней равномерного квантования значений ортогональных компонент поля голограммы

Это своеобразный аналог описанного выше метода квантования с добавлением псевдослучайного шума [68]. Однако это не лучший способ добиться устойчивости голограммы к квантованию. В ряде работ (например, [16]) предложено пользоваться так называемыми «регулярными» диффузорами, которые давали бы тот же эффект «размазывания» информации

по площади голограммы, что и случайный диффузор, но не давали случайного шумового узора *на* восстановленном изображении. Идея «регулярного» диффузора может найти наиболее полное воплощение при цифровом синтезе голограмм, поскольку здесь меньше всего ограничений, связанных с проблемой его физической реализации.

В [49] описан легко реализуемый при цифровом синтезе голограмм способ мультиплицирования. Синтезированная голограмма разбивается по площади на несколько участков с различной интенсивностью сигнала. Затем сигнал на центральном, обычно наиболее интенсивном участке, ослабляется в каждой точке во столько раз, во сколько максимум сигнала на этом участке больше максимума сигнала на соседнем, менее интенсивном участке. Этот ослабленный участок повторяется столько же раз на площади голограммы и суммируется с сигналом на соседних участках голограммы. Эта процедура может повторяться многократно, после чего получается мультиплицированная цифровая голограмма с гораздо более узким динамическим диапазоном значений. Она и подвергается квантованию для вывода на устройство записи голограмм.

Эксперименты по мультиплицированию голограмм [49] показали, что при надлежащем наборе параметров мультиплицирования (числа и размеров мультиплицируемых фрагментов голограмм) этот способ дает хорошие результаты.

2.5. ОБЗОР МЕТОДОВ КОДИРОВАНИЯ ИЗОБРАЖЕНИЙ

Кодирование изображений – это преобразование изображений в цифровой сигнал, т.е. последовательность чисел для хранения изображений в цифровых запоминающих устройствах и передачи в цифровых системах связи. Кодирование выполняется тем эффективнее, чем меньше объем цифрового сигнала, требуемого для описания изображения с заданной точностью. Объем цифрового сигнала характеризуется количеством двоичных единиц на элемент (отсчет) изображения. Считается, что для высококачественного воспроизведения изображения требуется, чтобы на 1 элемент изображения приходилось не менее 8 дв. ед. (бит.). С помощью специальных методов кодирования удается уменьшить эту величину в несколько раз и обеспечить приемлемое качество воспроизведения изображений из цифрового сигнала объемом до 1 дв. ед. на элемент изображения. Соответственно уменьшается требуемая емкость запоминающих устройств для хранения изображений и требуемая пропускная способность канала для передачи изображений.

В настоящее время существует большое разнообразие методов кодирования изображений, и они достаточно широко освещены в литературе. Практически все известные методы кодирования могут быть упорядочены в схему, показанную на рис. 2.16.

Большинство из них предполагают трехступенчатую процедуру кодирования: раздельную дискретизацию, квантование отсчетов полученного дискретного представления и последующее статистическое кодирование.

Дискретизация производится обычно как разложение (преобразование) сигнала по некоторым базисным функциям Преобразованию может подвергаться сразу все изображение (кадр) или фрагменты изображения (блоки). В качестве базисных функций выбирают как классические ортогональные семейства (базисные функции преобразований Фурье, Уолша, Хаара), так и другие, специально созданные для использования в кодировании изображений, такие как «слэнт»-преобразование, косинусное преобразование, линейный ортогональный базис и другие, а также неортогональные линейно-независимые базисы, используемые в так называемом групповом кодировании [41] и кодировании путем *декорреляции предсказанием.* При *групповом кодировании,* квантованию подвергаются средние значения, вычисляемые по блокам изображения, и поэлементные разности значений элементов и этих средних значений. При декорреляции предсказанием квантованию подвергаются разности между значениями элементов изображения и значениями, предсказанными по соседним ближайшим к ним элементам путем суммирования их значений с весами.

Блочные преобразования, как правило, более эффективны в отношении кодирования, чем преобразования всего кадра. Оптимальный размер блоков при двумерных преобразованиях – от нескольких десятков до нескольких сотен элементов (отсчетов) изображения на блок. Типичные размеры блоков: 4x4; 8x8; 16x16 элементов.

Из рекомендуемых для кодирования преобразований особо следует отметить как наиболее эффективные косинусное и «слэнт»-преобразование. Для использования в специализированных кодирующих устройствах, работающих в темпе развертки изображения, наиболее удобны скользящие преобразования (преобразования по сдвиговым базисам). При этом линейное преобразование сводится к декорреляции видеосигнала предсказанием. Предсказание может быть как одномерным, когда для предсказания используются только прошлые значения, расположенные на одной строке развертки с текущим элементом, и двумерным, при котором используются также значения сигнала на предыдущих строках развертки. Метод декорреляции предсказанием в сочетании с соответствующим квантованием декор рели ров а иного сигнала называют *дифференциальной импульсно-кодовой модуляцией (ДИКМ).*

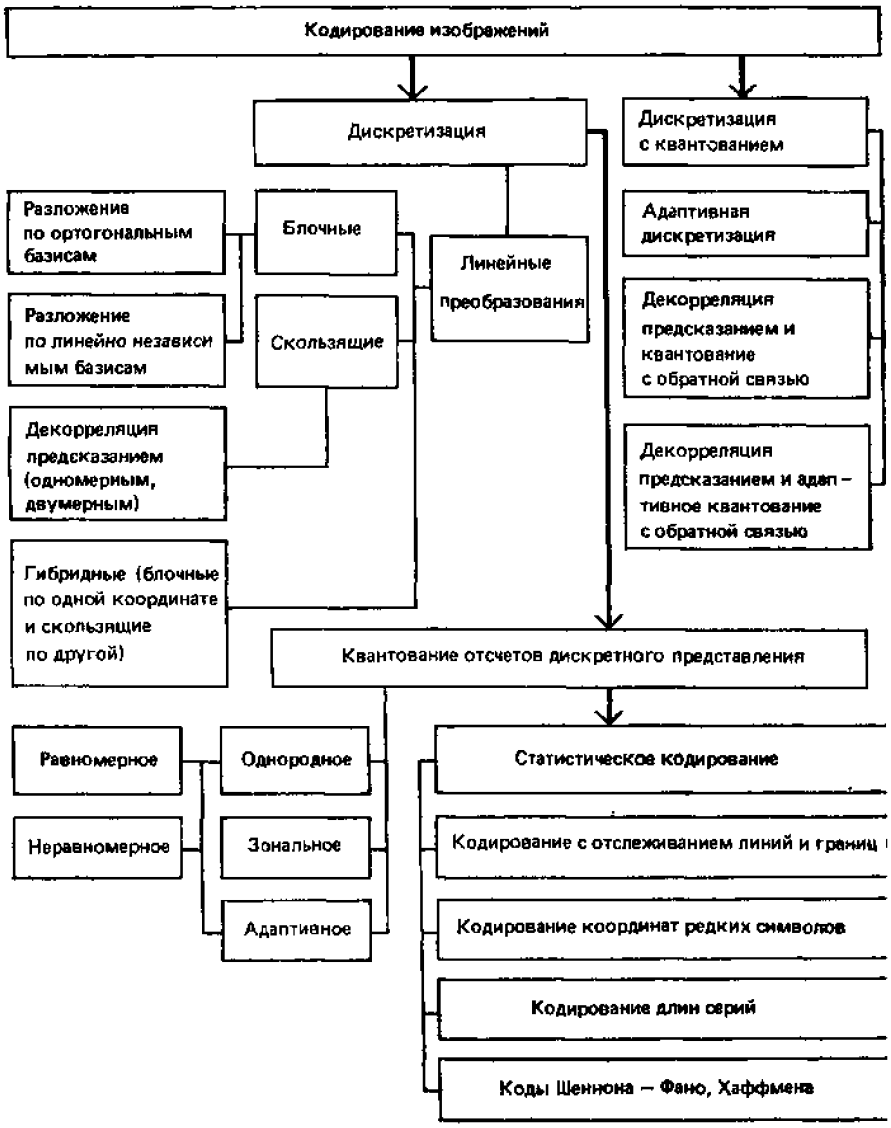


Рис. 2.16. Классификация методов кодирования изображений

В параллельно-последовательных (параллельных по одной координате и с разверткой по другой) системах считывания и передачи изображений удобно использовать *гибридный метод преобразования* – блочное преобразование по координате с параллельным считыванием и преобразование декорреляции предсказанием по координате, вдоль которой осуществляется развертка.

После получения отсчетов изображения в результате его дискретизации отсчеты подвергаются квантованию – замене их значений фиксированными заранее и пронумерованными значениями (уровнями квантования). Простейший' метод – *однородное квантование,* когда правило квантования (количество и расположение уровней квантования по шкалам значений сигнала) одинаково для всех отсчетов, полученных в результате дискретизации. Однако этот способ квантования наименее эффективен. Он, как правило,

применяется при так называемом первичном квантовании при формировании цифрового представления без сокращения его избыточности. При этом обычно достаточно 256 уровней квантования, расположенных по логарифмической шкале, и квантованию подвергаются непосредственно отсчеты самого сигнала изображения.

При кодировании с преобразованиями используются, как правило, *неоднородное квантование* – зональное и адаптивное.

Однородное и зональное квантование может быть равномерным (с равномерным расположением уровней квантования) и неравномерным. В последнем случае количество уровней квантования можно уменьшить, а эффективность кодирования соответственно повысить при сохранении качества восстановленного--после декодирования изображения.

В некоторых случаях оказывается удобным и возможным (с точки зрения реализации кодирующих устройств) построить смешанную процедуру дискретизации и квантования. Таковы адаптивная дискретизация растрированием и декорреляция предсказанием и квантование с обратной связью.

При адаптивной дискретизации растрированием расположение отсчетов сигнала определяется результатом квантования (как правило, на 2 уровня) ошибки восстановления непрерывного сигнала по значению ближайшего отсчета, как в упоминавшемся способе представления двумерных сигналов линиями равных-значений.

Заключительным этапом является *статистическое кодирование* результатов квантования отсчетов, при котором получают значительное сокращение объема цифрового описания изображений за счет неравномерности появления отдельных значений квантованного сигнала. При этом используются неравномерные коды-Шеннона – Фано и Хаффмена, кодирующие короткими кодовыми последовательностями часто встречающиеся значения отсчетов и длинными кодовыми последовательностями редко встречающиеся значения, а также различные методы кодирования редких символов в сочетании с кодами Шеннона – Фано и Хаффмена. Последние применяются в основном в системах с ДИКМ и с адаптивной дискретизацией.

Среди наиболее эффективных методов кодирования редких символов следует отметить методы кодирования с прослеживанием линий (например, линий равных значений). Методы кодирования редких символов находят также широкое применение в устройствах преобразования графических изображений в цифровой сигнал.

Глава 3
ДИСКРЕТНОЕ ПРЕДСТАВЛЕНИЕ ПРЕОБРАЗОВАНИЙ СИГНАЛОВ

3.1. ПРИНЦИПЫ ЦИФРОВОГО ПРЕДСТАВЛЕНИЯ ПРЕОБРАЗОВАНИИ

При описании и построении цифровых преобразований непрерывных сигналов необходимо соблюдать *принцип соответствия между непрерывными и цифровыми преобразованиями.* Непрерывное и цифровое преобразование соответствуют друг другу, если одинаковые входные сигналы они преобразуют в одинаковые выходные. Таким образом, способы цифрового представления непрерывных преобразований должны основываться на способах цифрового представления сигналов с учетом процедур дискретизации, квантования и восстановления непрерывных сигналов из цифровых (рис. 3.1).

В § 1.3 были выделены два класса преобразований сигналов: поэлементные и линейные. Цифровое представление поэлементных преобразований основывается на их определении как функций от значений коэффициентов дискретного представления сигналов:

$$\mathbf{Ta} = \{T_k(\alpha_k)\}$$

где $\{\alpha_k\}$—квантованные значения коэффициентов представления сигнала *a(x)* по надлежащим образом выбранному базису дискретизации. Поскольку значения $\{\alpha_k\}$ квантованы, функции $T_k(\cdot)$ проще всего задаются в табличной форме. Объем таблицы определяется количеством квантованных значений α_k (количеством уровней квантования). Если это количество велико, прибегают к другим способам задания функции, чаще всего — заданию ее в виде полинома

$$T_k(\alpha_k) = \sum_{r=0}^{N-1} c_r \alpha_k^r.$$

Цифровое представление линейных преобразований разбивается на два этапа: дискретизацию и квантование полученного дискретного представления. Дискретное представление непрерывных линейных преобразований основывается на дискретном представлении непрерывных сигналов.

Рис. 3.1. Принцип соответствия между непрерывными и цифровыми преобразованиями

Пусть $\{\varphi_k = \varphi_k(x)\}$, $\{\psi_k = \psi_k(x)\}$, $\{\eta_k = \eta_k(\xi)\}$ и $\{\Theta_k = \Theta_k(\xi)\}$ — соответственно базисы восстановления и базисы дискретизации входных и выходных сигналов линейного преобразования L; $\{\alpha_k\}$ и $\{\beta_k\}$ — коэффициенты представления сигналов по этим базисам, определяемые формулой (1.4), так что входные и выходные сигналы *a(x)* и $b(\xi)$ преобразования аппроксимируются функциями $\tilde{a}(x)$ и $\tilde{b}(\xi)$:

$$\tilde{a}(x) = \sum_{k=0}^{N_a-1} \alpha_k \varphi_k(x) \qquad (3.1a)$$

$$\tilde{b}(\xi) = \sum_{k=0}^{N_b-1} \beta_k \eta_k(\xi) \qquad (3.1б)$$

Найдем связь $\{\beta_k\}$ и $\{\alpha_k\}$, считая, что $\tilde{b}(\xi)$ как аппроксимация $b(\xi)$ может быть получена в результате действия преобразования L на сигнал $\tilde{a}(x)$, аппроксимирующий *a(x).* Так как

$$\beta_k = (\tilde{b},\ \Theta_k) = (L\tilde{a},\ \Theta_k),$$

то, подставив сюда (3.1а), получим

$$\beta_k = \sum_{n=0}^{N_a-1} h_{k,n}\alpha_n \tag{3.2}$$

где

$$h_{k,n} = (L\varphi_n,\ \Theta_k) = \int_X [L\varphi_n(x)]\,\Theta_k(\xi)\,d\xi \tag{3.3}$$

Матрица $H=\{h_{k,n}\}$ является дискретным представлением непрерывного линейного преобразования **L**. Формула (3.2) является, очевидно, дискретным аналогом формулы (1.46), а (3.3)—формулы (1.47). Если базисы входных и выходных сигналов совпадают, то дискретное представление линейного преобразования **L** согласно (3.3) есть матрица чисел, представляющих собой отсчеты проекции импульсной реакции фильтра $h(x,\xi)$ на базисные функции:

$$h_{k,n} = (L\varphi_n,\ \psi_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,\ \xi)\,\varphi_n(\xi)\,\psi_k(x)\,dx\,d\xi \tag{3.4}$$

Если базис входных и выходных сигналов является собственным базисом преобразования **L**, так что

$$L\varphi_k(x) = h_k\varphi_k(x),$$

матрица $\{h_{k,n}\}$ вырождается в диагональную матрицу:

$$h_{k,n} = h_k\delta_{k,n} \tag{3.5}$$

где $\delta_{k,n}$ —символ Кронекера (1.3).

Дискретное линейное преобразование, описываемое диагональной матрицей, будем называть *скалярным фильтром.*

Формула (3.3) показывает, как описывается дискретное линейное преобразование, представляющее заданное непрерывное преобразование по базисам $\{\varphi_n\}$ и $\{\Theta_k\}$. Следуя принципу соответствия между непрерывными и дискретными преобразованиями (рис. 3.1), найдем импульсную реакцию непрерывного преобразования, эквивалентного дискретному преобразованию, описываемому матрицей чисел $\{h_{k,n}\}$. Подставив (3.2) в (3.1 а) и выразив $\{a_n\}$ через $a(x)$ по (1.4), получим:

$$\tilde{b}(\xi) = \int_X a(x)\left(\sum_{k=0}^{N_b-1}\sum_n h_{k,n}\eta_k(\xi)\,\psi_n(x)\right)dx$$

Отсюда импульсная реакция непрерывного преобразования, соответствующего дискретному преобразованию с матрицей $\{h_{k,n}\}$, равна:

$$\tilde{h}(x,\ \xi) = \sum_{k=0}^{N_b-1}\sum_n h_{k,n}\eta_k(\xi)\,\psi_n(x). \tag{3.6}$$

Частотная характеристика эквивалентного непрерывного преобразования, очевидно, определяется выражением

$$\tilde{H}(f,\ p) = \sum_{k=0}^{N_b-1}\sum_n h_{k,n}\mathcal{H}_k(-p)\,\Psi_n(f), \tag{3.7}$$

где $\mathcal{H}_k(p)$ и $\Psi_n(f)$— преобразования Фурье функций $\eta_k(\xi)$ и $\psi_n(x)$ соответственно, т.е частотные характеристики устройства восстановления непрерывных сигналов из дискретных и устройства дискретизации.

3.2. ЦИФРОВЫЕ ФИЛЬТРЫ

Одним из самых важных в приложениях является класс линейных преобразований, для которых аргументы сигнала до и после преобразования совпадают. Такие преобразования будем называть *линейной фильтрацией,* а их *цифровую реализацию* — цифровыми фильтрами. Цифровое представление линейных фильтров строится через дискретное представление по сдвиговым базисным функциям.

Самыми распространенными и наиболее простыми в реализации являются фильтры, инвариантные к сдвигу. Они определены для инфинитных сигналов, и их импульсная реакция зависит только от разности аргументов. Соответственно и их дискретное представление по сдвиговым базисным функциям дискретизации $\lambda_д(x—k\Delta x)$ и восстановления $\lambda_в(x—k\Delta x)$

$$h_{k,n} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x-\xi)\lambda_д(x-k\Delta x)\lambda_в(\xi-n\Delta x)\,dx\,d\xi =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h[x-\xi+(k-n)\Delta x]\lambda_д(x)\lambda_в(\xi)\,dx\,d\xi = h_{k-n}.$$

также зависит только от разности номеров отсчетов входного и выходного сигналов. Следовательно, для цифровых фильтров получаем из (3.2) соотношение

$$b_k = \sum_{n=-\infty}^{\infty} a_n h_{k-n} \qquad (3.8)$$

При цифровой обработке количество слагаемых в (3.8) должно быть ограничено. Удобно это ограничение связывать не с протяженностью сигнала, как в (3.8), а с протяженностью импульсной реакции фильтра, которая, как правило, меньше протяженности сигнала. Заменив в (3.8) индексы суммирования и введя ограничение по количеству отсчетов импульсной реакции фильтра, получим конечномерное приближение к непрерывному фильтру

$$b_k = \sum_{n=0}^{N-1} h_n a_{k-n} \qquad (3.9)$$

Эта формула называется формулой *цифровой свертки.*

Импульсную реакцию непрерывного фильтра, соответствующего этому цифровому, можно в соответствии с (3.6) записать как

$$\tilde{h}(x,\ \xi) = \sum_{k=0}^{N_b-1}\sum_{n=0}^{N-1} h_n\varphi_0[x-(k-n)\Delta x]\psi_0(\xi-n\Delta x) \qquad (3.10)$$

где N_b—количество отсчетов $\{b_k\}$, используемых для восстановления непрерывного сигнала $b(x)$, а частотную характеристику — в соответствии с (3.7) как

$$\tilde{H}(f,\ p)=\left[\sum_{n=0}^{N-1} h_n \exp\left(i2\pi n\Delta x\right)\right]\left[\sum_{k=0}^{N_b-1} \exp\left[i2\pi(f-p)\,k\Delta x\right]\right]\times$$

$$\times H_в(-p)\,H_д(f)=\left[\sum_{n=0}^{N-1} h_n \exp\left(i2\pi n\Delta x\right)\right]\times$$

$$\times\frac{\sin\pi(f-p)(N_b-1)\,\Delta x}{\sin\pi(f-p)\,\Delta x}\exp\left[i\pi(f-p)(N_b-1)\,\Delta x\right]\times$$

$$\times H_в(-p)\,H_д(f).$$

Отбросив в этой формуле несущественный экспоненциальный множитель, определяемый сдвигом выходного сигнала дискретного фильтра относительно входного, получим окончательно

$$\tilde{H}(f,\ p)=\left[\sum_{n=0}^{N-1} h_n \exp\left(i2\pi np\Delta x\right)\right](N_b-1)\,\mathrm{sincd}\left[(N_b-1);\right.$$

$$\left.\pi(f-p)\,\Delta x\right]H_в(-p)\,H_д(f), \qquad (3.11)$$

где $H_д(f)$, $H_в(p)$—преобразования Фурье функций $\lambda_в(\xi)$, $\lambda_д(x)$, а

$$\mathrm{sincd}\,(N;\ x)=(\sin Nx)/N\sin x \qquad (3.12)$$

— дискретный аналог функции sinc x.

В частном случае, когда $\lambda_д(x)$ и $\lambda_в(\xi)$ —идеальные отсчетные функции вида $\mathrm{sinc}\,(\pi x/\Delta x)$

$$H_д(f)=\Delta x H_в(f)=\mathrm{rect}\,(f\Delta x+1/2),$$

гак что

$$\tilde{H}(f,\ p)=\begin{cases}\left[\sum_{n=0}^{N-1} h_n \exp\left(i2\pi n\Delta x\right)\right](N_b-1)\,\Delta x\,\mathrm{sincd}\left[(N_b-1)\right.\\ \left.\pi(f-p)\,\Delta x\right],\ f,p\in\left[-\frac{1}{2\Delta x};\ \frac{1}{2\Delta x}\right]\\ 0,\ f,\ p\notin\left[-\frac{1}{2\Delta x};\ \frac{1}{2\Delta x}\right].\end{cases} \qquad (3.13)$$

Сумма по k к множитель $(N_b-1)\,\mathrm{sincd}\left[(N_b-1),\ \pi(f-p)\,\Delta x\right]$ в (3.11) учитывают конечность последовательности отсчетов $\{b_k\}$ на выходе дискретного фильтра, из которых восстанавливается непрерывный сигнал. Множители же $H_в(-p)$ и $H_д(f)$ учитывают процедуру дискретизации входного сигнала и восстановления непрерывного выходного сигнала по его отсчетам.

Из формулы (3.11) вытекает, что непрерывный фильтр, эквивалентный дискретному фильтру (3.9), не является инвариантным к сдвигу. Связано это с тем, что количество членов N_b в сумме по k в (3.11) конечно, т.е. с краевыми эффектами. Они будут рассмотрены ниже. При $N\to\infty$ функция $(N_b-1)\mathrm{sincd}\,[(N_b-1);\ \pi(f-p)\,\Delta x]$ в (3,13) стремится к дельта-функции $\delta(f-p)$, т.е.

$$\lim_{N_b\to\infty}\tilde{H}(f,\ p)=\left[\sum_{n=0}^{N-1} h_n \exp\left(i2\pi pn\Delta x\right)\right]H_д(f)\,H_в(-p)\,\delta(f-p),$$

так что фильтр становится инвариантным к сдвигу.

По аналогии с непрерывной формой записи дискретного сигнала (1.23) можно ввести функцию

$$\tilde{\tilde{h}}(x)=\sum_{n=0}^{N-1} h_n\delta(x-n\Delta x) \qquad (3.14)$$

зависящую только ог отсчетов импульсной реакции цифрового фильтра. Эта функция описывает только сам дискретный фильтр н не учитывает краевые эффекты и процедуры

дискретизации и восстановления непрерывных сигналов. Назовем ее *дискретной импульсной реакцией* цифрового фильтра (3.9), заданного своими коэффициентами $\{h_n\}$, а его преобразование Фурье — сумму по n в (3.11)

$$\tilde{\tilde{H}}(p) = \sum_{n=0}^{N-1} h_n \exp(i\,2\pi p n \Delta x) \tag{3.15}$$

— *дискретной частотной характеристикой* цифрового фильтра. Соответственно выражения (ЗЛО) и (3.11) можно назвать непрерывной *импульсной реакцией* и *непрерывной частотной характеристикой* дискретного фильтра.

Знание частотных характеристик фильтров необходимо для синтеза фильтров и сопоставления их с непрерывными фильтрами, которые они аппроксимируют. Рассмотрим в качестве примера часто используемый в обработке изображений фильтр, вычисляющий текущее среднее последовательности отсчетов по окрестности из *(2N* + 1) отсчетов:

$$\bar{a}_k = \frac{1}{2N+1} \sum_{n=-N}^{N} a_{k-n} \tag{3.16}$$

Для этого фильтра

$$h_n = 1/(2N+1) \tag{3.17}$$

Подставив (3.17) в (3.15), найдем, что его дискретная частотная характеристика определяется выражением:

$$\tilde{\tilde{H}}(p) = \frac{1}{2N+1} \sum_{n=-N}^{N} \exp(i\,2\pi n p \Delta x) = \operatorname{sincd}[(2N+1);\ \pi p \Delta x] \tag{3.18}$$

Эта частотная характеристика, построенная для $N = 2$ и 128, показана на рис. 3.2, а, *б* соответственно. Отметим, что непрерывные частотные характеристики такого фильтра при идеальных отсчетных функциях в качестве базисов дискретизации и восстановления отличаются от показанных на рисунке тем, что за пределами интервала частот ($\pm 1/2\Delta x$) они равны нулю. Штриховой линией на

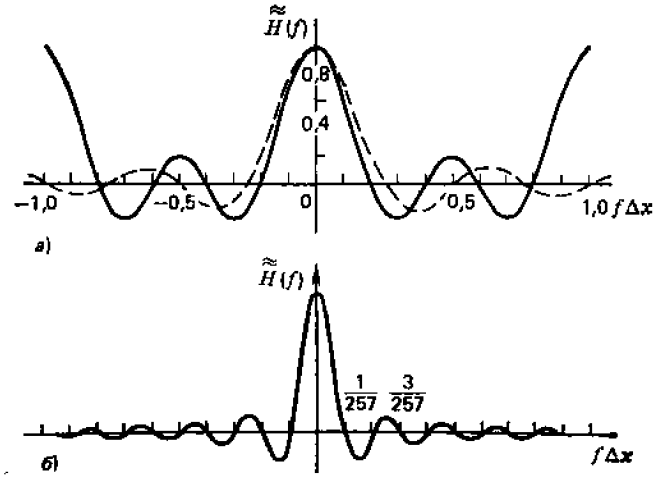


Рис. 3.2. Дискретные частотные характеристики фильтра (3.26):
а) *N=2;* б) jV = 128; — — — — частотная характеристика непрерывного фильтра (3.19)

рис. 3.2, *а* показана частотная характеристика $H(f) = \operatorname{sinc}(\pi N f \Delta x)$ непрерывного фильтра с импульсной реакцией

$$h(x) = \frac{1}{2N\Delta x}\ \mathrm{rect}\left(\frac{x}{2N\Delta x} + \frac{1}{2}\right) \qquad (3.19)$$

отсчетами которой являются весовые коэффициенты $\{h_n\}$ дискретного фильтра (3.16). Нетрудно видеть, что при малом N различие между этими характеристиками может быть заметным. Эти частотные характеристики лучше совпадают, если интервал суммирования для непрерывного фильтра простирается от -(2N+1}Δx/2 до (2N+1) Δ*x*/2, т.е шире на половину интервала дискретизации в ту и другую сторону интервала суммирования дискретного фильтра.

Все соотношения этого параграфа очевидным образом обобщаются на двумерный мучай заменой аргументов и индексов соответствующими парами аргументов и индексов для растрирования сигналов в прямоугольной системе координат. Так, двумерная цифровая свертка определяется при дискретизации на прямоугольном растре выражением

$$b_{k,l} = \sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1} h_{m,n}\, a_{k-m,\ l-n} \qquad (3.20)$$

дискретная частотная характеристика (3.15) фильтра — выражением

$$\tilde{\tilde{H}}(f_1,\ f_2) = \sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1} h_{m,n} \exp\left[i\,2\pi\left(f_1 m\Delta x_1 + f_2 n\Delta x_2\right)\right], \qquad (3\text{-}21)$$

а непрерывная частотная характеристика — выражением

$$\tilde{H}(f_1,\ p_1,\ f_2,\ p_2) = \sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1} h_{m,n} \exp\left[i\,2\pi\left(f_1 m\Delta x_1 + f_2 n\Delta x_2\right)\right] \times$$
$$\times\,(N_{b,1}-1)\,(N_{b,2}-1)\,\mathrm{sincd}\left[(N_{b,1}-1);\ \pi(f_1 - p_1)\Delta x_1\right] \times$$
$$\times\,\mathrm{sincd}\left[(N_{b,2}-1);\ \pi(f_2 - p_2)\Delta x_2\right] H_{\text{л}}(f_1;\ f_2)\, H_{\text{в}}(-p_1,\ -p_2). \qquad (3.22)$$

3.3. ДИСКРЕТНЫЕ ПРЕОБРАЗОВАНИЯ ФУРЬЕ

Непрерывные прямое и обратное преобразования Фурье, определяемые формулами (1.25), являются линейными преобразованиями с ядром соответственно $H_1(f,\ x) = \exp(i2\pi fx)$ и $H_2(x,f) = \exp(-i2\pi fx)$. Их дискретные представления по базисам из идеальных отсчетных функций

$$\varphi_k(x) = \mathrm{sinc}\left[\pi(x - k\Delta x)/\Delta x\right] \qquad (3.23a)$$
$$\chi_r(f) = \mathrm{sinc}\left[\pi(f - r\Delta f)/\Delta f\right], \qquad (3.23б)$$

имеют в соответствии с (3.3) следующий вид:

$$a_r = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k \exp\left(i\,2\pi kr/N\right) \qquad (3.24a)$$

$$a_k = \frac{1}{\sqrt{N}} \sum_{r=0}^{N-1} a_r \exp\left(-i\,2\pi kr/N\right) \qquad (3.24б)$$

где $\{\alpha_k\}$—последовательность отсчетов сигнала:

$$a_k = \frac{1}{\Delta x} \int_{-\infty}^{\infty} a(x)\,\mathrm{sinc}\left[\pi(x - k\Delta x)/\Delta x\right] dx; \qquad (3.25a)$$

$\{\alpha_k\}$ — последовательность отсчетов спектра сигнала:

$$\alpha_r = \frac{1}{\Delta f}\int_{-\infty}^{\infty} \alpha(f)\,\mathrm{sinc}\,[\pi(f - r\Delta f)/\Delta f]\,df \quad ; \qquad (3.256)$$

Δx и Δf — интервалы растрирования сигнала и его спектра Фурье; $N=1/\Delta x \Delta f$.

Соотношения (3,24а) и (3.246) называются *прямым* и *обратным дискретными преобразованиями Фурье (ДПФ)* последовательности.

Аналогично получаются двумерные дискретные преобразования Фурье:

$$\alpha_{r,s} = \frac{1}{\sqrt{N_1 N_2}} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right], \qquad (3.26а)$$

$$a_{k,l} = \frac{1}{\sqrt{N_1 N_2}} \sum_{r=0}^{N_1-1} \sum_{s=0}^{N_2-1} \alpha_{r,s} \exp\left[-i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] \qquad (3.266)$$

для растрирования сигналов и их спектров на прямоугольном растре. Нетрудно видеть, что двумерные ДПФ сводятся к двум одномерным.

Приведем важнейшие свойства ДПФ. В следующих ниже формулах латинскими буквами с индексами обозначены отсчеты сигналов, греческими — отсчеты их ДПФ.

Цикличность. Последовательность коэффициентов ДПФ является периодической последовательностью

$$\alpha_r = \alpha_{(r) \bmod N}; \qquad (3.27а)$$

Последовательность отсчетов сигнала, полученная обратным ДПФ из ее коэффициентов ДПФ, также определяется на номера отсчетов, выходящие за пределы заданного интервала (0, N—1), как периодическая последовательность

$$a_k = \text{ДПФ}(\{\alpha_r\}) = a_{(k) \bmod N} \qquad (3.276)$$

Это свойство вытекает из периодичности ядра преобразования $\{\exp(i2\pi kr/N)\}$. Вспомним также, что растрирование сигнала соответствует периодическому продолжению его спектра Фурье (см. §2.2).

Симметрия.

$$\{a_{N-k}\} \leftarrow \text{ДПФ} \rightarrow \{\pm \alpha_{N-r}\}; \qquad (3.28)$$

$$\{a_k^*\} \leftarrow \text{ДПФ} \rightarrow \{\alpha_{N-r}^*\}; \qquad (3.29а)$$

$$\{a_k = \pm a_k^*\} \leftarrow \text{ДПФ} \rightarrow \{\alpha_r = \pm \alpha_{N-r}^*\} \qquad (3.296)$$

Эти соотношения также показывают, что понятия симметрии четности и нечетности для последовательностей, порождаемых ДПФ, в отличие от непрерывных сигналов определены не относительно точки 0, а относительно точки с номером $N/2$. Таким образом, в силу целочисленности номеров элементов последовательностей смысл четности и нечетности этих последовательностей зависит от того, является количество элементов последовательностей N четным или нечетным.

Варианты симметрии для одномерных и двумерных ДПФ при четных и нечетных N_1 и N_2 иллюстрируются на рис. 3.3.

Соотношение Парсеваля:

$$\sum_{k=0}^{N-1} a_k b_k^* = \sum_{r=0}^{N-1} \alpha_r \beta_r \qquad (3.30)$$

Эго соотношение является, очевидно, частным случаем формулы (1.5).

Теорема сдвига:

$$\{a_{(k+k_0) \bmod N}\} \leftarrow \text{ДПФ} \rightarrow \alpha_r \exp(-i2\pi k_0 r/N) \qquad (3.31)$$

$$\{a_k \exp(i 2\pi k r_0/N)\} \leftarrow \text{ДПФ} \rightarrow \alpha_{(r+r_0) \bmod N}$$

Теорема о циклической свертке:

$$\left\{\sum_{k=0}^{N-1} a_n b_{(k-n) \bmod N}\right\} \leftarrow \text{ДПФ} \rightarrow \sqrt{N}\alpha_r\beta_r \tag{3.32}$$

Эта теорема является дискретным аналогом теоремы о свертке интегрального преобразования Фурье. Но в отличие от интегрального преобразования Фурье ДПФ от произведения дискретных спектров двух сигналов дает не арифметическую свертку этих сигналов, а циклическую свертку, т.е свертку сигналов, периодически продолженных за пределы заданного интервала. Дискретная теорем а отсчете в:

$$\{a_{(i) \bmod N};\ i=0,\ 1,\ \ldots,\ LN-1\} \leftarrow \text{ДПФ} \rightarrow \sqrt{L}\alpha_{(r) \bmod N} \times \times \delta((r) \bmod L) \tag{3.33а}$$

где $\delta(r)$— символ Кронекера. Эта теорема связана с теоремой отсчетов для непрерывных сигналов (см. § 2.2) и означает, что периодическое L-кратное продолжение последовательности приводит к раздвиганию отсчетов ее ДПФ с шагом L, причем появляющие-

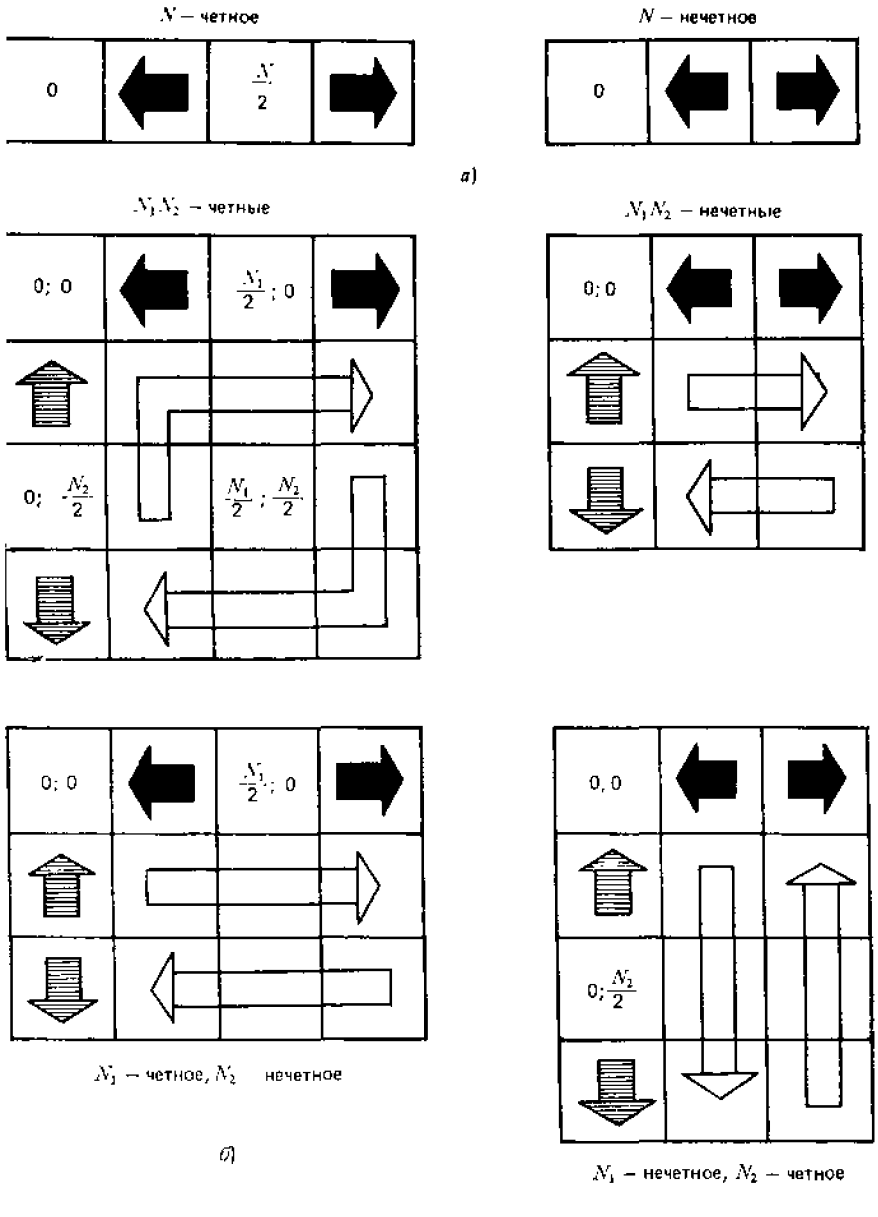


Рис. 3.3. Типы симметрии для одномерного (а) и двумерного *(б)* ДПФ при четных и нечетных N_1 и N_2

ся при этом дополнительные отсчеты имеют нулевое значение. Аналогичное соотношение имеется и для периодического продолжения ДПФ-спектра:

$$\{a_{(l) \bmod N}\delta((l) \bmod L);\ l=0,\ 1,\ \ldots,\ LN-1\} \leftarrow$$
$$\leftarrow \text{ДПФ} \rightarrow \{(1/\sqrt{L})\,a_{(r) \bmod N};\ r=0,\ 1,\ \ldots,\ LN-1\} \qquad (3.336)$$

Теорема об интерполяции. При N-четном:

$$\left\{\left[1-\operatorname{rect}\frac{l-N/2}{(L-1)N}\right]a_{(l) \bmod N};\ l=0,\ 1,\ \ldots,\ LN-1\right\} \leftarrow$$
$$\leftarrow \text{ДПФ} \rightarrow \left\{\frac{N-1}{N\sqrt{L}}\sum_{s=0}^{N-1}\alpha_s\,\operatorname{sincd}[(N-1);\ \pi(r-sL)/LN];\right.$$
$$r=0,\ 1,\ \ldots,\ LN-1\}. \qquad (3.34а)$$

При N-нечетном:

$$\left\{\left[1-\operatorname{rect}\frac{l-(N+1)/2}{(L-1)N+1}\right]a_{(l) \bmod N};\ l=0,\ 1,\ \ldots,\ LN-1\right\} \leftarrow$$
$$\leftarrow \text{ДПФ} \rightarrow \left\{\frac{1}{\sqrt{L}}\sum_{s=0}^{N-1}\alpha_s\,\operatorname{sincd}[N;\ \pi(r-sL)/LN];\ r=0,\ 1,\ \ldots\right.$$
$$\ldots,\ LN-1\}. \qquad (3.346)$$

Эта теорема также тесно связана с теоремой отсчетов для непрерывных сигналов, и смысл ее состоит в том, что спектр последовательности, полученной раздвиганием и дополнением нулями элементов некоторой исходной последовательности [способы дополнения нулями для четного и нечетного N несколько отличаются, как это видно из формул (3.34)], образуется с помощью интерполяции отсчетов ДПФ исходной последовательности функцией sincd(·) (3.12), являющейся дискретным аналогом отсчетной функции sinc x

Теорема о перестановках. Если P не имеет общих делителей с N, то

$$\{a_{(Pk) \bmod N}\} \leftarrow \text{ДПФ} \rightarrow \{\alpha_{(Qr) \bmod N}\} \qquad (3.35)$$

где Q определяется из условия {PQ)mod N= 1. Эта теорема является аналогом теоремы о масштабах интегрального преобразования Фурье. Но если изменение масштаба сигнала, скажем, растяжение сигнала по координате в P раз, приводит к сжатию его спектра Фурье в P раз, то для последовательностей и их ДПФ это соответствует перестановкам элементов последовательностей.

Аналогичные соотношения и теоремы справедливы и для двумерных ДПФ с той лишь разницей, что для сопоставления с непрерывным случаем двумерные последовательности, порождаемые ДПФ, следует рассматривать как периодические по двум индексам (двум координатам):

$$a_{k,l} = a_{(k) \bmod N_1,\ (l) \bmod N_2}$$

$$a_{r,s} = a_{(r) \bmod N_1,\ (s) \bmod N_2}.$$

Базисные функции (3.23) и соответствующие им двумерные функции, использованные выше для построения дискретного представления непрерывного преобразования Фурье, таковы, что нулевые отсчеты сигналов и их спектров попадают в начало координат. Между тем в принципе возможен произвольный сдвиг базисных функций относительно координат сигнала и спектра, так как растр дискретизации определяется работой устройств дискретизации и не зависит от сигнала.

Если в одномерном случае сдвиг нулевого отсчета сигнала относительно начала его координат равен $u\Delta x$, а сдвиг нулевого отсчета спектра сигнала относительно начала координат спектра равен $v\Delta f$, то дискретные представления непрерывных прямого и обратного преобразования Фурье получаются в виде [46, 49]:

(3.36а)

$$a_r^{u,v} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k \exp\left[i\,2\pi (k+u)(r+v)/N\right];$$

$$a_k^{u,v} = \frac{1}{\sqrt{N}} \sum_{r=0}^{N-1} a_r^{u,v} \exp\left[-i\,2\pi (k+u)(r+v)/N\right] \qquad (3.36б)$$

Индексы u и v над a_k в (3.36б) поставлены для того, чтобы подчеркнуть, что в отличие от исходной последовательности $\{a_k\}$ отсчетов сигнала $a(x)$, определенной для $k = 0, 1, \ldots, N-1$, $\{a_k^{u,v}\}$ определена для любых k. При $k \in [0, N-1]$ $\{a_k^{u,v}\}$ совпадает с $\{a_k\}$.

В выражения (3.36) входят постоянные множители $\exp(\pm i2\pi uv/N)$, не влияющие на свойства преобразований, но требующие дополнительных затрат на их вычисление. Исключив их, получим следующую пару преобразований:

$$a_r^{u,v} = \frac{1}{\sqrt{N}} \left[\sum_{k=0}^{N-1} a_k \exp\left[i\,2\pi k(r+v)/N\right]\right] \exp(i\,2\pi ru/N); \qquad (3.37а)$$

$$a_k^{u,v} = \frac{1}{\sqrt{N}} \left[\sum_{r=0}^{N-1} a_r^{u,v} \exp\left[-i\,2\pi r(k+u)/N\right]\right] \exp(-i2\pi kv/N) \qquad (3.37б)$$

Будем называть их *сдвинутыми дискретными преобразованиями Фурье* СДПФ (u, v). Эти преобразования выражаются через стандартное ДПФ (3.24), например:

$$a_r^{u,v} = \left[\sum_{k=0}^{N-1} \left(a_k \exp\left(i2\pi \frac{kv}{N}\right)\right) \exp\left(i2\pi \frac{kr}{N}\right)\right] \times$$

$$\times \exp\left(i2\pi \frac{ru}{N}\right)$$

Приведем без вывода некоторые наиболее важные свойства СДПФ.

Обобщенная цикличность. При целых h и g

$$a_{k+hN}^{u,v} = a_k \exp(-i\,2\pi hv)$$

$$a_{r+gN}^{u,v} = a_r^{u,v} \exp(i\,2\pi gu) \qquad (3.38)$$

Здесь в отличив от простого периодического продолжения (цикличности), характерного для стандартного ДПФ, от периода к периоду происходит поворот комплексных чисел, составляющих последовательность, на угол, пропорциональный сдвигу.

Симметрия:

$$\left\{a_{N-k}^{u,v} \exp\left[i2\pi\left(v - \frac{2kv}{N}\right)\right]\right\} \leftarrow \text{СДПФ} \rightarrow$$

$$\rightarrow \left\{a_{N-r}^{u,v} \exp\left[-i2\pi\left(u - \frac{2ru}{N}\right)\right]\right\}. \qquad (3.39)$$

Если $2u$, $2v$ — целые числа, то

$$\{(-1)^{2v}a^{u,v}_{N-2u-k}\} \leftarrow \text{СДПФ} \rightarrow \{(-1)^{2u}\alpha^{u,v}_{N-2v-r}\} \tag{3.40а}$$

$$\left\{a^{u,v}_k = \pm\, a^{u,v}_{N-k} \exp\left[i2\pi\left(v - \frac{2kv}{N}\right)\right]\right\} \leftarrow \text{СДПФ} \rightarrow$$

$$\rightarrow \left\{\alpha^{u,v}_r = \pm\, \alpha^{u,v}_{N-r} \exp\left[-i2\pi\left(u - \frac{2ru}{N}\right)\right]\right\}; \tag{3.40б}$$

$$\{a^{u,v}_k = \pm\,(-1)^{2v}a^{u,v}_{N-2u-k}\} \leftarrow \text{СДПФ} \rightarrow \{\alpha^{u,v}_r = \pm\,(-1)^{2u} \times$$

$$\times\, \alpha^{u,v}_{N-2v-r}\} \tag{3.40в}$$

$$\{a^{u,v}_k = \pm\,(-1)^{2v}a^{u,v}_{N-2u-k} = \pm\,(a^{u,v}_k)^*\} \leftarrow \text{СДПФ} \rightarrow$$

$$\rightarrow \{\alpha^{u,v}_r = \pm\,(-1)^{2u}\alpha^{u,v}_{N-2v-r} = \pm\,(-1)^{2u}\,(\alpha^{u,v}_{N-2v-r})^*\} \tag{3.40г}$$

$$\left\{a^*_k \exp\left(-i2\pi\frac{2kv}{N}\right)\right\} \leftarrow \text{СДПФ} \rightarrow \{(a^{u,v}_{N-r})^* \exp(i2\pi u)\} \tag{3.41а}$$

Если-2v — целое число, то

$$\left\{a^*_k \exp\left(i2\pi\frac{2uv}{N}\right)\right\} \leftarrow \text{СДПФ} \rightarrow \{(\alpha^{u,v}_{N-2v-r})^* \exp(i2\pi u)\} \tag{3.41б}$$

$$\{a^*_{N-k} \exp\{-i2\pi v)\} \leftarrow \text{СДПФ} \rightarrow \left\{(\alpha^{u,v}_r)^* \exp\left(i2\pi\,\frac{2ru}{N}\right)\right\} \tag{3.41в}$$

Если $2u$ — целое число, то

$$\{a^*_{N-2u-k} \exp\{-i2\pi v)\} \leftarrow \text{СДПФ} \rightarrow$$

$$\rightarrow \left\{(a^{u,v}_r)^* \exp\left(-i2\pi\,\frac{2uv}{N}\right)\right\}. \tag{3.41г}$$

Здесь проявляется характерная особенность сдвинутых дискретных преобразований Фурье: полуцелые параметры сдвига выделяются среди других (см. также [47]).

Теорема сдвига:

$$\{a^{u,v}_{k+k_0}\} \leftarrow \text{СДПФ} \rightarrow \{\alpha^{u,v}_r \exp\left[-i2\pi\,\frac{k_0(r+v)}{N}\right]\} \tag{3.42а}$$

$$\left\{a^{u,v}_k \exp\left[-i2\pi\,\frac{(k+u)r_0}{N}\right]\right\} \leftarrow \text{СДПФ} \rightarrow \{\alpha^{u,v}_{r+r_0}\} \tag{3.42б}$$

Теорема об интерполяции. Если прямое СДПФ последовательности $\{a_k\}$ выполнить с параметрами сдвига u, v, затем сдвинуть спектр $\{\alpha^{u,v}_r\}$ на r_0 и выполнить обратное СДПФ с параметрами p, q по M отсчетам сдвинутого спектра $\{\alpha^{u,v}_{r+r_0}\}$, то полученная последовательность $\{{}_{r_0}a^{p,q\,|\,u,v}_k\}$ окажется связанной с исходной последовательностью $\{a_k\}$ интерполяционным соотношением:

$$
{}^{r_0}a_k^{p,\,q\,|\,u,\,v} = \frac{1}{\sqrt{N}} \sum_{r=0}^{M-1} \alpha_{r+r_0}^{u,\,v} \exp\left(-i2\pi \frac{pr}{N}\right) \times
$$

$$
\times \exp\left[-i2\pi \frac{(r+q)\,k}{N}\right] = \sum_{n=0}^{N-1} a_n \times
$$

$$
\times \exp\left[i2\pi \frac{(n+u)(r_0+v-(M-1)/N)}{N}\right] \times
$$

$$
\times \frac{M}{N} \operatorname{sincd}\left(M;\ \pi \frac{n+u-k-p}{N}\right) \times
$$

$$
\times \exp\left[-i2\pi (k+p)\left(q+\frac{M-1}{2}\right)/N\right] \times
$$

$$
\times \exp\left[i2\pi (pq-uv)/N\right]. \tag{3.43}
$$

При $q = -(M-1)/2;\ v = -r_0 - (M-1)/2;\ M = N;\ pq = (uv) \bmod N$:

$$
{}^{r_0}a_k^{p,\,q|u,\,v} = \sum_{n=0}^{N-1} a_n \operatorname{sincd}\left(N;\ \pi \frac{n+u-k-p}{N}\right) \tag{3.44а}
$$

При $q = -(M-2)/2;\ v = -r_0 - (M-2)/2;\ M = N-1;\ pq = (uv) \bmod N$:

$$
{}^{r_0}a_k^{p,\,q|u,\,v} = \sum_{n=0}^{N-1} a_n \frac{N-1}{N} \operatorname{sincd}\left[(N-1);\ \pi \frac{n+u-k-p}{N}\right] \tag{3.44б}
$$

Это значит, что с помощью обратного СДПФ, меняя параметры сдвига по координате k, можно получать интерполированные произвольно расположенные промежуточные отсчеты последовательности. Это свойство СДПФ аналогично свойствам интегрального преобразования Фурье, для которого также можно получать сдвинутые копии сигнала с помощью сдвига по координате при обратном преобразовании Фурье. Оно показывает, что введение «аналогового» параметра сдвига придает СДПФ свойства, которыми не обладает стандартное ДПФ и которые сближают СДПФ с интегральным преобразованием Фурье.

Теорема о свертке. Пусть $\{\alpha_r^{u_a,\,v_a}\}$ — спектр СДПФ $(u_a,\ v_a)$ последовательности $\{a_k\}$; $\{\beta_r^{u_b,\,v_b}\}$ —спектр СДПФ (u_b, v_b)' последовательности $\{b_k\}$.

Тогда

$$
{}^{r_a,\,r_b}c_k^{u_c,\,v_c} = \frac{1}{\sqrt{N}} \sum_{r=0}^{M-1} \alpha_{r+r_a}^{u_a,\,v_a} \beta_{r+r_b}^{u_b,\,v_b} \exp\left[-i2\pi \frac{r(k+u_c)}{N}\right] \times
$$

$$
\times \exp\left(-i2\pi \frac{kv_c}{N}\right) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} a_n^{u_a,\,v_a} \exp\left[i2\pi (n+u_a)(r_a+\right.
$$

$$
\left.+\,v_a - v_c)/N\right] \tilde{b}_{k-n} \exp\left[-i2\pi u_a (v_a - v_c)/N\right], \tag{3.45а}
$$

где

$$
\tilde{b}_l = \sum_{m=0}^{N-1} b_m^{u_b,\,v_b} \exp\left[i2\pi (m+u_b)(r_b+v_b+(M-1)/2)/N\right] \frac{M}{N} \times
$$

$$
\times \operatorname{sincd}\left[M;\ \pi \frac{m-(l+u_c-u_a-u_b)}{N}\right] \times
$$

$$
\times \exp\left[i2\pi \frac{(u_c-u_a)\,v_c - u_b v_b}{N}\right] \times
$$

$$
\times \exp\left[-i2\pi \frac{(l+u_c-u_a)(v_c+(M-1)/2)}{N}\right]. \tag{3.45б}
$$

При $r_a = 0;\ v_a = v_c$:

$$ {}^{r_a, r_b}C_k^{u_c, v_c} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} a_n^{u_a, v_a} \tilde{b}_{k-n} \qquad (3.45в) $$

Таким образом, перемножив сдвинутые спектры СДПФ двух последовательностей и выполнив обратное СДПФ, получим последовательность, являющуюся обобщённой циклической сверткой исходных последовательностей. Сворачиваемые последовательности комплексных чисел являются здесь обобщенно-циклическими в соответствии с соотношениями (3.38).

Особое место среди разновидностей СДПФ занимает так называемое *косинусное преобразование [1]:*

$$ \begin{cases} \alpha_r = \dfrac{2}{\sqrt{2N}} \sum\limits_{k=0}^{N-1} a_k \cos\left(\pi \dfrac{(k+1/2)\, r}{N}\right); \quad r = 1, 2, \ldots, N-1 \\ \alpha_0 = \dfrac{1}{\sqrt{2N}} \sum\limits_{k=0}^{N-1} a_k \end{cases} \qquad (3.46) $$

являющееся СДПФ (1/2, 0) четной последовательности, определяемой соотношениями

$$ a_k = a_{2N-1-k}; \quad a_{k+2Nh} = a_k; \quad k = 0, 1, \ldots, 2N-1 \qquad (3.47) $$

Другой известной разновидностью СДПФ является так называемое *«синусное преобразование»* [64]:

(3.48а)

$$ \alpha_r = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} a_k \sin\left[\pi \frac{(k+1)(r+1)}{N+1}\right] $$

$$ a_k = \sqrt{\frac{2}{N}} \sum_{r=0}^{N-1} \alpha_r \sin\left[\pi \frac{(k+1)(r+1)}{N+1}\right] \qquad (3.48б) $$

которое, как нетрудно видеть, является СДПФ (1,1), нечетным образом продолженной по длине (2N+2) последовательности длиной N членов:

$$ a_k = -a_{2N-k}; \quad a_N = a_{2N+1} = 0 \qquad (3.49) $$

Двумерные СДПФ определяются аналогично одномерным:

$$ \alpha_{r,s}^{u,v;w,z} = \frac{1}{\sqrt{N_1 N_2}} \left\{ \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k,l} \exp\left[i 2\pi \left(\frac{k(r+v)}{N_1} + \right.\right.\right. $$

$$ \left.\left.\left. + \frac{l(s+z)}{N_2} \right) \right] \right\} \exp\left[i 2\pi \left(\frac{ru}{N_1} + \frac{sw}{N_2} \right) \right] \qquad (3.50а) $$

$$ a_{k,l}^{u,v;w,z} = \frac{1}{\sqrt{N_1 N_2}} \left\{ \sum_{r=0}^{N_1-1} \sum_{s=0}^{N_2-1} \alpha_{r,s}^{u,v;w,z} \exp\left[-i 2\pi \left(\frac{r(k+u)}{N_1} + \right.\right.\right. $$

$$ \left.\left.\left. + \frac{s(l+w)}{N_2} \right) \right] \right\} \exp\left[-i 2\pi \left(\frac{kv}{N_1} + \frac{lz}{N_2} \right) \right] \qquad (3.50б) $$

Вследствие разделимости двумерных СДПФ их свойства легко вытекают из свойств одномерных СДПФ.

В двумерном случае особый интерес представляют СДПФ, построенные для дискретизации сигналов и/или спектров по шестиугольному (гексагональному) растру. При гексагональном растре сетка отсчетов по одной координате сдвигается на половину расстояния между отсчетами для каждого нечетного номера отсчета но другой координате (см. рис. 2.5,г).

Поэтому двумерное СДПФ для расположения отсчетов сигнала и спектра по гексагональному растру определяется как

$$\alpha_{r,s}^{r,r} = \frac{1}{\sqrt{N_1 N_2}} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k,l} \times$$
$$\times \exp\left\{ i 2\pi \left[\frac{\left(k + \frac{1+(-1)^l}{4}\right)\left(r + \frac{1+(-1)^s}{4}\right)}{N_1} + \frac{ls}{N_2} \right] \right\} \qquad (3.51)$$

а при гексагональном растре только сигнала или только спектра соответственно, как:

$$\alpha_{r,s}^{r,n} = \frac{1}{\sqrt{N_1 N_2}} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k,l} \exp\left\{ i 2\pi \left[\frac{\left(k + \frac{1+(-1)^l}{4}\right) r}{N_1} + \frac{ls}{N_2} \right] \right\} \qquad (3.52)$$

$$\alpha_{r,s}^{n,r} = \frac{1}{\sqrt{N_1 N_2}} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k,l} \exp\left\{ i 2\pi \left[\frac{k\left(r + \frac{1+(-1)^s}{4}\right)}{N_1} + \frac{ls}{N_2} \right] \right\} \qquad (3.53)$$

Из этих формул видно, что от стандартного двумерного ДПФ СДПФ на гексагональном растре отличаются тем, что для каждой четной строки одно из одномерных преобразований, на которые распадаются ДПФ и СДПФ, являются СДПФ (1/2, 1/2), СДПФ (1/2,0) или СДПФ (0, 1/2) соответственно для гексагонального растра сигнала и спектра [формула (3.51)], гексагонального растра сигнала и прямоугольного растра спектра [формула (3.52)] и прямоугольного растра сигнала и гексагонального растра спектра [формула (3.53)].

Гексагональные дискретные спектры и сигналы связаны с прямоугольными спектрами и сигналами интерполяционными соотношениями, подобными (3.43). Так, отсчеты гексагонального спектра, заданного на прямоугольном растре, связаны с отсчетами его прямоугольного спектра соотношением

$$\alpha_{r,s}^{n,r} = \sum_{p=0}^{N-1} \alpha_{r,s}^{n,n} \operatorname{sincd}\left[N;\ \pi\left(r - p + \frac{1+(-1)^s}{4} \right) \right] \times$$
$$\times \exp\left[i\pi \frac{N-1}{N} \left(r - p + \frac{1+(-1)^s}{4} \right) \right]. \qquad (3.54)$$

3.4. ДИСКРЕТНЫЕ ПРЕОБРАЗОВАНИЯ ФРЕНЕЛЯ

Преобразование Френеля (1.31) связано с преобразованием Фурье (см. § 1.2), в силу чего при построении дискретного представления преобразования Френеля можно воспользоваться дискретным представлением преобразования Фурье. При этом особенно важно учитывать возможный сдвиг растра отсчетов сигнала $a(x)\exp(-i\pi x^2/D^2)$ и его фурье-спектра $\alpha_D(f)\exp(i\pi f^2/D^2)$ относительно начала координат сигнала и спектра, которые являются центрами симметрии экспоненциальных множителей $\exp(-i\pi x^2/D^2)$ и $\exp(i\pi f^2/D^2)$
Таким образом, для базисов отсчетных функций

$$\{\varphi_k(x) = \operatorname{sinc}[\pi(x - (k+u)\Delta x)/\Delta x]\}$$
$$\{\chi_r(f) = \operatorname{sinc}[\pi(f - (r+v)\Delta f)/\Delta f]\} \qquad (3.55)$$

(где *u, v, как* и выше, — сдвиг растра отсчетов относительно начала координат по *x* и *f*) получим следующее дискретное представление прямого и обратного преобразований Френеля:

$$\alpha_r^x = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k \exp\left\{ -i \frac{\pi}{N} \left[(k+u)\varkappa - \frac{r+v}{\varkappa} \right]^2 \right\} \qquad (3.56a)$$

$$a_k = \frac{1}{\sqrt{N}} \sum_{r=0}^{N-1} \alpha_r^{\varkappa} \exp\left\{ i \frac{\pi}{N} \left[(k+u)\varkappa - \frac{r+v}{\varkappa} \right]^2 \right\} \tag{3.56б}$$

где $N = 1/\Delta x \Delta f D^2$; $\{a_k\}$—последовательность отсчетов сигнала:

$$a_k = \frac{1}{\Delta x} \exp\left[i \pi \frac{(k+u)^2 \varkappa^2}{N} \right] \times$$
$$\times \int_{-\infty}^{\infty} a(x) \exp(-i \pi x^2/z^2) \operatorname{sinc}[\pi (x-(k+u)\Delta x)/\Delta x]\, dx \tag{3.57}$$

$\{\alpha_r^{\varkappa}\}$ — последовательность отсчетов спектра Френеля:

$$\alpha_r^{\varkappa} = \frac{1}{\Delta f} \exp\left[-i \pi \frac{(r+v)}{\varkappa^2} \right] \int_{-\infty}^{\infty} a_D(f) \times$$
$$\times \exp(i \pi f^2/D^2) \operatorname{sinc}[\pi (f-(r+v)\Delta f)/\Delta f]\, df \tag{3.58}$$

Δx и Δf—интервалы дискретизации сигнала и его спектра; $\varkappa$ — безразмерный параметр:

$$\varkappa = \sqrt{\Delta f/\Delta x}. \tag{3.59}$$

Это представление предполагает следующую аппроксимацию непрерывного сигнала и его спектра Френеля:

$$a(x) \approx \exp\left(i \pi \frac{x^2}{D^2} \right) \sum_{k=0}^{N-1} a_k \exp\left[-i \pi \frac{(k+u)^2 \varkappa^2}{N} \right] \times$$
$$\times \operatorname{sinc}[\pi (x-(k+u)\Delta x)/\Delta x] \tag{3.60а}$$

$$a_D(f) \approx \exp\left(-i \pi \frac{f^2}{D^2} \right) \sum_{r=0}^{N-1} \alpha_r^{\varkappa} \exp\left[i \pi \frac{(r+v)^2}{\varkappa^2} \right] \times$$
$$\times \operatorname{sinc}[\pi (f-(r+v)\Delta f)/\Delta f]. \tag{3.60б}$$

Преобразования (3.56) зависят от трех параметров: u, v и $\varkappa$. Однако поскольку параметры сдвига входят аддитивно, их можно заменить одним параметром, переписав (3.56) в виде:

$$\alpha_r^{\varkappa, w} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k \exp\left[-i \frac{\pi}{N} \left(k\varkappa - \frac{r}{\varkappa} + w \right)^2 \right] \tag{3.61 а}$$

$$a_k = \frac{1}{\sqrt{N}} \sum_{r=0}^{N-1} \alpha_r^{\varkappa, w} \exp\left[i \frac{\pi}{N} \left(k\varkappa - \frac{r}{\varkappa} + w \right)^2 \right] \tag{3.61б}$$

где ω — общий сдвиг.

Будем называть пару преобразований (3.61) соответственно прямым и обратным *дискретными преобразованиями Френеля* (ДПФР).

Очевидно, дискретное преобразование Френеля вычисляется через СДПФ ($w\varkappa$, 0):

$$\alpha_r^{\varkappa, w} = \frac{1}{\sqrt{N}} \left\{ \sum_{k=0}^{N-1} a_k \exp\left[-i \frac{\pi}{N} (k\varkappa + w)^2 \right] \times \right.$$
$$\left. \times \exp\left[i \frac{2\pi}{N} (k + w\varkappa) r \right] \right\} \exp\left(-i \frac{\pi r^2}{N\varkappa^2} \right).$$

Важнейшие свойства ДПФР приведены в табл. 3.1. Для анализа этих свойств удобно ввести периодическую (с периодом JV) функцию (см. табл. 3.1):

$$\mathrm{frinc}\,(N;\ q;\ r) = \frac{1}{N}\sum_{k=0}^{N-1} \exp\left(\mathrm{i}\,\frac{\pi}{N}\,qk^2\right)\exp\left(-\mathrm{i}\,\frac{2\pi}{N}\,kr\right) \tag{3.62}$$

Свойства функции frinc (N; *q; r):*

$$\mathrm{frinc}\,(N;\ q;\ r) = \frac{1}{N}\sum_{k=0}^{N-1} \exp\left(\mathrm{i}\,\frac{\pi}{N}\,k^2 q\right)\exp\left(-\mathrm{i}\,\frac{2\pi}{N}\,kr\right);$$

$$\mathrm{frinc}\,(N;\ q;\ r) = \mathrm{frinc}\,(N;\ q;\ (r) \bmod N);$$

$$(\mathrm{frinc}\ (N;\ q;\ r))^* = \mathrm{frinc}\,(N;\ -q;\ N-r);$$

$$\mathrm{frinc}\,(N;\ q;\ N-r) = \mathrm{frinc}\,(N;\ -q;\ r+(N-1)\,q)\times$$

$$\times \exp\{\mathrm{i}\,\pi\,(N-1)\,[(N-1)\,q+2r]/N\};$$

$$\mathrm{frinc}\,(N;\ 0;\ r) = \mathrm{sincd}\left(N;\ \pi\,\frac{r}{N}\right)\exp\left(-\mathrm{i}\,\pi\,\frac{N-1}{N}\,r\right);$$

$$\lim_{Q\to\infty}\frac{1}{2Q}\int_{-Q}^{Q}\mathrm{frinc}\,(N;\ q;\ r)\,dq = 1/N;$$

$$\int_{-\infty}^{\infty}\mathrm{frinc}\,(N;\ q;\ r)\exp(\mathrm{i}\,2\pi\sigma q)\,dq = \left\{\frac{1}{N}\sum_{k=0}^{N-1}\exp\left(-\mathrm{i}\,\frac{2\pi}{N}\,kr\right)\times\right.$$

$$\left.\times\,\delta\left(\left(\frac{k^2}{2}+\sigma\right)/N\right)\right\}\mathrm{rect}\left(\frac{\sigma+(N-1)^2/N}{(N-1)^2/N}\right).$$

Из этих свойств можно заключить, что для ДПФР frinc(N; *q; r)* играет ту же роль, что и sincd(N; *x)* (3.12) для ДПФ. Нетрудно усмотреть также связь между функцией frinc(N; *q; r)* и специальными функциями, получившими название интегралов Френеля ([45]):

$$F_r^{\pm}(z) = C(z) \pm \mathrm{i}\,S(z) = \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{z}}\exp(\pm\,\mathrm{i}\,t^2)\,dt.$$

Таблица 3.1. *Свойства дискретного преобразования Френеля* Сигнал ДПФР

Сигнал	ДПФР
$a_k = \frac{1}{\sqrt{N}}\sum_{r=0}^{N-1} a_r^{\varkappa,\,w}\times$ $\times\exp\left[\mathrm{i}\,\frac{\pi}{N}\left(k\varkappa-\frac{r}{\varkappa}+w\right)^2\right]$	$a_r^{\varkappa,\,w} = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} a_k\times$ $\times\exp\left[-\mathrm{i}\,\frac{\pi}{N}\left(k\varkappa-\frac{r}{\varkappa}+w\right)^2\right]$
$a_{k+k_0}\exp\left[-\mathrm{i}\,\frac{2\pi}{N}\left(k\varkappa+w+\frac{k_0\varkappa}{2}\right)k_0\varkappa\right]$	$a_r^{\varkappa,\,w}\exp\left(-\mathrm{i}\,2\pi\,\frac{k_0 r}{N}\right)$
$a_{k+Nh} = a_k\exp\left[\mathrm{i}\,2\pi\left(k\varkappa+\right.\right.$ $\left.\left.+w+\frac{Nh\varkappa}{2}\right)h\varkappa\right]$	$a_{r+gN}^{\varkappa,\,w} = a_r^{\varkappa,\,w}\exp\left[-\mathrm{i}2\pi\left(\frac{r}{\varkappa}+\right.\right.$ $\left.\left.+w+\frac{Ng}{2\varkappa}\right)\frac{g}{\varkappa}\right]$
δ_{k-k_0}	$\frac{1}{\sqrt{N}}\exp\left[-\mathrm{i}\,\frac{\pi}{N}\left(k_0\varkappa-\frac{r}{\varkappa}+w\right)^2\right]$
$\frac{1}{\sqrt{N}}\exp\left[\mathrm{i}\,\frac{\pi}{N}\left(k\varkappa-\frac{r_0}{\varkappa}+w\right)^2\right]$	$\delta\,(r-r_0)$

$a_k^* \exp\left\{i\frac{2\pi}{N}\left[\left(k-\frac{N-1}{2}\right)\varkappa+\right.\right.$ $\left.\left.+(w+w_0)]^2\right\}$, $w_0=(N-1)(\varkappa-1/\varkappa)/2$	$a^*_{N-1-r}\exp\left[-i\frac{2\pi}{N\varkappa^2}\left(r-\frac{N-1}{2}\right)^2\right]$
$a_{N-1-k}\exp\left[i\frac{4\pi\varkappa}{N}\left(k-\frac{N-1}{2}\right)(w+w_0)\right]$	$a_{N-1-r}\exp\left[-i\frac{4\pi}{N\varkappa}\left(r-\frac{N-1}{2}\right)\times\right.$ $\left.\times(w+w_0)\right]$

Сигнал	ДПФР
1	$\sqrt{N}\,\mathrm{frinc}\,(N;\ -\varkappa^2;\ r-\varkappa[w+$ $+\varkappa(N-1)])\exp\left\{-i\frac{\pi}{N}\left[\frac{r}{\varkappa}-(w+\right.\right.$ $\left.\left.+(N-1)\varkappa)\right]^2\right\}$
$\cos[\pi(k\nu+\zeta)^2/N]$	$\frac{1}{2}\sqrt{N}\,\mathrm{frinc}\,(N;\ \nu^2-\varkappa^2;\ r-\varkappa w_x+$ $+\nu w_y)\exp\left\{i\frac{\pi}{N}\left[w_y^2-\left(\frac{r}{\varkappa}-w\varkappa\right)^2\right]\right\}+$ $+\frac{1}{2}\sqrt{N}\,\mathrm{frinc}\,(N;\ -\nu^2-\varkappa^2;\ r-\varkappa w_x-$ $-\nu w_y)\exp\left\{-i\frac{\pi}{N}\left[w_y^2+\left(\frac{r}{\varkappa}-w\varkappa\right)^2\right]\right\};$ $w_x=\varkappa(N-1)+w;$ $w_y=\nu(N-1)+w$
$\cos[\pi(k\varkappa^2+\zeta)^2/N]$	$\frac{1}{2}\sqrt{N}\,\mathrm{sincd}\left[N;\ \pi\frac{r-\varkappa(w-\zeta)}{N}\right]\times$ $\times\exp\left\{-i\frac{\pi}{N}[r-\varkappa(w-\zeta)](N-1)+\right.$ $\left.+\frac{r-\varkappa(w+\zeta)}{\varkappa^2}\right\}+\frac{1}{2}\sqrt{N}\times$ $\times\,\mathrm{frinc}\,(N;\ -2\varkappa^2;\ r-2\varkappa^2(N-1)+$ $+\varkappa(w+\zeta))\exp\left\{-i\frac{\pi}{N}[(\varkappa(N-1)+\right.$ $\left.+\zeta)^2+(r/\varkappa-\varkappa(N-1)-w)^2]\right\}$

$$r,w a_k^{\nu,\varkappa}=\frac{1}{\sqrt{N}}\sum_{r=0}^{N-1}{}^{1}\alpha_r^{\nu,w}\exp\left[i\frac{\pi}{N}\left(k\nu-\frac{r}{\nu}+z\right)^2\right]=\left\{\sum_{n=0}^{N-1}a_n^{\nu,w}\exp\left[-i\frac{\pi}{N}(n\varkappa+\right.\right.$$
$$\left.\left.+w)^2\right]\mathrm{frinc}\,(N;\ q;\ k-n-p)\right\}\exp\left[i\frac{\pi}{N}(k\nu+z)^2\right];$$
$$q=1/\nu^2+1/\varkappa^2;\quad p=w/\varkappa-z/\nu.$$

$$\nu,w a_k^{\nu,\varkappa}=\left\{\sum_{n=0}^{N-1}a_n^{\nu,w}\exp\left[-i\frac{\pi}{N}(n\varkappa+w)^2\right]\mathrm{sincd}\left(N;\ \pi\frac{k-n-(w-z)/\varkappa}{N}\right)\times\right.$$
$$\left.\times\exp\left[-i\pi\frac{N-1}{N}\left(k-n-\frac{w-z}{\varkappa}\right)\right]\right\}\exp\left[i\frac{\pi}{N}(k\varkappa+z)^2\right].$$

Действительно, при N→ ∞

$$\lim_{N\to\infty} \mathrm{frinc}\,(N;\ q;\ r) = \int_0^1 \exp\left[-i\pi\frac{(r-p)^2}{q_0}\right]\times$$
$$\times \exp\left[i\pi q_0\left(x-\frac{r-p}{q_0}\right)^2\right]dx = \frac{2}{\sqrt{2q_0}}\times$$
$$\times \exp\left[-i\pi\frac{(r-p)^2}{q_0}\right]\left\{\mathrm{Fr}^{\mathrm{sign}\,q_0}\left(\pi|q_0|\left(1-\frac{(r-p)}{q_0}\right)^2\right)-\right.$$
$$\left.-\mathrm{Fr}^{\mathrm{sign}\,q_0}\left(\frac{\pi}{|q_0|}(r-p)^2\right)\right\},$$

где $x = \lim_{N\to\infty}(k/N)$; $q_0 = \lim_{N\to\infty} qN$ — безразмерная величина, меняющаяся от -∞ до ∞.

Дискретное представление усеченного преобразования Френеля, очевидно, является усеченным вариантом ДПФР:

$$\hat{a}_r^{x,\,w} = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} a_k \exp\left[-i\frac{\pi}{N}(kx+w)^2\right]\exp\left(i\frac{2\pi}{N}kr\right) \qquad (3.63а)$$

Будем называть это преобразование *частичным дискретным преобразованием Френеля* (ЧДПФР). Ему, как нетрудно проверить, соответствует обратное преобразование:

$$a_k = \frac{1}{\sqrt{N}}\exp\left[i\frac{\pi}{N}(kx+w)^2\right]\sum_{r=0}^{N-1}\hat{a}_r^{x,\,w}\exp\left(-i2\pi\frac{kr}{N}\right) \qquad (3.63б)$$

Связь ЧДПФР с ДПФ очевидна. Некоторые свойства ЧДПФР приведены в табл. 3.2.

Таблица 3.2. *Свойства частичного дискретного преобразования Френеля*

Сигнал	ЧДПФР
$a_k = \frac{1}{\sqrt{N}}\exp\left[i\frac{\pi}{N}(kx+w)^2\right]\sum_{r=0}^{N-1}\hat{a}_r^{x,\,w}\times \times\exp\left(-i2\pi\frac{kr}{N}\right)$	$\hat{a}_r^{x,\,w} = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1} a_k\exp\left[-i\frac{\pi}{N}(kx+w)^2\right]\times \times\exp\left(i\frac{2\pi}{N}kr\right)$
$a_k\exp\left(i\frac{2\pi}{N}kr_0\right)$	$\hat{a}_{r+r_0}^{x,\,w}$
—	$\hat{a}_r^{x,\,w} = \hat{a}_{(r)\bmod N}^{x,\,w}$

Сигнал	ЧДПФР
δ_{k-k_0}	$\exp\left(i\frac{2\pi}{N}k_0r\right)\exp\left[-i\frac{\pi}{N}(k_0x+w)^2\right]$
$\exp\left(-i2\pi\frac{kr_0}{N}\right)\times \times\exp\left[i\frac{\pi}{N}(kx+w)^2\right]$	δ_{r-r_0}
$a_k^*\exp\left[i\frac{2\pi}{N}(kx+w)^2\right]$	$\left(\hat{a}_{N-r}^{x,\,w}\right)^*$
$a_{N-k}\exp\left[i\frac{\pi}{N}(2w+ +Nx)(2kx+Nx)\right]$	$\hat{a}_{N-r}^{x,\,w}$
1	$\sqrt{N}\,\mathrm{frinc}\,(N;\ -x^2;\ r-w_x x)\times \times\exp\left\{i\frac{\pi}{N}[2r(N-1)-w_x^2]\right\}$; $w_x = w+(N-1)x$

$\cos\left[\frac{\pi}{N}(k\nu+z)^2\right]$	$\frac{\sqrt{N}}{2}\,\mathrm{frinc}\,(N;\ \nu^2-\varkappa^2;\ r-\varkappa w_x+\nu w_y)\times$ $\times\exp\left\{i\frac{\pi}{N}[w_y^2-w_x^2+2r(N-1)]\right\}+$ $+\frac{\sqrt{N}}{2}\,\mathrm{frinc}\,(N;\ -(\varkappa^2+\nu^2);\ r-\varkappa w_x-\nu w_y)\times$ $\times\exp\left\{-i\frac{\pi}{N}[w_y^2+w_x^2-2r(N-1)]\right\};$ $w_x=w+\varkappa(N-1);\quad w_y=z+\nu(N-1)$
$\cos\left[\frac{\pi}{N}(k\varkappa+z)^2\right]$	$\frac{\sqrt{N}}{2}\,\mathrm{sincd}\left(N;\ \pi\frac{r-\varkappa(w-z)}{N}\right)\times$ $\times\exp\left[i\frac{2\pi\varkappa}{N}(w-z)(N-1)\right]+\frac{\sqrt{N}}{2}\,\mathrm{frinc}\,(N;\ 2\varkappa^2;$ $r-\varkappa[(w+z)+2(N-1)\varkappa])\exp\left\{-i\frac{\pi}{N}[[\varkappa(N-1)+$ $+w]^2+[\varkappa(N-1)+z]^2-2r(N-1)]\right\}$

$$w,\ z\hat{\alpha}^{\varkappa,\ \nu}=\sum_{s=0}^{N-1}z\hat{\alpha}_s^{\nu}\,\mathrm{frinc}\,(N;\ \nu^2-\varkappa^2;\ r-s-\varkappa w+\nu z)\exp\left[i\pi(z^2-w^2)/N\right]$$

Глава 4
ЭФФЕКТИВНЫЕ ВЫЧИСЛИТЕЛЬНЫЕ ПРОЦЕДУРЫ ЦИФРОВОЙ ФИЛЬТРАЦИИ

4.1. МЕТОДЫ ВЫЧИСЛЕНИЙ ДИСКРЕТНЫХ ПРЕОБРАЗОВАНИЙ ФУРЬЕ

Понятие об алгоритмах быстрого преобразования Фурье. Рассмотрим дискретное преобразование Фурье:

$$a_r = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k \exp\left(i 2\pi \frac{kr}{N}\right) \qquad (4.1)$$

последовательности $\{a_k\}$, k = 0, 1, ..., N—1. Если выполнять вычисления непосредственно по этой формуле, то для нахождения всех N коэффициентов $\{a_r\}$ необходимо проделать примерно N^2 комплексных операций (N^2 умножений и $N(N—1)$ комплексных сложений). Это число может оказаться очень большим. Практическое значение при цифровой обработке сигналов ДПФ приобрело только после изобретения так называемых быстрых алгоритмов преобразования Фурье (БПФ) [1].

Возможность сокращения числа операций при вычислении ДПФ станет очевидной, если рассмотреть двумерное ДПФ. Действительно, для двумерного массива размера N_1N_2 благодаря тому, что двумерное ДПФ факторизуется на два одномерных, количество операций равно приблизительно $N_1^2N_2$ для ДПФ по строкам плюс $N_1N_2^2$ для ДПФ по столбцам, т.е. $N_1N_2(N_1+N_2)$ операций, а не $N_1^2N_2^2 = N_1N_2 \times N_1N_2$, как это было бы для одномерного массива того же объема. Очевидно, для n-мерного массива размерностью $N_1N_2, ..., N_n$ количество требуемых операций будет примерно порядка $\left(\prod_{p=1}^{n} N_p\right)\left(\sum_{p=1}^{n} N_p\right)$, т.е. в $\prod_{p=1}^{n} N_p / \sum_{p=1}^{n} N_p$ раз меньше, чем для одномерного массива того же объема. Таким образом, если представить одномерное ДПФ в виде многомерного, то число операций резко сокращается. Такое представление возможно, если объем массива является составным числом:

$$N = N_1N_2 \ldots N_n$$

Разберем эту возможность на примере двух сомножителей:

$$N = N_1N_2$$

Представим индексы k и r в системе счисления с основаниями N_1 и N_2:

$$k = k_2N_1 + k_1; \; r = r_1N_2 + r_2, \qquad (4.3)$$

где $k_1, r_1 = 0, 1, ..., N_1—1$; $k_2, r_2 = 0, 1, ..., N_2—1$. Подставив (4.3) в (4.1), получим

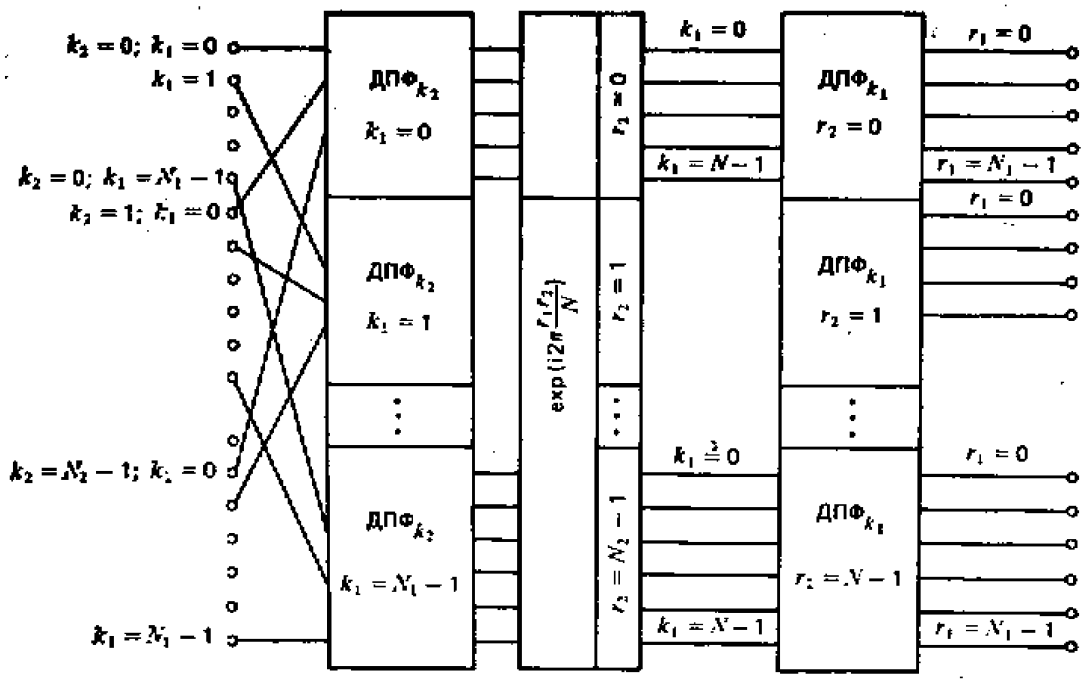


Рис. 4.1. Принцип БПФ

$$a_{r_1, r_2} = \frac{1}{\sqrt{N_1 N_2}} \sum_{k_2=0}^{N_2-1} \sum_{k_1=0}^{N_1-1} a_{k_2, k_1} \times$$

$$\times \exp\left[i 2\pi \frac{(k_2 N_1 + k_1)(r_1 N_2 + r_2)}{N_2 N_1} \right] =$$

$$= \frac{1}{\sqrt{N_1 N_2}} \sum_{k_1=0}^{N_1-1} \exp\left[i 2\pi \left(\frac{k_1 r_1}{N_1} + \frac{k_1 r_2}{N_1 N_2} \right) \right] \times$$

$$\times \sum_{k_2=0}^{N_2-1} a_{k_2, k_1} \exp\left(i 2\pi \frac{k_2 r_2}{N_2} \right) =$$

$$= \text{ДПФ}_{k_1} \left\{ \exp\left(i 2\pi \frac{k_1 r_2}{N} \right) \text{ ДПФ}_{k_2} \{a_{k_2, k_1}\} \right\}.$$

Таким образом, исходное ДПФ оказалось сведенным к двум ДПФ, производимым над уменьшенными массивами (рис. 4.1). Заметим, что при такой организации вычислений их результат получается в так называемом инверсном порядке, определенном инверсией разрядов в позиционно-численном представлении их номеров по сложному основанию (4.2).

Очевидно, что требуемое при таком способе вычислений ДПФ $\{a_{k2,k1}\}$ количество операций умножения равно $N(N_1+N_2)$, сложения — $N(N_1+N_2—2)$ вместо $N^2 = NN_1N_2$ умножений и $N(N_1\ N_2—1)$ сложений при прямом вычислении (4.1). В общем случае, если выполняется (4.2), такая процедура позволяет уменьшить общее число операций при вычислении ДПФ до величины порядка $N \sum_{p=1}^{n} N_p$ вместо $N^2 = N \prod_{p=1}^{n} N_p$. Если $N = 2^n$, то количество требуемых операций уменьшается с N^2 до $2Nn = 2N \log_2 N$, т.е. при больших N уменьшается в десятки — сотни раз. Алгоритмы ДПФ, основанные на описанном принципе, и называются алгоритмами БПФ.

В настоящее время имеется большое разнообразие алгоритмов БПФ, отличающихся различными способами упорядочивания входного массива чисел и результата преобразования, а также способами использования внешних запоминающих устройств. Общая теория таких алгоритмов будет рассмотрена в гл. 5.

Совмещенные алгоритмы. ДПФ последовательностей вещественных чисел обладает двукратной избыточностью: коэффициенты Фурье с номерами, дополняющими друг друга до длины последовательности, являются комплексно-сопряженными числами:

$$\alpha_r = \alpha^*_{N-r} \qquad (4.4)$$

Это означает, что достаточно вычислить α_r только для $r = 0, 1, ..., N/2$, а остальные α_r можно найти без вычислений по формуле (4.4). Это обстоятельство можно использовать для сокращения примерно вдвое числа операций, требуемых для вычисления ДПФ последовательностей вещественных чисел.

Известны два способа реализации указанного выигрыша: совмещенное преобразование двух последовательностей и разбиение одной последовательности с четным числом членов на две подпоследовательности, совмещение преобразования этих подпоследовательностей и пересчет результата на всю последовательность.

Первый способ состоит в том, что из двух последовательностей $\{a_k\}$ и $\{b_k\}$ вещественных чисел длиной N ($k = 0, 1, 2, ..., N—1$) образуют последовательность

$$c_k — a_k + ib_k$$

и вычисляют ее ДПФ

$$\gamma_r = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} (a_k + i b_k) \exp\left(i 2\pi \frac{kr}{N}\right),$$

а ДПФ $\{\alpha_r\}$, $\{\beta_r\}$ исходных последовательностей находят, пользуясь соотношениями:

$$\alpha_r = (\gamma_r + \gamma^*_{N-r})/2; \quad \beta_r = -i(\gamma_r - \gamma^*_{N-r})/2 \qquad (4.5)$$

Таким образом, выполнив $(N+2)$ сложения вещественных чисел, можно вычислить все коэффициенты ДПФ последовательностей $\{a_k\}$ и $\{b_k\}$ из результата преобразования совмещенной последовательности $\{a_k+ib_k\}$, причем вычисления по (4,5) достаточно выполнять только для $r = 0, 1, ..., N/2$, а остальные α_r,иβ_r можно найти по (4.4).

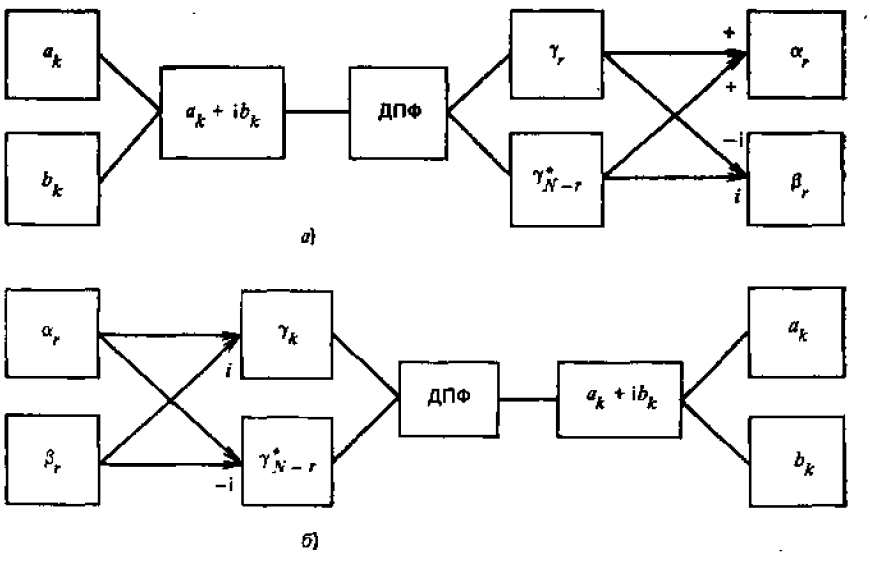


Рис. 4.2, Совмещенные алгоритмы ДПФ

Описанная процедура вычислений иллюстрируется рис. 4.2, *а.* Такой способ совмещенного ДПФ целесообразно применять, например, при преобразовании двумерных массивов (изображений), когда в качестве $\{a_k\}$ и $\{b_k\}$ удобно выбирать соседние пары строк массива.

При преобразовании одномерных последовательностей вещественных чисел используется второй способ сокращения числа операций, вытекающий из следующего соотношения для ДПФ последовательности $\{a_k\}$ вещественных чисел длиной $2N$:

$$\alpha_r = \frac{1}{\sqrt{2N}} \sum_{k=0}^{N-1} a_k \exp\left(i 2\pi \frac{kr}{2N}\right) =$$
$$= \frac{1}{\sqrt{2N}} \left\{ \sum_{k=0}^{N-1} a_{2k} \exp\left(i 2\pi \frac{kr}{N}\right) + \right.$$
$$\left. + \left[\sum_{k=0}^{N-1} a_{2k+1} \exp\left(i 2\pi \frac{kr}{N}\right) \right] \exp\left(i \pi \frac{r}{N}\right) \right. \tag{4.6}$$

Таким образом, ДПФ всей последовательности $\{a_k\}$ можно выполнить, вычислив ДПФ от двух ее подпоследовательностей, составленных соответственно из четных и нечетных членов исходной последовательности, и затем просуммировав полученные результаты по формуле

$$\alpha_r = \frac{1}{\sqrt{2}} [\alpha_r^{(ч)} + \exp(i \pi r/N) \alpha_r^{(н)}]$$

где $\alpha_r^{(ч)}$ и $\alpha_r^{(н)}$ — ДПФ от $\{\alpha_{2k}\}$, $\{\alpha_{2k+1}\}$ соответственно.

Коэффициенты $\alpha_r^{(ч)}$ и $\alpha_r^{(н)}$ можно найти, пользуясь первым способом совмещенного ДПФ по формулам (4.5). Эти алгоритмы совмещенного преобразования Фурье легко обратить и получить алгоритмы вычисления ДПФ последовательностей, элементы которых, симметричные относительно ее середины, являются комплексно-сопряженными числами, как в (4.4). Граф такого алгоритма для двух последовательностей $\{\alpha_r=\alpha_{N-r}^*\}$, $\{\beta_r=\beta_{N-r}^*\}$показан на рис. 4.2, *б*.

Использование алгоритмов .совмещенных преобразований возможно и для вычисления двумерного преобразования Фурье массивов вещественных или комплексно-сопряженных чисел, выполняемых как два одномерных преобразования. В самом деле, при преобразовании двумерных массивов вещественных чисел первое преобразование Фурье можно выполнять, совмещая ДПФ пар строк массива, а второе преобразование Фурье полученного массива комплексных чисел выполнять только до половины столбцов, вторую же половину находить, не вычисляя, как комплексно-сопряженную первой. При преобразовании комплексно-сопряженных массивов нужно поступать в обратном порядке: первое преобразование Фурье выполнять только до половины массива, дополнив его числами, комплексно-сопряженными с результатом первого преобразования, после чего второе преобразование Фурье выполнять с помощью описанного алгоритма совмещенного преобразования двух последовательностей с попарно комплексно-сопряженными элементами.

Совмещенный алгоритм косинусного преобразования. Как было указано в § 3.3,. косинусное преобразование сводится к СДПФ (1/2,0) четных последовательностей вида:

$\{a_k — a_{2N-1-k}\}$, $k = 0,$ 1, ... 2N – 1.

В соответствии со свойствами СДПФ (§ 3.3) для таких последовательностей выполняются следующие соотношения:

$$\{a_k = a_{2N-1-k}\} \leftarrow \text{СДПФ} \rightarrow \{\gamma_r^{1/2,0} = -\alpha_{2N-r}^{1/2,0} = -\alpha_{2N+r}^{1/2,0}\};$$
$$\left\{a_k \exp\left[i 2\pi \frac{N(k+1/2)}{N}\right] = i(-1)^k a_k\right\} \leftarrow \text{СДПФ} \rightarrow \{\alpha_{r+N}^{1/2,0}\}$$

Их можно использовать для построения совмещенного алгоритма косинусного преобразования [46]. Алгоритм состоит в том, что из двух четных последовательностей с четным числом членов:

$a_k=a_{2N-1-k}$; $b_k=b_{2N-1-k}$; k=0,1,…,2N-1

образуется последовательность

$$c_k = a_k + i(-1)^k b_k$$

и вычисляется ее СДПФ (1/2,0) $\gamma_r^{1/2,0}$, а коэффициенты $\{\alpha_r^{1/2,0}\}$ и $\{\beta_r^{1/2,0}\}$ косинусного преобразования исходных последовательностей находятся по $\{\gamma_r^{1/2,0}\}$ для r = 0, 1, ..., N—1 из следующих соотношений:

$$\alpha_r^{1/2,0} = (\gamma_r^{1/2,0} - \gamma_{2N-r}^{1/2,0})/2; \quad \beta_r^{1/2,0} = -(\gamma_{N+r}^{1/2,0} + \gamma_{N-r}^{1/2,0})/2$$

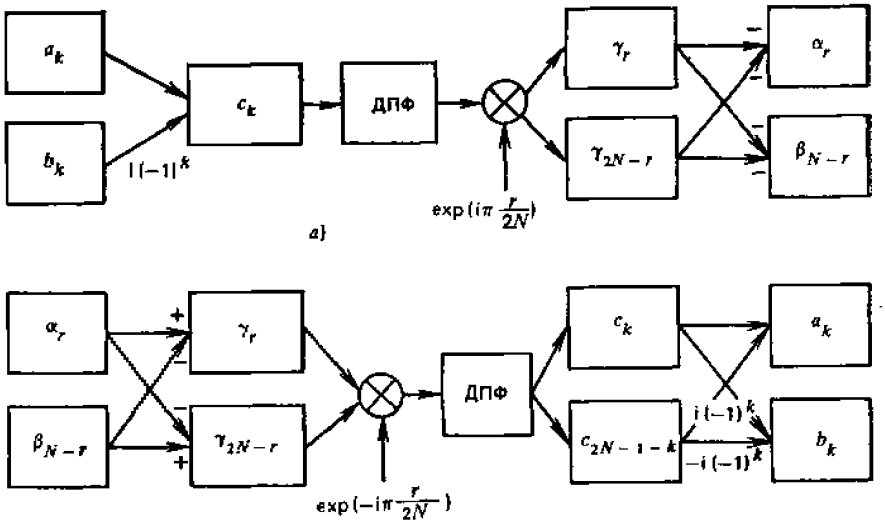


Рис. -4.3. Графы совмещенных алгоритмов преобразования СДПФ (1/2,0) четных последовательностей: *а* — прямое, *б* — обратное преобразование

В результате такого совмещения затраты времени на преобразование четных последовательностей ненамного превышают затраты на преобразование последовательностей половинной длины. Описанная процедура иллюстрируется на рис. 4.3, а. Ее нетрудно обратить и получить процедуру обратного преобразования (рис. 4.3, б).

Косинусное преобразование—преобразование по вещественному базису. Поэтому вещественная и мнимая части $\{\alpha_k^{1/2,\ 0}\}$ и $\{\beta_k^{1/2,\ 0}\}$ являются соответственно вещественной и мнимой частями $\{a_k\}$ и $\{b_k\}$. Это открывает простую возможность дополнительного совмещения косинусного преобразования для последовательностей вещественных чисел путем формирования из двух четных последовательностей вещественных чисел одной четной последовательности комплексных чисел. Тогда, пользуясь описанным приемом совмещения СДПФ (1/2,0), можно осуществить преобразование четных последовательностей вещественных чисел за время, почти вчетверо меньшее времени преобразования одной комплексной последовательности той же длины.

Рекуррентный алгоритм вычисления ДПФ. При обработке изображений и других двумерных сигналов иногда необходимо вычислять спектры следующих друг за другом сильно перекрывающихся фрагментов сигнала, или так называемые локальные спектры. В этом случае вычисление спектра каждого фрагмента целесообразно производить рекуррентно, используя спектр предшествующего ему при обработке фрагмента. Пусть фрагменты состоят из $N_1 x N_2$ элементов и размещаются на изображении с шагом p и q по двум координатам. Свяжем ДПФ двух соседних перекрывающихся фрагментов. Предположим, $\alpha_{r,s}^{(0,\ 0)}$—двумерный спектр первого из этих фрагментов. Тогда спектр $\alpha_{r,s}^{(p,\ q)}$ второго фрагмента, смещенного по координатам на p и q отсчетов относительно первого, можно записать как

$$\alpha_{r,s}^{(p,q)} = \frac{1}{\sqrt{N_1N_2}} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k+p,\,l+q} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] =$$

$$= \frac{1}{\sqrt{N_1N_2}} \left\{ \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] - \right.$$

$$- \sum_{k=0}^{p-1} \sum_{l=0}^{q-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] -$$

$$- \sum_{k=p}^{N_1-1} \sum_{l=0}^{q-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] -$$

$$- \sum_{k=0}^{p-1} \sum_{l=q}^{N_2-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] +$$

$$+ \sum_{k=N_1}^{N_1+p-1} \sum_{l=q}^{N_2-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] +$$

$$+ \sum_{k=N_1}^{N_1+p-1} \sum_{l=N_2}^{N_2+q-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] +$$

$$\left. + \sum_{k=p}^{N_1-1} \sum_{l=N_2}^{N_2+q-1} a_{k,\,l} \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] \right\} \times$$

$$\times \exp\left[-i2\pi\left(\frac{pr}{N_1} + \frac{qs}{N_2}\right)\right],$$

или после группировки и замены переменных

$$\alpha_{r,s}^{(p,q)} = \alpha_{r,s}^{(0,0)} \exp\left[-i2\pi\left(\frac{pr}{N_1} + \frac{qs}{N_2}\right)\right] +$$

$$+ \frac{1}{\sqrt{N_1N_2}} \left\{ \sum_{k=0}^{p-1} \sum_{l=0}^{q-1} (a_{k+N_1,\,l+N_2} - a_{k,\,l}) \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] + \right.$$

$$+ \sum_{k=p}^{N_1-1} \sum_{l=0}^{q-1} (a_{k,\,l+N_2} - a_{k,\,l}) \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] +$$

$$\left. + \sum_{k=0}^{p-1} \sum_{l=q}^{N_2-1} (a_{k+N_1,\,l} - a_{k,\,l}) \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] \right\} \times$$

$$\times \exp\left[-i2\pi\left(\frac{pr}{N_1} + \frac{qs}{N_2}\right)\right].$$

Таким образом, $\alpha_{k,s}^{(p,q)}$ вычисляется через $\alpha_{r,s}^{(0,0)}$ и три усеченных ДПФ (ДПФ укороченных последовательностей). В частном случае смещения только в одном направлении (например, вдоль 1}

$$\alpha_{r,s}^{(0,q)} = \left\{ \alpha_{r,s}^{(0,0)} + \frac{1}{\sqrt{N_1N_2}} \sum_{k=0}^{N_1-1} \sum_{l=0}^{q-1} (a_{k,\,l+N_2} - a_{k,\,l}) \times \right.$$

$$\left. \times \exp\left[i2\pi\left(\frac{kr}{N_1} + \frac{ls}{N_2}\right)\right] \right\} \exp(-i2\pi qs/N_2).$$

Если смещение равно одному элементу, как при «скользящей» обработке,

$$\alpha_{r,s}^{(0,1)} = \left\{ \alpha_{r,s}^{(0,0)} + \frac{1}{\sqrt{N_1N_2}} \sum_{k=0}^{N_1-1} (a_{k,\,N_2} - a_{k,\,0}) \times \right.$$

$$\left. \times \exp\left(i2\pi \frac{kr}{N_1}\right) \right\} \exp\left(-i2\pi \frac{s}{N_2}\right).$$

В одномерном случае, очевидно,

$$a_r^{(p)} = \left\{ a_r^{(0)} + \frac{1}{\sqrt{N}} \sum_{k=0}^{p-1} (a_{k+N} - a_k) \exp\left(i\,2\pi \frac{kr}{N} \right) \right\} \times$$

$$\times \exp\left(-i\,2\pi \frac{pr}{N} \right),$$

а при «скользящей» обработке

$$a_r^{(1)} = \left[a_r^{(0)} + \frac{1}{\sqrt{N_1}} (a_N - a_0) \right] \exp(-i\,2\pi r/N)$$

Используя приведенные соотношения, можно уменьшить количество операций, необходимых для вычисления локальных спектров. Степень уменьшения зависит от размеров фрагментов, их перекрытия, а также от доступной емкости запоминающего устройства для хранения промежуточных результатов.

4.2. ИСПОЛЬЗОВАНИЕ ДИСКРЕТНЫХ ПРЕОБРАЗОВАНИЙ ФУРЬЕ ДЛЯ ВЫЧИСЛЕНИЯ СВЕРТКИ, ИНТЕРПОЛЯЦИИ, СПЕКТРАЛЬНОГО АНАЛИЗА СИГНАЛОВ

Благодаря существованию алгоритмов БПФ дискретные преобразования Фурье (см. § 3.3) находят широкое применение для свертки, спектрального анализа и интерполяции сигналов.

Измерение фурье-спектров сигналов. Как следует из связи ДПФ с интегралами Фурье (§ 3.3), значения дискретных спектров Фурье последовательностей отсчетов сигналов могут рассматриваться с известным приближением как отсчеты непрерывного спектра Фурье непрерывного сигнала. Расположение растра отсчетов спектра в спектральной плоскости определяется при этом пара метра ми сдвига СДПФ. Меняя эти параметры, можно получить отсчеты спектра, расположенные произвольно относительно системы координат непрерывного спектра. Выполняя несколько последовательных СДПФ с соответствующими сдвигами, можно полу* чать оценки спектров сигналов с произвольно подробной дискретизацией по оси частот.

Пусть $\{a_r^{p,q}\}$ —отсчеты спектра сигнала $\{a_k\}$, полученные с помощью СДПФ *(p, q)*. Если над первыми M отсчетами того же сигнала, сдвинутого на k_0, выполнить СДПФ *(u, v)*:

$$a_r^{u,v} = \frac{1}{\sqrt{N}} \sum_{k=0}^{M-1} a_{k+k_0} \exp\left(i\,2\pi \frac{kv}{N} \right) \exp\left[i\,2\pi \frac{(k+u)\,r}{N} \right], \qquad (4.7)$$

то, как вытекает из формулы (3.43), записанной для спектров, отсчеты спектра $\{a_r^{u,v}\}$ будут сдвинуты относительно отсчетов спектра $\{a_r^{p,q}\}$ на *(q*—у)-ю часть интервала дискретизации по частоте и связаны с ними следующим интерполяционным соотношением:

$$a_r^{u,v} = \sum_{s=0}^{N-1} a_s^{p,q} \exp\left[-i\,2\pi (s+v) \left(k_0 + \frac{M-1}{2} + u \right) / N \right] \times$$

$$\times \operatorname{sincd}\left(M;\ \frac{r-s+q-v}{N} \right) \exp\Big[i\,2\pi (r+$$

$$+ q) \left(p + \frac{M-1}{2} \right) / N \Big] (M/N) \exp(i\,2\pi (uv - pq)/N) \qquad (4.8)$$

Возможны разные варианты выбора начальных параметров; сдвига p и q. Если выбрать $p=q=0$, что соответствует использованию стандартного ДПФ для вычисления начальных, опорных отсчетов спектра, то при четном N можно рекомендовать следующие значения параметров сдвига сигнала k_0 и преобразования u, а также числа отсчетов сигнала M для спектрального анализа: $k_0=u= —\ (N—2)/2$; $M = N—1$. Получающиеся при таком преобразовании спектральные коэффициенты связаны с опорными следующим образом:

$$\alpha_r^{u,\,v} = \left(\sum_{s=0}^{N-1} \alpha_s^{0,\,0}\, \mathrm{sincd}\left[(N-1);\ \frac{r-s+v}{N} \right] \right) \frac{N-1}{N} \times$$

$$\times \exp\left(i\pi \frac{N-2}{N} v \right).$$

Соответствующая схема анализа показана на рис. 4.4. С соответствующей заменой одномерного индекса k двумерным индексом (k, l), параметров сдвига $\&_0$ и p двумерными параметрами (k_0, l_0), (p_1, p_2) и одномерного СДПФ двумерным СДПФ по массиву из N_1N_2 отсчетов она справедлива и для двумерного случая. В двумерном случае описанный способ спектрального анализа можно, в частности, применять для перехода по формуле (3.54) от прямо-

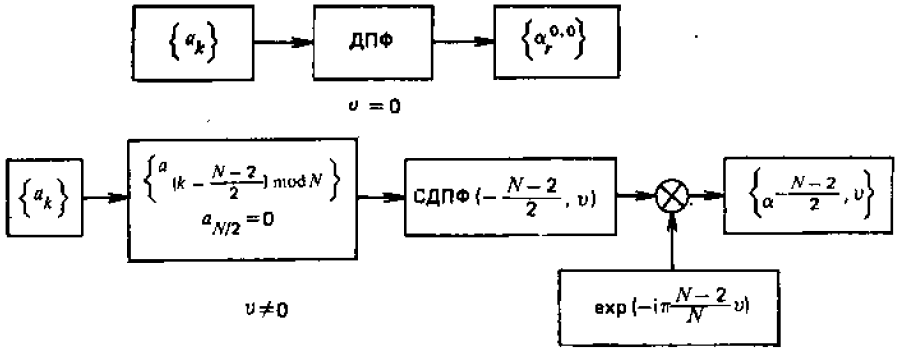


Рис. 4.4. Схема спектрального анализа (вариант 1)

угольного растра дискретизации сигнала к гексагональному растру дискретизации спектра Фурье. Такой гексагональный спектр удобнее обычного прямоугольного для оценки спектров изотропных сигналов, а также для фильтрации в частотной области с помощью изотропных двумерных фильтров.

Для того чтобы отсчеты дискретного спектра лучше соответствовали отсчетам непрерывного спектра, можно применять известные способы маскирования сигналов (см., например, [8]). Хорошие результаты может дать также четное продолжение сигнала с использованием для оценки спектров СДПФ (1/2,0), т.е. спектральный анализ по косинусному преобразованию. В этом случае $p = 1/2$, u = 0, и в (4.7) следует выбирать $M = N = 2N_0$, где N_0— длина исходной последовательности отсчетов сигнала до его четного продолжения; $k_0 = N_0$; $u = -(N_0 — 1/2)$. Как вытекает из (4.8), при таком выборе параметров преобразования

$$\alpha_r^{-(2N_0-1)/2;\,v} = \sum_{s=0}^{2N_0-1} \alpha_s^{1/2;\,0}\, \mathrm{sincd}\left(2N_0;\ \pi \frac{r-s+v}{2N_0} \right) \times$$

$$\times \exp\left[i\pi \frac{2N_0-1}{2N_0} v \right].$$

Отсюда следует схема спектрального анализа, показанная на рис. 4.5.

Известен еще один способ спектрального анализа с помощью ДПФ. Он основан на теореме интерполяции (3.34) и заключается в том, что для получения отсчетов спектра сигнала протяженностью в N отсчетов сигнал дополняется $(L—1)N$ нулевыми отсчетами по (3.34), после чего выполняется LN-точечное ДПФ. Этот способ дает такую же оценку спектра, что и описанный выше вариант рис. 4.4, но он требует больше вычислительных операций $(NL\log_2 NL$ вместо $NL\log_2 2N)$. Кроме того, этот алгоритм требует в L раз большей емкости запоминающего устройства процес-

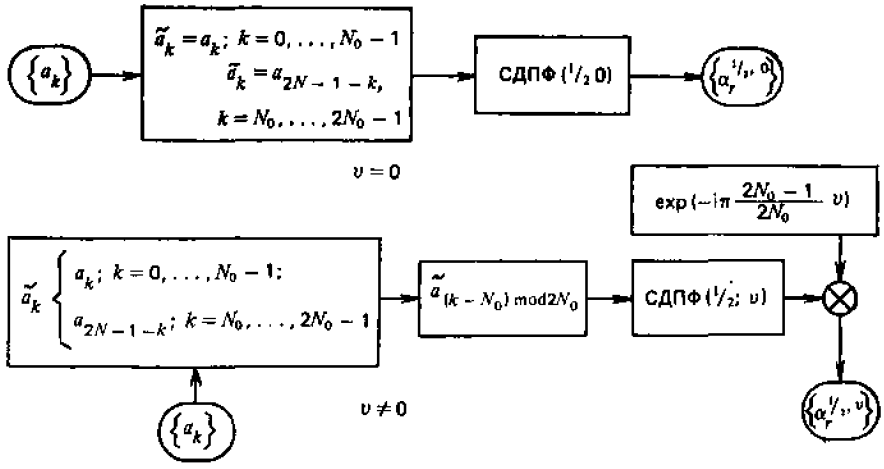


Рис. 4.5. Схема спектрального анализа (вариант 2)

сора. Отметим также, что СДПФ позволяет получить не только равноотстоящие друг от друга отсчеты, но и отсчеты с произвольным сдвигом относительно растра опорных отсчетов.

Интерполяция сигналов. Дискретные преобразования Фурье являются удобным инструментом для нахождения незаданных промежуточных отсчетов сигналов по его заданным отсчетам (интерполяции сигналов). Для непрерывных сигналов, удовлетворяющих теореме отсчетов, оптимальная интерполяция определяется как

$$a(k + \delta) = \sum_{n=-\infty}^{\infty} a_n \operatorname{sinc} \pi (k - n + \delta); \quad |\delta| \leqslant 1/2$$

По теореме об интерполяции (3.43), (3.44) для последовательностей конечной длины это соотношение можно аппроксимировать с помощью пары преобразований СДПФ и ОСДПФ с соответственно подобранными значениями параметров сдвига *(u, v; p, q)*.

Если количество отсчетов сигнала N нечетно, обратное СДПФ следует выполнить по всем отсчетам спектра сигнала $\{\alpha^{u, v}_r\}$. Если N — четное число, как обычно при использовании для выполнения СДПФ алгоритмов БПФ, то при обратном преобразовании следует выбирать M в формуле (3.43) равным N—1, т.е. последний $(N$—1) отсчет спектра $\alpha^{u, v}_{N-1}$ полагать равным нулю. Схемы вычислений для интерполяции сигналов совпадают со схемами спектрального анализа (см. рис. 4.4, 4.5), с той разницей, что показанные там операции производятся не над сигналами, а над их ДПФ (рис. 4.4) и косинусными (рис. 4.5) спектрами. Последний вариант предпочтительнее, так как благодаря принудительному четному продолжению сигнала, выполняемому при косинусном преобразовании, уменьшаются ошибки интерполяции на краях последовательности.

В двумерном случае при обработке изображений особый интерес представляет использование СДПФ для перехода от прямоугольного растра отсчетов к гексагональному, и наоборот. Это можно сделать с помощью двумерных СДПФ, описываемых формулами (3.51), (3.54). Например, для нахождения отсчетов спектра двумерного сигнала по гексагональному растру по его прямоугольным отсчетам необходимо выполнить прямое СДПФ по формуле (3.54), а обратное — по (3.51) со знаком «минус» в экспоненте.

Как и для спектрального анализа, для интерполяции, сигналов может применяться алгоритм с дополнением спектра нулями, основанный на теореме интерполяции (3.34). Но, как уже указывалось, он менее эффективен в вычислительном отношении, чем алгоритм, использующий СДПФ.

Вычисление свертки. Одно из важнейших применений дискретных преобразований Фурье — вычисление с их помощью цифровой •свертки сигналов. Оно основано на теоремах о свертке (3.32) и (3.45) для дискретных преобразований Фурье (см. § 3.3). Такая фильтрация сигналов называется *фильтрацией в частотной области.*

Особенность реализации цифровой свертки, определяемой формулой (3.9), с помощью циклических сверток (3.32) и (3.45) состоит в способе доопределения сворачиваемых последовательностей за краями заданного интервала отсчетов (О, N—1): сворачиваемые последовательности продолжаются по правилу, определяемому свойствами (3.27) и (3.38)

цикличности дискретных преобразований Фурье. Так, использование ДПФ дает циклическую свертку

$$c_k = \sum_{n=0}^{N-1} h_n a_{(k-n) \bmod N}$$

соответствующую простому периодическому продолжению сигнала. В результате оказывается, что на краях последовательности (при k, близких к 0 и W—1) в свертке участвуют не соседние отсчеты сигнала, близкие к отсчету с номером k, а удаленные от него на всю длину последовательности N. Для борьбы с этими краевыми эффектами применяется принудительное удлинение последовательности с доопределением значений сигнала.

Удобным и достаточно простым способом доопределения является четное дополнение до двойной длины по правилу

$$\tilde{a}_k = \begin{cases} a_k; & k = 0, 1, \ldots, N-1; \\ a_{2N-1-k}; & k = N, \ldots, 2N-1 \end{cases} \qquad (4.9)$$

Если при этом четным образом по правилу

$$\tilde{h}_n = \begin{cases} h_n; & n = 0, 1, \ldots, N-1; \\ h_{2N-n}; & n = N, N+1, \ldots, 2N-1 \end{cases} \qquad (4.10)$$

дополнить до двойной длины импульсную реакцию фильтра и для вычисления прямого преобразования использовать СДПФ (1/2,0)

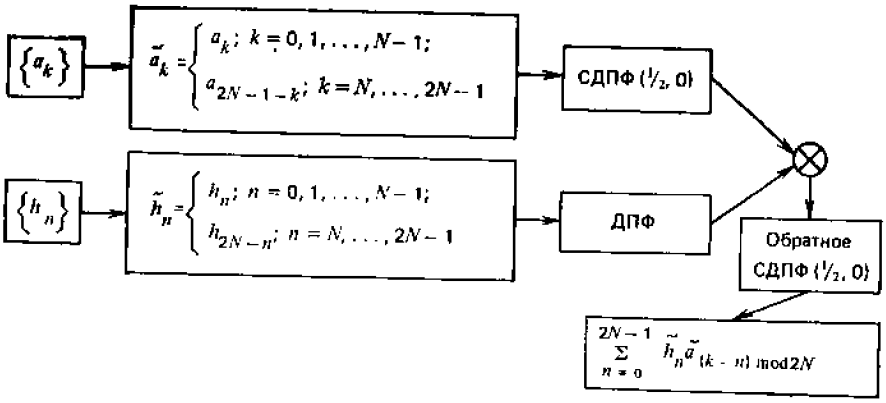


Рис. 4.6. Схема свертки в частотной области с четным продолжением сигналов

для последовательности $\{a_k\}$ и ДПФ для последовательности $\{h_n\}$, а при обратном преобразовании — обратное СДПФ (1/2,0), как показано па схеме рис. 4.6, то для прямого и обратного СДПФ (1/2,0) применим описанный выше совмещенный алгоритм, так что количество операций на фильтрацию почти не увеличивается за счет удвоения длины последовательности. Для выполнения ДПФ последовательности можно рекомендовать усеченные алгоритмы ДПФ [46, 49], описанные в гл. 5, пользуясь тем, что, как правило, протяженность импульсной реакции фильтра намного меньше протяженности сигнала, т.е. последовательность $\{\tilde{h}_n\}$ содержит большое количество нулей. Кроме того, следует учитывать, что для заданного фильтра ДПФ вычисляется один раз для всех возможных обрабатываемых сигналов. Приведенные рекомендации справедливы и для двумерной свертки. В этом случае четное дополнение (4.9) и (4.10) выполняется по двум координатам.

Поскольку СДПФ и ДПФ выполняются с помощью алгоритма БПФ, количество отсчетов сигналов, как правило, выбирается равным целой степени двух. Если фактически имеющееся количество отсчетов не есть целая степень двух, то сигнал нужно дополнить до ближайшего такого числа отсчетов, делая это с учетом правил циклического продолжения (3.27) и (3.38) так, чтобы на краях новой продолженной последовательности не было разрывов

сигнала. Хорошие результаты получаются при определении недостающих отсчетов линейной интерполяцией между крайними отсчетами периодически продолженной последовательности. Этот способ дополнения последовательностей иллюстрируется рис. 4.7 для случая простого периодического продолжения, соответствующего стандартному ДПФ.

Количество дополнительных отсчетов должно быть не менее двойного количества ненулевых отсчетов импульсной реализации в формуле (3.9). Если это не так, последовательность сигнала с найденными дополнительными отсчетами нужно четным образом

Рис. 4.7. Доопределение последовательности при вычислении свертки с помощью ДПФ

дополнить до двойной длины и далее работать с этой удвоенной последовательностью, как было описано выше.

Если количество ненулевых отсчетов импульсной реакции фильтра намного меньше количества отсчетов сигнала, свертку сигнала с этой импульсной реакцией можно вычислять быстрее, если делать это не для всей реализации сигнала, а по частям, разбивая исходную последовательность на фрагменты. При этом вопрос о доопределении сигнала возникает только для крайних фрагментов на границах последовательности сигнала. Остальные фрагменты нужно выбирать либо с перекрытием на количество ненулевых отсчетов импульсной реакции фильтра, а при стыковке результатов свертки лишние отсчеты (на половину длины перекрытия с каждой стороны) отбрасывать, либо дополнять фрагменты по краям нулями на двойную длину импульсной реакции, а при стыковке результатов свертки складывать перекрывающиеся участки. Такой способ свертки с использованием ДПФ называется секционированием. Он дает выигрыш в числе операций по сравнению со сверткой по всей последовательности сигнала, равный по порядку величины (log.N)/logNф, где N — количество отсчетов во всей реализации сигнала, Nф — количество отсчетов во фрагменте.

Следует отметить, что использование дискретных преобразований Фурье для вычисления свертки эффективно в вычислительном отношении только для больших последовательностей сигналов и протяженных импульсных реакций фильтров. Точные рекомендации зависят от типа используемого процессора. Ориентировочной пороговой точкой числа отсчетов импульсной реакции фильтра, ниже которой применение ДПФ нецелесообразно, может служить число 10.

4.3. АЛГОРИТМЫ ЦИФРОВОЙ ФИЛЬТРАЦИИ В ПРОСТРАНСТВЕННОЙ ОБЛАСТИ

Рекурсивная цифровая фильтрация. Согласно (3.9) отсчеты выходного сигнала линейного инвариантного к сдвигу фильтра можно найти как взвешенную сумму входных отсчетов. Для этого требуется, вообще говоря, выполнить N операций умножения и $(N—!)$ операций сложения. Такая реализация фильтра называется *нерекурсивной,* или *трансверсальной.*

Существует класс фильтров, вид импульсной реакции которых позволяет преобразовать (3.9) в рекурсивное соотношение:

$$b_k = h_0 a_k + \sum_{s=1}^{N_p} g_s b_{k-s} \qquad (4.11)$$

Вычисление по этой формуле требует меньшего числа операций, чем по (3.9). Например, при $g_1 \neq 0$; $g_2 = g_3 = \ldots = 0$

$$b_k = h_0 a_k + g_1 b_{k-1}$$

т.е. каждое значение b_k может быть вычислено посредством всего двух умножений и одного сложения, тогда как в прямой сумме (3.9) будет, вообще говоря, бесконечное число членов, так как их вес $h_n = h_0 g_1^n$ может только более или менее быстро убывать при $g_1 < 1$ и никогда не обращается в нуль. Рекурсивным фильтрам соответствуют схемы вычислений с обратной связью.

Соотношение (4.11) описывает простейший *рекурсивный цифровой фильтр.* В общем случае такой фильтр определяется соотношением

$$b_k = \sum_{n=0}^{N_т-1} h_n a_{k-n} + \sum_{n=1}^{N_р} g_n b_{k-n} \qquad (4.12)$$

где $N_т$ и $N_р$ — количество слагаемых в трансверсальной и рекурсивной частях формулы соответственно.

Хороший пример возможности рекурсивного представления дает фильтр (3.16), осуществляющий скользящее усреднение сигнала. Действительно, нетрудно показать, что в этом случае

$$\bar{a}_k = (a_{k+N} - a_{k-N+1})/(2N+1) + \bar{a}_{k-1} \qquad (4.13)$$

Таким образом, оказывается, что, построив вычисления по (4.13), можно находить текущее среднее сигнала не за *2N* сложений на один отсчет среднего, как в (3.16), а только за три. Замечательно, что количество операций здесь не зависит от количества отсчетов, по которым происходит усреднение.

Вообще при рекурсивной фильтрации количество операций на вычисление результата или, соответственно, количество сумматоров и перемножителей в схемной реализации, как правило, намного меньше, чем при нерекурсивной. Это преимущество в быстродействии заставляет всегда искать возможность аппроксимации требуемого фильтра рекурсивным. Для такой аппроксимации удобнее всего пользоваться частотными характеристиками фильтров. Чтобы вывести выражение для частотной характеристики рекурсивного фильтра, преобразуем (4.12) в равенство

$$b_k - \sum_{n=1}^{N_р} g_n b_{k-n} = \sum_{n=0}^{N_т-1} h_n a_{k-n}$$

Теперь, пренебрегая краевыми эффектами для последовательности отсчетов выходного сигнала $\{b_k\}$, т.е. считая, что при восстановлении непрерывного сигнала $b(x)$ количество отсчетов $\{b_k\}$ бесконечно велико, получим, что входной и выходной сигналы фильтра (4.12) связаны соотношением

$$\left(1 - \sum_{n=1}^{N_р} g_n \delta(x - n\Delta x)\right) * b(x) =$$
$$= \sum_{k=-\infty}^{\infty} \sum_{n=0}^{N-1} h_n \varphi_0[x - (k-n)\Delta x]\, \psi_0(\xi - n\Delta x)] * a(x)$$

где $*$ означает операцию свертки. Отсюда следует, что непрерывная частотная характеристика рекурсивного фильтра (4.12) определяется выражением

$$\tilde{H}(f) = \frac{\sum_{n=0}^{N_т-1} h_n \exp(i\,2\pi f n \Delta x)}{1 - \sum_{n=1}^{N_р-1} g_n \exp(i\,2\pi f n \Delta x)} H_g(f) H_п(-f) \qquad (4.14)$$

а дискретная частотная характеристика соответственно выражением

$$\tilde{\tilde{H}}(f) = \frac{\sum_{n=0}^{N_т-1} h_n \exp(i\,2\pi f n \Delta x)}{1 - \sum_{n=1}^{N_р-1} g_n \exp(i\,2\pi f n \Delta x)} \qquad (4.15)$$

Ввиду того, что рекурсивные фильтры являются фильтрами с обратной связью, они могут быть неустойчивыми. Эта неустойчивость может проявляться в том, что при рекурсивной фильтрации малые ошибки вычислений предыдущих значений, например ошибки вследствие квантования значений b_k, могут дать неизмеримо большие ошибки в вычислении текущих значений. Поэтому при синтезе рекурсивных фильтров следует проверять их на устойчивость [39].

В двумерном случае также возможно построение рекурсивных фильтров, но при этом необходимо задаться направлением рекурсии. Если считать «прошлыми» значения сигнала слева и сверху от данного отсчета на прямоугольном растре (рис. 4.8), то формулу (4.12) одномерного рекурсивного фильтра можно обобщить на двумерный случай следующим образом:

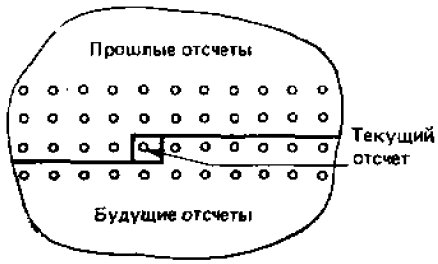


Рис. 4.8. Прошлые и будущие отсчеты на прямоугольном растре при двумерной фильтрации

$$b_{k,\,l} = \sum_{m=0}^{N_{1т}-1} \sum_{n=0}^{N_{2т}-1}{}' h_{m,\,n} a_{k-m,\,l-n} + \\ + \sum_{m=1}^{N_{1р}} \sum_{n=-N_{2р}}^{N_{2р}} g_{m,\,n} b_{k-m,\,l-n} + \sum_{n=1}^{N_{2р}} g_{0,\,n} b_{k,\,l-n} \qquad (4.16)$$

Соответственно дискретная частотная характеристика двумерного рекурсивного фильтра определяется выражением

$$\tilde{\tilde{H}}(f_1, f_2) = \left\{ \sum_{m=0}^{N_{1т}-1} \sum_{n=0}^{N_{2т}-1} h_{m,\,n} \exp\left[i\,2\pi\left(mf_1\Delta x_1 + nf_2\Delta x_2\right)\right] \right\} \times \\ \times \left\{ 1 - \sum_{m=1}^{N_{1р}} \sum_{n=-N_{2р}}^{N_{2р}} g_{m,\,n} \exp\left[i\,2\pi\left(mf_1\Delta x_1 + nf_2\Delta x_2\right)\right] - \right. \\ \left. - \sum_{n=1}^{N_{2р}} g_{0,\,n} \exp\left(i\,2\pi n f_2 \Delta x_2\right) \right\}. \qquad (4.17)$$

Выражения для дискретных частотных характеристик цифровых фильтров можно сделать менее громоздкими, если произвести в них замену $z = \exp(-i2\pi f\Delta x)$. Подставим, например, значение г в (4.13):

$$\tilde{\tilde{H}}(z) = \sum_{n=0}^{N-1} h_n z^{-n} \qquad (4.18)$$

Эта формула называется z-*преобразованием* последовательности чисел $\{h_n\}$. Таким образом, дискретные частотные характеристики как функции z выражаются через z-преобразования отсчетов импульсных реакций.

Двумерные разделимые фильтры. Существует еще один класс двумерных цифровых фильтров, представляющих интерес с точки зрения экономии вычислительных затрат — *двумерные разделимые фильтры.* Это фильтры, импульсная реакция которых $h_{n,m}$ может быть представлена в виде поэлементного произведения одномерных последовательностей

$$h_{n,\,m} = h_m^{(1)} h_n^{(2)} \qquad (4.19)$$

Для таких фильтров формула (3.20) переходит в

$$b_{k,\,l} = \sum_{m=0}^{N_1-1} h_m^{(1)} \sum_{n=0}^{N_2-1} h_n^{(2)} a_{k-m,\,l-n} \qquad (4.20)$$

Для вычисления одного отсчета $b_{k,l}$ по (4.20) требуется выполнить $N_1 + N_2$ операций умножения и $N_1 + N_2$—2 операций сложения, а не, соответственно, N_1N_2 и $(N_1$—1) $(N_2$—1) операций, требуемых для вычислений по формуле (3.20).

Представляя двумерный фильтр в разделимой форме и используя рекурсивное представление одномерных фильтров, можно получить громадную экономию количества требуемых операций.

Примером такого фильтра с возможностью рекурсивного представления одномерных фильтров является двумерный фильтр, используемый для получения значений текущего среднего значения сигнала по прямоугольной окрестности:

$$\bar{a}_{k,\,l} = \left(\sum_{m=-N_1}^{N_1} \sum_{n=-N_2}^{N_2} a_{k-n,\,l-m} \right) / (2N_1 + 1)\,(2N_2 + 1) \qquad (4.21)$$

Представление (4.19) не всегда возможно. Но можно аппроксимировать требуемую функцию $h_{m,n}$ суммой разделимых функций:

$$h_{m,\,n} = \sum_{r=0}^{R-1} h_m^{(r)} g_n^{(r)} \qquad (4.22)$$

Если число членов R в этой сумме невелико, такая замена одного фильтра несколькими также может быть эффективней в вычислительном отношении, чем фильтрация по (3.20). Задача о наилучшем представлении (4.22) родственна задаче о наилучшем конечномерном приближении сигналов.

Представление (4.22) импульсной реакции фильтра в виде суммы импульсных реакций более простого вида соответствует тому, что фильтрация сигнала осуществляется параллельно несколькими фильтрами, и результаты фильтрации складываются. Такое представление можно назвать *параллельным* (рис. 4.9,а).

Последовательная (каскадная) цифровая фильтрация. Возможно также последовательное (каскадное) представление цифровых фильтров, когда фильтрация сигнала осуществляется последовательно набором действующих друг за другом простых фильтров (рис. 4.9,6). В этом случае требуемая импульсная реакция экви-

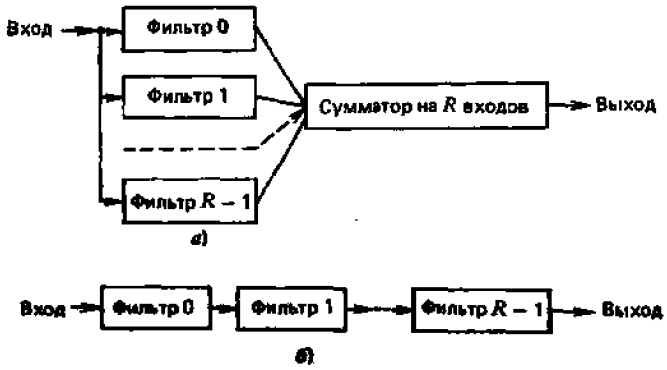


Рис. 4.9. Параллельное (а) и последовательнее *(б)* представление фильтров

валентного фильтра записывается в виде свертки импульсных реакций более простого вида

$$h_n = \sum_{n_{R-1}} h_{n_{R-1}}^{(R-1)} \cdots \sum_{n_1} h^0_{\,n - \sum_{r=1}^{R-1} n_r} h_{n_1}^{(1)} \qquad (4.23)$$

в соответствии с последовательным пропусканием сигнала через несколько фильтров.

В одномерном случае последовательное (каскадное) представление фильтра не дает выигрыша в числе операций, если фильтры каскадов не могут быть построены как рекурсивные, и даже дает некоторый проигрыш по сравнению с однокаскадными представлениями. Действительно, пусть N— количество отсчетов (протяженность) импульсной реакции фильтров в каждом каскаде в формуле (4.23). Тогда протяженность импульсной реакции h_n при R каскадах составит $[R(N—1) + 1]$ отсчет. Это значит, что если производить фильтрацию по формуле (3.9), то потребуется выполнить $[R(N—1) + 1]$ операций на каждый отсчет преобразованного сигнала, а при последовательной фильтрации R фильтрами RN операций, т.е. больше. Но двумерные последовательные (каскадные) фильтры могут быть значительно выгоднее однокаскад-ных в отношении вычислительных затрат, так как в двумерном случае однокаскадный фильтр требует затраты $[R(N_1—1) + 1]$x[R(N_2—1) + 1] операций вместо RN_1N_2 для эквивалентного последовательного (каскадного) фильтра с протяженностью импульсной реакции каждого каскада N_1N_2 отсчетов. Поэтому последовательное (каскадное) представление двумерных фильтров представляет значительный резерв экономии машинного времени при обработке двумерных сигналов в цифровых ЭВМ, и, соответственно, аппаратных средств в специализированных процессорах.

Параллельная и рекурсивная организация цифровых фильтров. Главным принципиальным резервом повышения производительности цифровой обработки сигналов является распараллеливание вычислительных операций, требуемых для обработки.

Ниже описан способ параллельной рекурсивной организации цифровой одномерной и многомерной линейной фильтрации. Способ основан на приближенном представлении импульсной реакции требуемого цифрового фильтра с помощью разложения ее по линейно-независимой системе базисных функций, коэффициенты разложения для которых могут вычисляться рекурсивно при движении апертуры фильтра вдоль обрабатываемой последовательности [50]. При этом обеспечивается распараллеливание цифрового фильтра на число каналов, равное количеству базисных функций в разложении импульсной реакции цифрового фильтра. Сложность вычислений в каждом канале определяется благодаря рекурсии не протяженностью простирания базисной функции (в двумерном случае площадью), а протяженностью ее границы — количеством концевых точек.

Рассмотрим одномерные цифровые фильтры:

$$b_k = \sum_{n=-N_1^{(h)}}^{N_2^{(h)}} h_n a_{k-n} \qquad (4.24)$$

где $\{a_k\}$ и $\{ b_k \}$ — последовательности отсчетов входного и выходного сигнала фильтра соответственно; $\{h_n\}$—отсчеты импульсной реакции цифрового фильтра; $(N_1^{(h)} + N_2^{(h)} + 1)$ — протяженность импульсной реакции.

Пусть последовательность $\{h_n\}$ можно приближенно с заданной точностью разложить по системе линейно-независимых базисных функций $\{\varphi_r(n)\}$. r = 0, 1,..., R— 1; R< $N_1^{(h)} + N_2^{(h)}$ +1:

$$h_n \approx \sum_{r=0}^{R-1} \eta_r \varphi_r(n) \qquad (4.25)$$

Тогда $\{b_k\}$ можно приближенно представить в виде:

$$b_k \approx \sum_{n=-N_1^{(h)}}^{N_2^{(h)}} \sum_{r=0}^{R-1} \eta_r \varphi_r(n)\, a_{k-n} = \sum_{r=0}^{R-1} \eta_r \sum_{n=-N_1^{(r)}}^{N_2^{(r)}} \varphi_r(n)\, a_{k-n}$$

где в последнем равенстве учтено то обстоятельство, что, в принципе, протяженность функции $\{\varphi_r(n)\}$ может быть меньше протяженности последовательности $\{h_n\}$, т.е. $N_1^{(r)} \leqslant N_1^{(h)}$, $N_2^{(r)} \leqslant N_2^{(h)}$ Обозначим

$$\beta_r(k) = \sum_{n=-N_1^{(r)}}^{N_2^{(r)}} \varphi_r(n)\, a_{k-n} \tag{4.26}$$

Тогда

$$b_k \approx \sum_{r=0}^{R-1} \eta_r \beta_r(k) \tag{4.27}$$

где точность приближения определяется точностью разложения (4.25). Тем самым цифровой фильтр (4.24) сведен к R параллельным цифровым фильтрам (4.26) и сумматору (4.27) с R входами, вычисляющему сумму выходных сигналов фильтров (4.26). В тривиальном случае, когда $\varphi_r(n) = \delta(n-r)$, где $\delta(\cdot)$— символ (дельта-функция) Кронекера, (4.27) тождественно (4.24), что соответствует распараллеливанию (4.24) на $(N_1^{(h)} + N_2^{(h)} + 1)$ фильтров, каждый из которых выполняет только одну операцию умножения на отсчет сигнала вместо $(N_1^{(h)} + N_2^{(h)} + 1)$ для фильтра (4.1).Уменьшить количество фильтров можно, перейдя к другим базисам, которые могут более экономно описать импульсную реакцию требуемого цифрового фильтра. Однако в общем случае количество операций для каждого параллельного фильтра (4.26) может быть примерно таким же, что и для исходного фильтра (4.24). Средством упрощения вычислительной сложности цифровых фильтров является рекурсивная организация фильтрации. Поэтому, потребуем, чтобы фильтры (4.26) были рекурсивными.

В простейшем и наиболее быстродействующем рекурсивном фильтре первого порядка каждый данный отсчет выходного сигнала зависит только от одного отсчета, найденного на предыдущем шаге. Выразим $\beta_r(k)$ через $\beta_r(k-1)$:

$$\beta_r(k) = \sum_{n=-N_1^{(r)}}^{N_2^{(r)}} \varphi_r(n)\, a_{k-1-(n-1)} = \sum_{m=-N_1^{(r)}-1}^{N_2^{(r)}-1} \varphi_r(m+1)\, a_{k-1-m} =$$

$$= \sum_{m=-N_1^{(r)}}^{N_2^{(r)}-1} \varphi_r(m+1)\, a_{k-1-m} + \varphi_r(-N_1^{(r)})\, a_{k+N_1^{(r)}}.$$

Пусть теперь $\varphi_r(m)$ таково, что

$$\varphi_r(m+1) = \varphi_r^{(0)} \varphi_r(m) \tag{4.28}$$

Тогда

$$\beta_r(k) = \varphi_r^{(0)} \sum_{m=-N_1^{(r)}}^{N_2^{(r)}-1} \varphi_r(m)\, a_{k-1-m} + \varphi_r(-N_1^{(r)})\, a_{k+N_1^{(r)}} =$$

$$= \varphi_r^{(0)} \beta_r(k-1) + \varphi_r(-N_1^{(r)})\, a_{k+N_1^{(r)}} - \varphi_r^{(0)} \varphi_r(N_2^{(r)})\, a_{k-1-N_2^{(r)}} \tag{4.29}$$

т.е. при выполнении условия (4.28) фильтры (4.26) являются рекурсивными, причем количество операций для них на один отсчет выходного сигнала составляет три операции умножения и две операции сложения независимо от протяженности базисных функций.

Определим базисные функции, удовлетворяющие условию (4.28), для чего рассмотрим их дискретное преобразование Фурье по m:

$$\varphi_r(m) = \sum_{p=0}^{M-1} \xi_p^{(r)} \exp(i\, 2\pi p m / M); \quad M = N_1^{(r)} + N_2^{(r)} + 1$$

и наложим на них условие (4.28):

$$\varphi_r^{(m+1)} = \sum_{p=0}^{M-1} \xi_p^{(r)} \exp(i\,2\pi p/M) \exp(i\,2\pi pm/M) = \varphi_r^{(0)} \varphi_r(m) =$$
$$= \varphi_r^{(0)} \sum_{p=0}^{M-1} \xi_p^{(r)} \exp(i\,2\pi pm/M). \tag{4.30}$$

Из необходимости почленного равенства в (4.30) вытекают следующие условия для $\xi_p^{(r)}$ и $\varphi_r^{(0)}$:

$$\xi_p^{(r)} = \delta(p - p_0); \quad \varphi_r^{(0)} = \exp(i\,2\pi p_0/M)$$

Таким образом, условию (4.28) удовлетворяют два базиса: базис дискретных экспоненциальных функций (ДЭФ)

$$p_0 = r; \quad \varphi_r(m) = \exp(i\,2\pi rm/M); \quad M = N_1^{(h)} + N_2^{(h)} + 1 \tag{4.31}$$

и линейно-независимый базис из прямоугольных функций:

$$p_0 = 0; \quad \varphi_r(m) = \mathrm{rect}\left(\frac{m + N_1^{(r)}}{N_1^{(r)} + N_2^{(r)} + 1}\right) \tag{4.32}$$

Схема организации вычислений в соответствии с описанным способом показана на рис. 4.10.

При использовании базиса (4.31) каждый из R параллельных рекурсивных фильтров (4.3) выполняет локальный дискретный Фурье-анализ на частоте r обрабатываемой последовательности, зеркально обращенной по k. Получаемые при этом спектральные коэффициенты последовательности перемножаются поэлементно на соответствующие спектральные коэффициенты импульсной реакции фильтра, т.е. на отсчеты его дискретной частотной характеристики, и складываются для получения результата фильтрации. Это аналог обработки сигнала в частотной области (см. § 4.2) с той разницей, что обратное преобразование Фурье заменяется суммированием (4.27), так как требуется вычислить только один (нулевой) член этого преобразования.

При использовании базиса (4.32) импульсная реакция цифрового фильтра как бы квантуется и представляется в виде суммы R прямоугольных функций разной протяженности. Это иллюстрируется рис. 4.11, на котором $h^{(r)}$, $h^{(r+1)}$ — границы r-го интервала

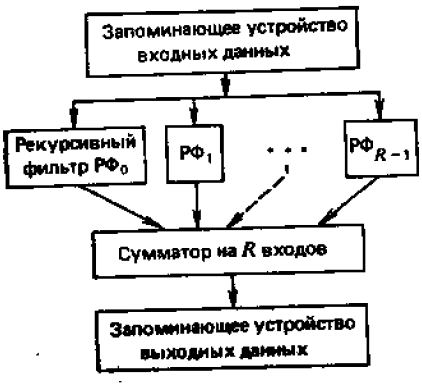


Рис. 4.10. Схема параллельной ре курсивной организации цифровых фильтров

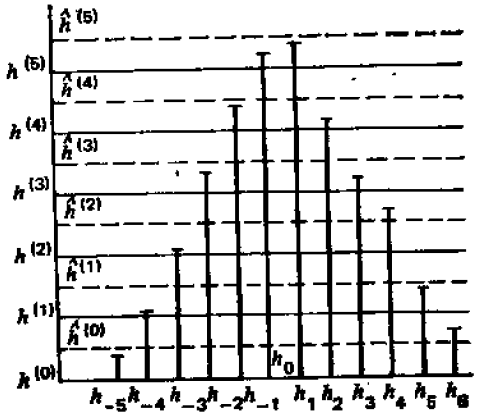


Рис. 4.11. Аппроксимация импульсной реакции фильтра в базисе прямоугольных функции

квантования, $\hat{h}^{(r)}$ — соответствующее квантованное значение r-го интервала, $h^{(r)} \leqslant \hat{h}^{(r)} < h^{(r+1)}$, $\hat{h}^{(r)} = \sum_{k=0}^{r} \eta_{|k}$. В этом случае фильтры (4.26) вычисляют локальные средние значения обрабатываемой последовательности на отрезке длиной в $N_1^{(r)} + N_2^{(r)} + 1$ элемент, и эти

локальные средние складываются с весами, равными приращениям значений «уровней квантования» импульсной реакции. -

Фильтры (4.26), соответствующие базису из прямоугольных функций, проще фильтров, соответствующих базису ДЭФ, однако при одинаковой точности аппроксимации $\{h_n\}$ с помощью (4.25) их может потребоваться больше.

Для двумерных цифровых фильтров

$$b_{k,l} = \sum_{n=-N_1^{(h)}}^{N_2^{(h)}} \sum_{m=-M_1^{(h)}}^{M_2^{(h)}} h_{m,n} a_{k-n,\,l-m}$$

в качестве базисных функций $\varphi_{r,s}(n,\ m)$ разложения импульсной реакции

$$h_{n,m} \approx \sum_{r=0}^{R-1} \sum_{q=0}^{Q-1} \eta_{r,q} \varphi_{r,q}(n,m)$$

проще всего выбирать разделимые функции

$$\varphi_{r,q}(n,m) = \varphi_r^{(1)}(n)\,\varphi_q^{(2)}(m)$$

При этом соответствующие параллельные фильтры

$$\beta_{r,s}(k) = \sum_{m=-M_1^{(q)}}^{M_2^{(q)}} \varphi_q(m) \sum_{n=-N_1^{(r)}}^{N_2^{(r)}} \varphi_r(n)\, a_{k-n,\,l-m}$$

будут наиболее простыми. По отношению к одномерным функциям $\varphi_r^{(1)}(n)$ и $\varphi_q^{(2)}(m)$, очевидно, применимы полученные выше результаты. Отсюда следует, что в двумерном случае кроме «чистых» базисов ДЭФ и прямоугольных функций возможны также «смешанные» базисы — прямоугольный по одной координате и экспоненциальный по другой.

Существуют также неразделимые рекурсивные двумерные базисы. Можно указать, по крайней мере, один такой базис — базис из прямоугольных функций, ограниченных произвольной кривой (аналогичный разделимый базис ограничен прямоугольником, если считать, что двумерный сигнал задан на прямоугольном растре). Апертура такого двумерного фильтра показана на рис. 4.12 (его разделимый аналог иллюстрируется рис. 4.12,6). На этом рисунке заштрихованы пограничные элементы апертуры, а стрелки показывают направления сканирования по двумерным коорди

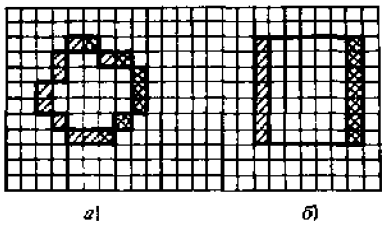


Рис. 4.12. Неразделимый *(а)* и разделимый (б) двумерный рекурсивный базис

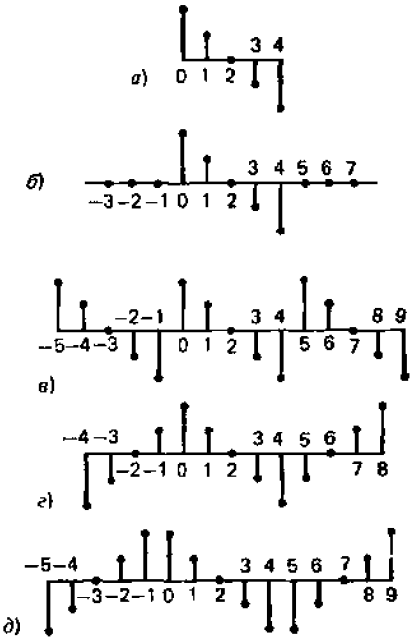


Рис. 4.13. Способы доопределения → сигналов:

а — исходная последовательность; б — дополнение нулями; в — экстраполяция повторением; *г* — четное продолжение; *д* — четное продолжение с повторением крайнего отсчета

натам сигнала *k* и *l*. Одинарной штриховкой показаны элементы, уходящие в процессе сканирования из апертуры, двойной — элементы, соответствующие $a'_{k+Nl}(r)$ (вновь поступающие).

Оценим выигрыш в производительности и аппаратурных затратах, а также точность представления импульсных реакций в параллельной и рекурсивной форме. В описанном методе организации цифровой линейной фильтрации выигрыш в производительности достигается за счет распараллеливания. В тривиальном случае распараллеливания фильтра с апертурой в $N = N_1^{(h)} + N_2^{(h)} + 1$ отсчетов на *N* параллельных фильтров производительность повышается в N раз но сравнению с однопроцессорным фильтром. При использовании рекурсивных базисов типа (4.28) количество операций на один отсчет сигнала уменьшается по сравнению с однопроцессорным фильтром в N/3 раза, и этот выигрыш не зависит от степени распараллеливания. От нее зависит только точность аппроксимации требуемой импульсной реакции.

Сокращение аппаратурных затрат по сравнению с тривиальным распараллеливанием в *g=N/R* раз достигается в описанном методе в обмен на точность представления импульсной реакции фильтра. В случае представления импульсной реакции фильтра по (4.25) с помощью таких ортонормальных базисов, как базис ДЭФ, среднеквадратическая погрешность аппроксимации при R<N может быть оценена по теореме Парсеваля:

$$|\varepsilon|^2 = \sum_{r=R}^{N-1} |\eta_r|^2$$

Для базиса из прямоугольных функций точность аппроксимации можно оценить методами, принятыми в теории оптимального квантования (см. §2.3):

$$|\varepsilon|^2 = \sum_{r=0}^{R-1} \sum_{h^{(r-1)} \leqslant h < h^{(r)}} |h - \hat{h}^{(r)}|^2$$

где суммирование во внутренней сумме проводится по всем значениям отсчетов импульсной реакции фильтра $\{h_n\}$, попадающим в заданный интервал квантования $\{h^{(r-1)}, h^{(r)}\}$ (см. рис. 4.11). При заданной степени распараллеливания *R,* равной количеству уровней квантования импульсной реакции, ошибку аппроксимации можно минимизировать оптимальным выбором границ квантования $\{h^{(r)}\}$ и квантованных значений $\{\hat{h}^{(r)}\}$ так же, как это делается при оптимальном поэлементном квантовании.

Краевые эффекты при цифровой фильтрации. В заключение остановимся на способах борьбы с «краевыми эффектами», т.е. искажениями результата фильтрации, связанными с тем, что количество отсчетов заданной последовательности сигнала конечно.

Вернемся к основной формуле цифровой фильтрации (3.9). Пусть отсчеты исходной последовательности входного сигнала $\{a_k\}$ нумеруются от 0 до N_a—1. Формула (3.9) определяет только отсчеты выходного сигнала с номерами от $k=N$—1 до $k=N_a$—N+1. Отсчеты с номерами от $k=0$ до $k=N$—2 и от $k=N_a$—N+2 до $k=N_a$—1 не определены, так как не заданы значения исходной последовательности при $k<0$ и $k>N_a$—1. Таким образом, при линейной цифровой фильтрации по (3.9) длина последовательности сокращается. Этого можно избежать, если доопределить недостающие отсчеты исходной последовательности. Такое доопределение естественно основывать на том или ином способе экстраполяции последовательности. Можно рекомендовать несколько способов экстраполяции (рис. 4.13).

1. *Дополнение средним значением.* Все недостающие отсчеты считаются равными среднему значению отсчетов последовательности: статистическому среднему или среднему по заданной последовательности. Наиболее часто встречающаяся разновидность этого способа — дополнение нулями. Недостатком его является то, что при экстраполяции средним значением почти всегда имеются большие разрывы в величине отсчетов сигнала на краях последовательности.

2. Доопределение повторением крайних отсчетов последовательности. Это *экстраполяция нулевого порядка.* Она не дает скачков на краях, но повторение крайних отсчетов при экстраполяции на большое число дополнительных отсчетов сильно искажает структуру сигнала и результат фильтрации на краях.

3. *Экстраполяция* более *высокого порядка:* недостающие отсчеты определяются как взвешенная сумма крайних заданных отсчетов последовательности. Веса могут выбираться из статистических соображений по критерию минимума среднеквадратической ошибки экстраполяции, если известна статистическая корреляционная функция сигнала, или классическими методами, сохраняющими значения первых, вторых и более высоких разностей сигнала. Метод неудобен сложностью вычислений.

4. *Четное продолжение.* Этот метод очень прост в вычислительном отношении и в то же время он хорошо сохраняет структуру сигнала и гарантирует отсутствие скачков на краях.

4.4. БЫСТРЫЕ АЛГОРИТМЫ ВЫЧИСЛЕНИЯ ДПФ И СВЕРТКИ СИГНАЛОВ С УМЕНЬШЕННЫМ ЧИСЛОМ УМНОЖЕНИИ

Выше упоминалось, что в существующих цифровых процессорах операции сложения и умножения двух чисел неравноценны по быстродействию. Чаще всего операции умножения требуют в несколько раз больше времени, чем операции сложения. Поэтому представляют интерес появившиеся в последнее время алгоритмы спектрального анализа и свертки сигналов, требующие меньшего числа умножений, чем алгоритмы, основанные на БПФ.

Простейший из них — алгоритм Винограда ускоренной свертки сигналов. Идея его состоит в том, чтобы сократить количество операций умножения ценой некоторого увеличения количества операций сложения.

Рассмотрим формулу дискретной свертки (3.9):

$$b_k = \sum_{n=0}^{N-1} h_n a_{k-n}$$

и будем считать, что заданы все необходимые $N+k$—1 отсчетов $\{a_n\}$. Пусть также N— четное число; если N — нечетное, то будем рассматривать сумму максимального четного числа слагаемых.

Вычислим сначала попарные произведения внутри последовательностей $\{h_h\}$ и $\{a_{k-n}\}$:

$$\eta_l = \sum_{n=0}^{[N/2]-1} h_{2n} h_{2n+1} \qquad (4.33)$$

$$\zeta_k = \sum_{n=0}^{(N/2)-1} a_{k-2n} a_{k-(2n+1)} \qquad (4.34)$$

Тогда, как легко проверить,

$$b_k = \sum_{n=0}^{(N/2)-1} (h_{2n} + a_{k-(2n+1)})(h_{2n+1} + a_{k-2n}) - \eta_i - \zeta_k \tag{4.35}$$

Таким образом, вычисление свертки последовательностей $\{h_n\}$ и $\{a_n\}$ сводится к вычислению свертки вдвое более коротких последовательностей, составленных из попарных сумм элементов исходных последовательностей. Что касается добавочных членов η в (4.3) и ξ_k в (4.3), то η не зависит от k, и ее достаточно вычислить один раз для всех элементов свертки, а ξ_k может вычисляться рекуррентно для четных и нечетных k. Действительно, при k четном ($k = 2k_1$)

$$\zeta_k = \zeta_{2k_1} = \sum_{n=0}^{(N/2)-1} a_{2(k_1-n)} a_{2(k_1-n)+1} = a_{2k_1} a_{2k_1+1} +$$
$$+ \sum_{n=0}^{(N/2)-1} a_{2(k_1-1-n)} a_{2(k_1-1-n)+1} - a_{2(k_1-N/2)} a_{2(k_1-N/2)+1} =$$
$$= a_k a_{k+1} - a_{k-N} a_{k-N+1} + \zeta_{k-2}.$$

Точно такая же формула справедлива для нечетных k.

Оценим количество операций, требуемых для вычисления K значений свертки (время на сложение и вычитание считаем одинаковым). Для вычислений K сумм (4.35) требуется выполнить (3N/2—1) K операций сложения и $KN/2$ операций умножения. Вычисление η требует $N/2$—1 операций сложения и $N/2$ операций умножения. Вычисление $K/2$ четных значений ξ_k требует $N/2$—1 операций сложения и $N/2$ операций умножения для ξ_0 плюс два сложения и два умножения на каждое из $K/2$—1 остальных ξ_k, т.е. $2(K+N/2$—3) операций сложения и $2(K+N/2—2)$ операций умножения. Итого общее количество сложений

$$N = \frac{3NK}{2}\left(1 + \frac{2K + 3N - 7}{3NK}\right)$$

и общее количество умножений

$$N = \frac{NK}{2}\left(1 + \frac{4K + 3N - 8}{KN}\right)$$

Отсюда видно, что по сравнению с прямым вычислением (3.9), требующим $K(N$—1) операций сложения и KN операций умножения, описанный алгоритм требует примерно вдвое меньше умножений и в полтора раза больше сложений. Граница применимости описанного алгоритма определяется соотношением времени сложения и времени умножения конкретного процессора.

Описанный прием в принципе можно использовать и при вычислении двумерной свертки (3.20), но дополнительный выигрыш за счет двумерности здесь получить трудно, поскольку обычно свертка вычисляется одномерно: сначала пробегаются все возможные значения k, после чего l изменяется на единицу.

Для вычисления двумерных сверток Нуссбаумер [26] предложил использовать так называемые полиномиальные преобразования, в которых значительная часть умножений заменяется циклическими сдвигами. По данным [67] использование этих преобразований для вычислений свертки массивов объемом 256x256 элементов дает почти двойную экономию в числе умножений и 30%-ное сокращение числа сложений по сравнению с алгоритмами, использующими БПФ. Правда, реальный выигрыш не так велик: программа на Фортране работает только на 20% быстрее. Более подробные сведения можно найти в [26].

Алгоритм вычисления ДПФ с уменьшенным числом умножений был также разработан Виноградом [20] для случая, когда число отсчетов сигнала является простым числом или степенью простого числа. Он основан на представлении ДПФ как свертки с помощью специальных перестановок и на ускоренном вычислении свертки с уменьшенным числом умножений. В настоящее время известны такие алгоритмы для числа отсчетов 2, 3, 4, 5, 7, 8, 9, 16 [20]. Если N является составным числом, оно раскладывается на простые сомножители из указанного ряда. В соответствии с этими сомножителями матрица ДПФ факторизуется, как в

§4.1, на несколько ДПФ с простым основанием, которые выполняются по ускоренному алгоритму Винограда.

Класс алгоритмов вычисления свертки сигналов, вообще не требующий операций умножения, основан на так называемых теоретико-числовых преобразованиях [25]. В этих алгоритмах вычисления выполняются не как в обычной арифметике, а по некоторому модулю. Благодаря этому умножения заменяются сдвигами и, кроме того, отсутствуют ошибки округления. Однако это преимущество достигается ценой жесткой зависимости значения модуля от длины последовательности, что сильно ограничивает область применения подобных алгоритмов.

Глава 5

ДИСКРЕТНЫЕ ОРТОГОНАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ И БЫСТРЫЕ АЛГОРИТМЫ В МАТРИЧНОМ ПРЕДСТАВЛЕНИИ

5.1. КЛАСС ДИСКРЕТНЫХ ПРЕОБРАЗОВАНИИ, ОБЛАДАЮЩИХ БЫСТРЫМИ АЛГОРИТМАМИ

В последнее время в цифровой обработке изображений кроме дискретных преобразований Фурье и Френеля качали широко использоваться другие ортогональные преобразования, такие как преобразование Хаара, «обобщенные» преобразования Хаара, преобразования Уолша–Адамара – Пэли, слэнт-преобразование, или преобразование по пилообразному базису, преобразование по функциям Виленкина–Крестенсона (ВКФ), гибридные и другие преобразования (см., например, [1]). Они находят применение в оптической пространственной фильтрации, синтезе элементов оптоэлектроннки, в изображающих системах с кодированной апертурой. Главные их применения в обработке изображений – обобщенная фильтрация, в частности фильтрация для подавления помех и коррекции искаженных изображений, кодирование, выделение признаков. Общим замечательным свойством этих преобразований, определяющим целесообразность их применения, является простота их вычислительной реализации. Для всех этих преобразований, как и для дискретного преобразования Фурье, существуют так называемые быстрые алгоритмы (быстрое преобразование Хаара, быстрое преобразование Уолша – Адамара, быстрое слэнт-преобразование и т.д.).

Эти преобразования связывает глубокое родство, которое нагляднее всего выявляется при их матричном представлении. Оно основано на использовании известных в матричной алгебре понятий кронекеровского произведения и прямой суммы матриц, а также на дополнительно введенных понятиях вертикальной суммы матриц, поэтажно-кронекеровских матриц и «элементарных матриц», рассматриваемых в дальнейшем как структурные блоки, из которых строятся матрицы преобразований. Начнем с необходимых определений и обозначений.

Определение 1. Вертикальной суммой двух матриц $\mathbf{M}^{(1)}_{r,s}$ и $\mathbf{M}^{(2)}_{p,s}$ называется матрица $\mathbf{M}_{r+p,s}$, первые r строк которой являются строками матрицы $\mathbf{M}^{(1)}_{r,s}$, а последние p строк – строками матрицы $\mathbf{M}^{(2)}_{p,s}$, Будем обозначать вертикаль ную сумму матриц знаком $\boxed{=}$:

$$\mathbf{M}_{r+p,\,s} = \mathbf{M}^{(1)}_{r,\,s} \boxed{=} \mathbf{M}^{(2)}_{p,\,s}$$

Определение 2. Поэтажно-кронекеровской матрицей называется вертикальная сумма кронекеровских матриц. Обозначения:

$\bigotimes_{n-1}$ – знак кронекеровского произведения двух матриц;

$\bigotimes_{r=0}$ – знак кронекеровского произведения п матриц, нумеруемых по г справа налево;

$\mathbf{M}^{[r]}$ – матрица М в r-й кронекеровской степени;

$\bigoplus_{n-1}$ – знак прямой суммы двух матриц;

$\bigoplus_{r=0}$– знак прямой суммы п матриц, нумеруемых по г справа налево;

$\boxed{\overset{n-1}{=}}_{r=0}$ – знак вертикальной суммы *n* матриц, нумеруемых по *r* сверху вниз;

L_P – единичная матрица размера PxP;

G_a–матрица-строка из а элементов;

$G_a^{(i)}$ –матрица-строка из а элементов с *i*-м элементом, равным единице, и остальными элементами, равными нулю, *i*=0, 1, ..., а-1;

$$\bar{\delta}(r) = \begin{cases} 0, & r = 0, \\ 1, & r \neq 0. \end{cases} \qquad \mathbf{M}^{\bar{\delta}(r)} = \begin{cases} 1, & r = 0; \\ \mathbf{M}, & r \neq 0. \end{cases}$$

Элементарные матрицы:

$$\mathbf{G}_2^0 = [10]; \quad \mathbf{G}_2^1 = [01]; \quad \mathbf{G}_2^2 = [11]; \quad \mathbf{G}_2^3 = [1-1].$$

Вспомогательные матрицы:

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \bar{\mathbf{I}}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad \mathbf{h}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix};$$

$$\mathbf{d}_u = \begin{bmatrix} 1 & 0 \\ 0 & \exp\left(i\dfrac{2\pi u}{2^n}\right) \end{bmatrix};$$

$$\mathbf{\Delta}_r = \delta(r)\,\mathbf{I}_2 = \begin{bmatrix} 0^r & 0 \\ 0 & 0^r \end{bmatrix};$$

$$\mathbf{har}_2^{(2)} = \begin{bmatrix} \mathbf{h}_2 \otimes \mathbf{G}_2^2 \\ \bar{\mathbf{I}}_2 h_2 \mathbf{d}_2 \otimes \mathbf{G}_2^3 \end{bmatrix};$$

$$\mathbf{Sl}_n = \begin{bmatrix} \sqrt{(2^{2n-2}-1)/(2^{2n}-1)} & 2^{n-1}\sqrt{3/2^{2n-1}} \\ -2^{n-1}\sqrt{3/(2^{2n}-1)} & \sqrt{(2^{2n-2}-1)/2^{2n}-1} \end{bmatrix}$$

Нетрудно видеть, что вспомогательные матрицы можно представить в виде вертикальных сумм элементарных матриц, умноженных на скалярную величину. Например, $\mathbf{d}_u = \mathbf{G}_2^0 \boxplus \exp(i\,2\pi u/2^n)\,\mathbf{G}_2^1$. Матрица Sl_n введена в [36].

С помощью этих элементарных и вспомогательных матриц и введенных матричных операций оказывается возможным представить матрицы рассматриваемого класса преобразований в виде поэтажно-кронекеровских матриц. Ниже приведено такое представление для размеров матриц 2^nx2^n.

Дискретное преобразование Фурье. Матрица ДПФ $\mathbf{FOUR}_{2^n} = \{w_n^{kr}\}$, где $w_n = \exp(i2\pi/2^n)$, может быть представлена как поэтажно-кронекеровская, если номера элементов ее строк записать в виде двоичного числа. Действительно, при $r = \sum_{j=0}^{n-1} r_j 2^j$

$$\mathbf{FOUR}_{2^n} = \left\{ w_n^{k\sum_{j=0}^{N-1} r_j 2^j} \right\} = \left\{ \prod_{j=0}^{n-1} w_n^{kr_j 2^j} \right\} = \boxplus_{k=0}^{2^n-1} \bigotimes_{j=0}^{n-1} [1\; w_n^{2^j k}] \tag{5.1}$$

Возможен также другой вариант представления матрицы ДПФ как поэтажно-кронекеровской:

$$\mathbf{FOUR}_{2^n} = \mathbf{M}_{2^n}^{\text{инв}} [\mathbf{G}_2^2 \otimes \mathbf{M}_{2^{n-1}}^{\text{инв}} \mathbf{FOUR}_{2^{n-1}}] \boxplus$$
$$\boxplus \left[\mathbf{G}_2^3 \otimes \mathbf{M}_{2^{n-1}}^{\text{инв}} \mathbf{FOUR}_{2^{n-1}} \left(\bigotimes_{j=0}^{n-2} \mathbf{d}_{2^j} \right) \right] = \mathbf{M}_{2^n}^{\text{инв}} \boxplus_{r=0}^{n} \left((\mathbf{G}_2^2)^{[n-r]} \otimes \right.$$
$$\left. \otimes \left(\mathbf{G}_2^3 \otimes (\mathbf{M}_{2^{r-1}}^{\text{инв}} \mathbf{FOUR}_{2^{r-1}}) \left(\bigotimes_{j=0}^{r-2} \mathbf{d}_{2^j} \right) \right)^{\bar{\delta}(r)} \right), \tag{5.2}$$

где $\mathbf{M}_{2^n}^{\text{инв}}$ –матрица перестановки по закону двоичной инверсии. Она осуществляет перестановку строк умножаемой на нее матрицы в соответствии с их двоично-инвертированным номером, т.е. номером, в котором двоичные разряды представлены в обратном порядке. Она также может быть представлена как коэтажно-кронекеровская матрица:

$$\mathbf{M}_{2^n}^{\text{инв}} = [\mathbf{M}_{2^{n-1}}^{\text{инв}} \otimes \mathbf{G}_2^0] \boxplus [\mathbf{M}_{2^{n-1}}^{\text{инв}} \otimes \mathbf{G}_2^1] = \boxplus_{r=0}^{n} (\mathbf{M}_{2^{r-1}}^{\text{инв}} \otimes \mathbf{G}_2^1)^{\bar{\delta}(r)} \otimes (\mathbf{G}_2^0)^{[n-r]}$$

Выражение (5.2) нетрудно получить из (5.1):

$$\mathbf{FOUR}_{2^n} = \boxplus_{k=0}^{2^n-1} \bigotimes_{j=0}^{n-1} [1 w_n^{2^j k}] = \boxplus_{k_{n-1}=0}^{1} \boxplus_{k_{n-2}=0}^{1} \cdots$$

$$\cdots \boxplus_{k_0=0}^{1} \bigotimes_{j=0}^{2^n-1} [1 w_n^{2^j k}] = \mathbf{M}_{2^n}^{\text{инв}} \boxplus_{k_0=0}^{1} \boxplus_{k_1=0}^{1} \cdots \boxplus_{k_{n-1}=0}^{1} \bigotimes_{j=0}^{n-1} [1 w_n^{2^j k}] =$$

$$= \mathbf{M}_{2^n}^{\text{инв}} \left(\left(\boxplus_{k_1=0}^{1} \cdots \boxplus_{k_{n-1}=0}^{1} \bigotimes_{j=0}^{n-1} [1 w_n^{2^{j+1}[k/2]}] \right) \boxplus \right.$$

$$\left. \boxplus \left(\boxplus_{k_1=0}^{1} \cdots \boxplus_{k_{n-1}=0}^{1} \bigotimes_{j=0}^{n-1} [1 w^{2^j (2[k/2]+1)}] \right) \right) =$$

$$= \mathbf{M}_{2^n}^{\text{инв}} \left(\left(\boxplus_{[k/2]=0}^{2^{n-1}-1} [11] \otimes \bigotimes_{j=0}^{n-2} [1 w_{n-1}^{2^j [k/2]}] \right) \boxplus \right.$$

$$\left. \boxplus \left(\boxplus_{[k/2]=0}^{2^{n-1}-1} [1-1] \otimes \bigotimes_{j=0}^{n-2} [1 w_{n-1}^{2^j [k/2]}] \begin{bmatrix} 1 & 0 \\ 0 & w_n^{2^j} \end{bmatrix} \right) \right) =$$

$$= \mathbf{M}_{2^n}^{\text{инв}} \left(\left([11] \otimes \left(\boxplus_{[k/2]=0}^{2^{n-1}-1} \bigotimes_{j=0}^{n-2} [1 w_{n-1}^{2^j k}] \right) \right) \boxplus \right.$$

$$\left. \boxplus \left([1-1] \otimes \left(\boxplus_{[k/2]=0}^{2^{n-1}-1} \bigotimes_{j=0}^{n-2} [1 w_{n-1}^{2^j}] \right) \left(\bigotimes_{j=0}^{n-2} \mathbf{d}_{2^j} \right) \right) \right) =$$

$$= \mathbf{M}_{2^n}^{\text{инв}} ([\mathbf{G}_2^2 \otimes \mathbf{M}_{2^{n-1}}^{\text{инв}}] \mathbf{FOUR}_{2^{n-1}}] \boxplus$$

$$\boxplus \left[\mathbf{G}_2^3 \otimes \mathbf{M}_{2^{n-1}}^{\text{инв}} \mathbf{FOUR}_{2^{n-1}} \bigotimes_{j=0}^{n-2} \mathbf{d}_{2^j} \right] \Big).$$

Преобразование Уолша – Адамара. Оно описывается матрицей Адамара, являющейся, как было показано еще Гудом[1], кронекеровской матрицей – кро-некеровским произведением вспомогательных матриц

$$\mathbf{HAD}_{2^n} = \mathbf{h}_2^{[n]} = \boxplus_{r=0}^{n} (\mathbf{G}_2^2)^{[n-r]} \otimes (\mathbf{G}_2^3 \otimes \mathbf{h}_2^{[r-1]})^{\bar{\delta}(r)}. \tag{5.3}$$

Модифицированное преобразование Адамара [1].

$$\mathbf{MHAD}_{2^n} = \boxplus_{r=0}^{n} (\mathbf{G}_2^2)^{[n-r]} \otimes (\mathbf{G}_2^3 \otimes 2^{(r-1)/2} \mathbf{I}_2^{[r-1]})^{\bar{\delta}(r)}. \tag{5.4}$$

Преобразования Уолша и Уолша – Пэли, Это производные от преобразования Уолша – Адамара, получающиеся из них с помощью перестановок столбцов матрицы [31]):

$$\mathbf{PAL}_{2^n} = \mathbf{M}_{2^n}^{\text{инв}} \mathbf{HAD}_{2^n};$$

$$\mathbf{WAL}_{2^n} = \mathbf{M}_{2^n}^{\text{гр/пр}} \mathbf{M}_{2^n}^{\text{инв}} \mathbf{HAD}_{2^n}, \tag{5.5}$$

где $\mathbf{M}_{2^n}^{\text{гр/пр}}$ – матрица перестановки, осуществляющая перестановку строк умножаемой на нее матрицы в соответствии с их номерами в прямом двоичном коде, которые получены из их двоичных номеров, рассматриваемых как записанные в коде Грэя. Матрицы $\mathbf{WAL}_{2n}$ и $\mathbf{PAL}_{2n}$ могут быть представлены в виде по-этажно-кронекеровских в рекурсивной и нерекурсивной форме:

$$\mathbf{PAL}_{2^n} = [\mathbf{PAL}_{2^{n-1}} \otimes \mathbf{G}_2^2] \boxplus [\mathbf{PAL}_{2^{n-1}} \otimes \mathbf{G}_2^3] = \boxplus_{r=0}^{n} (\mathbf{PAL}_{2^{r-1}} \otimes \mathbf{G}_2^3)^{\delta(r)} \otimes$$

$$\otimes (\mathbf{G}_2^2)^{[n-r]}; \tag{5.6}$$

$$\mathbf{WAL}_{2^n} = [\mathbf{WAL}_{2^{n-1}} \otimes \mathbf{G}_2^2] \boxplus [(\bar{\mathbf{I}}_2^{[n-1]} \mathbf{WAL}_{2^{n-1}}) \otimes \mathbf{G}_2^3] =$$

$$= \boxplus_{r=0}^{n} (\bar{\mathbf{I}}_2^{[r-1]} \mathbf{WAL}_{2^{r-1}}) \otimes \mathbf{G}_2^3)^{\bar{\delta}(r)} \otimes (\mathbf{G}_2^2)^{[n-r]}. \tag{5.7}$$

Преобразование Хаара. В принятых нами обозначениях матрица преобразования Хаара HAR_{2n} размера 2^nx2^n может быть записана в рекурсивной и нерекурсивной форме:

$$\mathbf{HAR}_{2^n} = [\mathbf{HAR}_{2^{n-1}} \otimes \mathbf{G}_2^2] \boxplus 2^{(n-1)/2} [\mathbf{I}_2^{[n-1]} \otimes \mathbf{G}_2^3] =$$

$$= \boxplus_{r=0}^{n} (2^{(r-1)/2} \mathbf{I}_2^{[r-1]} \otimes \mathbf{G}_2^3)^{\bar{\delta}(r)} \otimes (\mathbf{G}_2)^{[n-r]}. \tag{5.8}$$

Комплексное преобразование Хаара. Впервые оно было введено с помощью рекурсивного определения. В наших обозначениях записывается так:

$$\mathbf{CHAR}_{2^n} = [\mathbf{CHAR}_{2^{n-1}} \otimes \mathbf{G}_2^2] \prod 2^{(n-2)/2} [(\mathbf{h}_2\mathbf{d}_2 \otimes \mathbf{I}_{2^{n-2}} \otimes \mathbf{G}_2^3] =$$

$$= \prod_{r=0}^{n-1} (\mathbf{h}_2\mathbf{d}_2 \otimes 2^{r/2}\mathbf{I}_2^{[r]})^{\overline{\delta}(r-1)} \otimes \mathbf{G}_2^{3\overline{\delta}(r)} \otimes (\mathbf{G}_2^2)^{[n-r]}. \quad (5.9)$$

Обобщенное преобразование Хаара. Преобразование первого порядка:

$$\mathbf{CHAR}_{2^n}^{(1)} = [(\mathbf{HAR}_{2^{n-1}}^{(1)}) \otimes \mathbf{G}_2^2] \prod 2^{(n-2)/2} [(\overline{\mathbf{I}}_2\mathbf{h}_2\mathbf{d}_2) \otimes \mathbf{I}_2^{[n-2]} \otimes \mathbf{G}_2^3] =$$

$$= \prod_{r=0}^{n} ((\overline{\mathbf{I}}_2\mathbf{h}_2\mathbf{d}_2 \otimes 2^{r/2}\mathbf{I}_2^{[r]})^{\overline{\delta}(r-1)} \otimes \mathbf{G}_2^{3\overline{\delta}(r)} \otimes (\mathbf{G}_2^2)^{[n-r]}). \quad (5.10)$$

Преобразование второго порядка:

$$\mathbf{CHAR}_{2^n}^{(2)} = [(\mathbf{HAR}_{2^{n-1}}^{(2)}) \otimes \mathbf{G}_2^2] \prod [(\mathbf{har}_{2^2}^{(2)})\ (\mathbf{d}_6 \otimes \mathbf{d}_4)\ (\mathbf{M}_{2^2}^{\text{пр/пр}} \mathbf{M}_{2^2}^{\text{инв}}) \otimes$$

$$\otimes (2^{\frac{n-3}{2}} \mathbf{I}_2^{[n-3]}) \otimes \mathbf{G}_2^3] = \prod_{r=0}^{n} (((\mathbf{har}_{2^2}^{(2)})\ (\mathbf{d}_8 \otimes \mathbf{d}_4)\ \mathbf{M}_{2^2}^{\text{пр/пр}} \mathbf{M}_{2^2}^{\text{инв}})^{\overline{\delta}(r-2)} \otimes$$

$$\otimes \overline{\mathbf{I}}_2\mathbf{h}_2\mathbf{d}_2)^{\overline{\delta}(r-1)} \otimes \mathbf{G}_2^{3\overline{\delta}(r)} \otimes (\mathbf{G}_2^2)^{[n-r]}). \quad (5.11)$$

Преобразование Адамара – Хаара:

$$\mathbf{HDHR}_{2^n}^{(m)} = \mathbf{HAD}_{2^m} \otimes \mathbf{HAR}_{2^{n-m}} = \mathbf{h}_2^{[m]} \otimes$$

$$\otimes \prod_{r=0}^{n-m} (2^{(r-1)/2} \mathbf{I}_2^{[r-1]} \otimes \mathbf{G}_2^3)^{\overline{\delta}(r)} \otimes (\mathbf{G}_2^2)^{[n-m-r]}. \quad (5.12)$$

S -преобразования:

$$\mathbf{S}_{2^n} = \prod_{r_1=0}^{(n/2)-1} \left[\left(\prod_{r_2=0}^{1} \left[\frac{\Delta_{r_2}^{[2(r_1-1)]} \otimes \mathbf{G}_2^2}{(2^{r_1-1}\mathbf{I}_2^{[2(r_1-1)]}) \otimes \mathbf{G}_2^3} \right] \right) \otimes \mathbf{G}_2^3 \right]^{\overline{\delta}(r_1)} \otimes (\mathbf{G}_2)^{[n-2r_1]}. \quad (5.13)$$

Преобразование по ВКФ впервые введено под названием «обобщенные преобразования» с помощью кронекеровского произведения матриц, подробно исследовано в [31]. В матричной форме, очевидно, записывается подобно преобразованию Уолша – Адамара:

$$\mathbf{VCP}_{2^n,m}^{(m)} = (\mathbf{FOUR}_{2^n})^{[m]}. \quad (5.14)$$

Преобразование перехода HAR–PAL. Это преобразование, которое позволяет по спектру сигнала в базисе Хаара найти его спектр в базисе фукции Уолша в упорядочении Пэли (спектр преобразования Пэли):

$$\mathbf{M}_{2^n}^{\text{HR}-\text{P}} = \prod_{r=0}^{n} (\mathbf{G}_2^0)^{[n-r]} \otimes (\mathbf{G}_2^1 \otimes \mathbf{M}_{2^{r-1}}^{\text{HR}-\text{P}} \mathbf{D}_{2^{r-1}} \mathbf{HAR}_{2^{r-1}})^{\delta(r)}, \quad (5.15)$$

где

$$\mathbf{D}_{2^r} = 2^{r/2} \left(1 \otimes \bigotimes_{k=0}^{r-1} 2^{-k/2} \mathbf{I}_2^{[k]} \right). \quad (5.16)$$

Преобразование перехода MHAD–HAD. Это преобразование, позволяющее по спектру модифицированного преобразования Адамара найти спектр собственно преобразования Адамара:

$$\mathbf{M}_{2^n}^{\text{MH}-\text{H}} = \prod_{r=0}^{n} \mathbf{G}_2^0 \otimes (\mathbf{G}_2^1 \otimes \mathbf{M}_{2^{r-1}}^{\text{MH}-\text{H}} \mathbf{D}_{2^{r-1}} \mathbf{MHAD}_{2^{r-1}})^{\overline{\delta}(r)}. \quad (5.17)$$

Преобразование перехода MHAD–FOUR. Структура матрицы преобразования MHAD–FOUR перехода от спектра модифицированного преобразования
Адамара к спектру ДПФ ввиду очевидной аналогии между матрицами HAD_{2n} и FOUR_{2n} сходна со структурой матрицы $\mathrm{M}_{2n}^{\text{MH-M}}$:

$$\mathbf{M}_{2^n}^{\text{MH}-\text{F}} = \prod_{r=0}^{n} \mathbf{G}_2^0 \otimes \left(\mathbf{G}_2^1 \otimes (\mathbf{M}_{2^{r-1}}^{\text{MH}-\text{F}} \mathbf{D}_{2^{r-1}} \mathbf{MHAD}_{2^{r-1}}) \left(\bigotimes_{j=0}^{r-2} \mathbf{d}_{2^{n-j-2}} \right) \right)^{\overline{\delta}(r)}. \quad (5.18)$$

Слэнт-преобразование (преобразование по пилообразному базису, или наклонное преобразование). Впервые в общей форме введено для кодирования изображений. В матричной форме более подробно исследовано в [36]. С помощью введенной в этой работе вспомогательной матрицы $\mathbf{Sl}_n$ матрица слэнт-преобразования может быть представлена в рекуррентной и нерекуррентной форме как поэтажно-кронекеровская:

$$\mathbf{SL}_{2^n} = \mathbf{M}_{2^n}^{\text{инв}} \{[\mathbf{G}_2^2 \otimes \mathbf{M}_{2^{n-1}}^{\text{инв}}] \mathbf{SL}_{2^{n-1}} \mathbf{M}_{2^{n-1}}^{\text{пр/гр}}] \prod [\mathbf{G}_2^3 \otimes \mathbf{M}_{2^{n-1}}^{\text{инв}} \times$$

$$\times [\mathbf{sl}_n \otimes \mathbf{I}_{2^{n-1}}] \mathbf{SL}_{2^{n-1}} \mathbf{M}_{2^{n-1}}^{\text{пр/гр}}]\} \mathbf{M}_{2^n}^{\text{пр/гр}} = \mathbf{M}_{2^n}^{\text{инв}} \left(\prod_{r=0}^{n} (\mathbf{G}_2)^{[n-r]} \otimes (\mathbf{G}_2^3 \otimes \right.$$

$$\left. \otimes \mathbf{M}_2^{\text{инв}} (\mathbf{sl}_r \otimes \mathbf{I}_{2^{r-1}-2}) \mathbf{SL}_{2^{r-1}} \mathbf{M}_{2^{r-1}}^{\text{пр/гр}}) \right)^{\overrightarrow{\delta}_r}. \qquad (5.19)$$

Матрицы перестановок. Пусть $V_s=\{v_i\}$, i=0, 1, ..., s-1–вектор данных длиной s, причем s – составное число: $s = s_1 \cdot s_r \bullet\bullet\bullet s_n$. Индекс i можно представить в лексикографической форме по смешанному основанию $\{s_j\}$:

$$i = \sum_{j=0}^{n-1} l_{j+1} \prod_{k=0}^{j} s_k, \qquad (5.20)$$

где s_0=l, l_i=0, 1,..., s_j -1. В случае инверсной перестановки по смешанному основанию $\{s_j\}$ индекс i_p элемента, который должен после инверсной перестановки занять место i, выражается формулой:

$$i_p = \sum_{j=0}^{n-1} l_{n-j} \prod_{k=n-j+1}^{n+1} s_k, \qquad (5.21)$$

Здесь значения l_{n-j} такие же, как и в формуле (5.20), а s_{n+1}=l. Матричным аналогом выражения (5.20) будет выражение:

$$\mathbf{G}_s^{(i)} = \mathbf{G}_{s_n}^{(l_n)} \otimes \mathbf{G}_{s_{n-1}}^{(l_{n-1})} \otimes \ldots \otimes \mathbf{G}_{s_1}^{(l_1)}, \qquad (5.22)$$

а (5.21) – выражение

$$\mathbf{G}_s^{(i_p)} = \mathbf{G}_{s_1}^{(l_1)} \otimes \mathbf{G}_{s_2}^{(l_2)} \otimes \ldots \otimes \mathbf{G}_{s_n}^{(l_n)}. \qquad (5.23)$$

Пусть P–матрица инверсной перестановки элементов вектора $\mathbf{V_s}$, $s=s_1 \cdot s_2 \ldots s_n$ по смешанному основанию. Очевидно, ее можно представить как вертикальную сумму вида:

$$\mathbf{P} = \prod_{i=0}^{s-1} \mathbf{G}_s^{(i_p)}. \qquad (5.24)$$

Подставив в (5.24) формулу (5.23) и представив вертикальную сумму по i в виде n вертикальных сумм по l_j, в соответствии с (5.20) получим

$$\mathbf{P} = \prod_{l_n=0}^{s_n-1} \prod_{l_{n-1}=0}^{s_{n-1}-1} \ldots \prod_{l_1=}^{s_1-1} \mathbf{G}_{s_1}^{(l_1)} \otimes \mathbf{G}_{s_2}^{(l_2)} \otimes \ldots \otimes \mathbf{G}_{s_n}^{(l_n)}. \qquad (5.25)$$

Перестановка в (5.25) определяется расположением множителей s_1 s_2, ..., s_n в обратном порядке, что наглядно видно при сравнении (5.22) и (5.23). Естественно рассмотреть более широкий класс перестановок данных, которые определяются произвольным изменением порядка следования множителей s_1, s_2, ..., s_n. Точнее, пусть $[\alpha]=\{1, 2, \ldots, n\}$, а $\sigma[\alpha]=\{\sigma(1), \ldots, \sigma(n)\}$ описывает произвольную перестановку мест расположения элементов в системе [a], где $\sigma(i) \in \{1, 2, \ldots, n\}$. Тогда матрица перестановки $\mathbf{P}(\sigma[\alpha]/[\alpha])$, соответствующая $\sigma[\alpha]$, имеет вид:

$$\mathbf{P}_S(\sigma[\alpha]/[\alpha]) = \prod_{l_n=0}^{s_n-1} \prod_{l_{n-1}=0}^{s_{n-1}-1} \ldots \prod_{l_1=0}^{s_1-1} \mathbf{G}_{s_{\sigma(n)}}^{(l_{\sigma(n)})} \otimes \ldots \otimes \mathbf{G}_{s_{\sigma(1)}}^{(l_{\sigma(1)})}. \qquad (3.26)$$

Приведем пример использования этого класса перестановок. Пусть

$$\mathbf{B} = \bigotimes_{i=1}^{n} \mathbf{A}_{q_i, r_i}, \qquad (5.27)$$

а $\mathbf{B}\mathbf{p}$ – матрица, получаемая при перестановке сомножителей в (5.27) в соответствии с перестановкой в $\sigma[\alpha]$:

$$\mathbf{B}_p = \bigotimes_{i=1}^{n} \mathbf{A}_{q_{\sigma(i)} r_{\sigma(i)}}$$

Тогда матрицы B и B_p связаны следующим соотношением:

$$\mathbf{B} = \mathbf{P}_Q(\sigma[\alpha]/[\alpha])\,\mathbf{B}_p\mathbf{P}_R^{\mathrm{T}}(\sigma[\alpha]/[\alpha]) \quad (5.28)$$

где $Q = \prod_{i=1}^{n} q_i$, $R = \prod_{i=1}^{n} r_i$, а $\mathbf{P}_R^{\mathrm{T}}(\sigma[\alpha]/[\alpha])$ – транспонированная матрица перестановки.

Матрицы произвольных перестановок обладают свойством самосопряженности. Поэтому легко можно записать явную форму матрицы $\mathbf{P}_R^{\mathrm{T}}(\sigma[\alpha]/[\alpha])$. Так как транспонированная матрица перестановки совпадает с обратной, то она должна совпадать с матрицей перестановки, определяемой перестановкой $\sigma^{-1}[\sigma[\alpha]]$ элементов системы $\sigma[\alpha]$ в исходное расположение $[\alpha]$. Согласно выражению (5.26):

$$\mathbf{P}_R^{\mathrm{T}}(\sigma[\alpha]/[\alpha]) = \mathbf{P}_R^{-1}(\sigma[\alpha]/[\alpha]) = \mathbf{P}_R([\alpha]/\sigma[\alpha]) =$$

$$= \sum_{l_{\sigma(n)}=0}^{r_{\sigma(n)}-1} \sum_{l_{\sigma(n-1)}=0}^{r_{\sigma(n-1)}-1} \dots \sum_{l_{\sigma(1)}=0}^{r_{\sigma(1)}-1} \mathbf{G}_{r_n}^{(l_n)} \otimes \mathbf{G}_{r_{n-1}}^{(l_{n-1})} \otimes \dots \otimes \mathbf{G}_{r_1}^{(l_1)}.$$

Выражение (5.28), в частности перестановку местами в прямом произведении двух матриц, можно использовать для факторизации матриц транспонирования. Разбиение матрицы перестановки на несколько множителей необходимо для представления алгоритмов транспонирования больших матриц, не помещающихся целиком в оперативной памяти ЭВМ.

Приведенное представление матриц ортогональных преобразований благодаря своей наглядности создает удобную основу для систематизации ортогональных преобразований, имеющих быстрые алгоритмы. Например, сопоставляя формулы (5.2), (5.3), (5.4), (5.8) легко видеть, как, изменяя порядок элементарных матриц в кронекеровском произведении, можно из матрицы HAR_{2n} получить матрицу $\mathbf{MHAD}_{2n}$, затем, заменяя элементарные матрицы $\mathbf{I}_2$ на $\mathbf{h}_2$ получить матрицу HAD_{2n} и, наконец, вводя диагональные элементарные матрицы $\mathbf{d}_u$, получить матрицу $FOUR_{2n}$ с двоичной инверсией строк. Очень поучительно сопоставление матриц $\mathbf{M}_{2^n}^{\text{нив}}$ и $\mathbf{PAL}_{2^n}$, $\mathbf{M}_{2^n}^{\text{гр/пр}}$ и $\mathbf{WAL}_{2^n}$, $\mathbf{FOUR}_{2^n}$ и $\mathbf{SLANT}_{2^n}$ также наглядно показывающее, какую роль играют в построении матриц преобразований элементарные матрицы и т.д. Кроме того, как будет показано ниже, это представление наряду с теоремами факторизации, приведенными в следующем параграфе, позволяет легко строить быстрые алгоритмы в их наиболее удобном матричном представлении, пользуясь аппаратом матричной алгебры. Оно также позволяет глубже понять, как должны строиться дискретные преобразования, если желательно, чтобы для их выполнения существовали быстрые алгоритмы. В частности, из него вытекает, что новые виды преобразований, гарантированно обладающие быстрым алгоритмом, можно строить, вводя новые виды элементарных матриц или комбинируя их в одном преобразовании, а также комбинируя различные поэтажно-кронекеровские матрицы с помощью вертикальной суммы и кронекеровского произведения.

5.2. ЭЛЕМЕНТЫ МАТРИЧНОГО АППАРАТА ВЫВОДА БЫСТРЫХ АЛГОРИТМОВ

На матричном языке утверждение, что преобразование, описываемое матрицей размером NxN, может быть вычислено с помощью быстрых алгоритмов, означает, что матрица преобразования факторизуется в произведение $\log_2 N$ слабо заполненных матриц. Возможность факторизации матриц основана на следующих теоремах.

Теорема 1.

$$\sum_{k=0}^{n-1} (\mathbf{M}_{r_k, s_k}^{k} \mathbf{N}_{s_k, q}^{(k)}) = \left(\bigoplus_{k=0}^{n-1} \mathbf{M}_{r_k, s_k}^{(k)}\right)\left(\sum_{k=0}^{n-1} \mathbf{N}_{s_k, q}\right).$$

Доказательство этой теоремы очевидно из рис. 5.1, а.

Следствие 1.

$$\sum_{k=0}^{n-1} (\mathbf{M}_{r_k, s_k}^{(k)}) = \left(\bigoplus_{k=0}^{n-1} \mathbf{M}_{r_k, s_k}^{(k)}\right)\left(\sum_{k=0}^{n-1} \mathbf{I}_{s_k}\right).$$

Это следствие непосредственно вытекает из теоремы 1 в частном случае, когда $\mathbf{N}_{s_k,q} = \mathbf{I}_{s_k,q}$.

Теорема 2.

$$\mathbf{M}_{r,s} \otimes \mathbf{N}_{p,q} = (\mathbf{M}_{r,s} \otimes \mathbf{I}_p)\ (\mathbf{I}_s \otimes \mathbf{N}_{p,q})$$

Доказательство теоремы иллюстрируется рис. 5.1,6.

Из теорем 1, 2 и следствия 1 непосредственно вытекают:

Следствие 2. При $\mathbf{N}^{(k)}_{s_k,q} = \mathbf{N}_{s,q}$

$$\boxplus_{k=0}^{n-1} (\mathbf{M}^{(k)}_{r_k,s} \mathbf{N}_{s,q}) = \left(\boxplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r_k,s}\right) \mathbf{N}_{s,q}$$

(это соотношение легко также проверить непосредственно).

Следствие 3.

$$\boxplus_{k=0}^{n-1} (\mathbf{M}^{(k)}_{r_k,s_k} \otimes \mathbf{N}^{(k)}_{p_k,q_k}) = \left(\bigoplus_{k=0}^{n-1} (\mathbf{M}^{(k)}_{r_k,s_k} \otimes \mathbf{I}_{p,k})\right) \boxplus_{k=0}^{n-1} (\mathbf{I}_{s_k} \otimes \mathbf{N}^{(k)}_{p_k,q_k}).$$

Следствие 4. Из следствия 3 непосредственно вытекает:

$$\boxplus_{k=0}^{n-1} (\mathbf{M}^{(k)}_{r_k,s_k} \otimes \mathbf{G}^{(k)}_{q_k}) = \left(\bigoplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r_k,s_k}\right) \left(\boxplus_{r=0}^{n-1} (\mathbf{I}_{s_k} \otimes \mathbf{G}^{(k)}_{q_k})\right),$$

где $\mathbf{G}^{(k)}_{q_k}$ – матрицы-строки размером q_k, элементов.

Следствие 5. Пусть $\mathbf{M}_{r,s} = \boxplus_{k=0}^{r-1} \mathbf{G}^{(M),k}_{s}$; $\mathbf{N}_{p,q} = \boxplus_{n=0}^{p-1} \mathbf{G}^{(N),n}_{q}$.

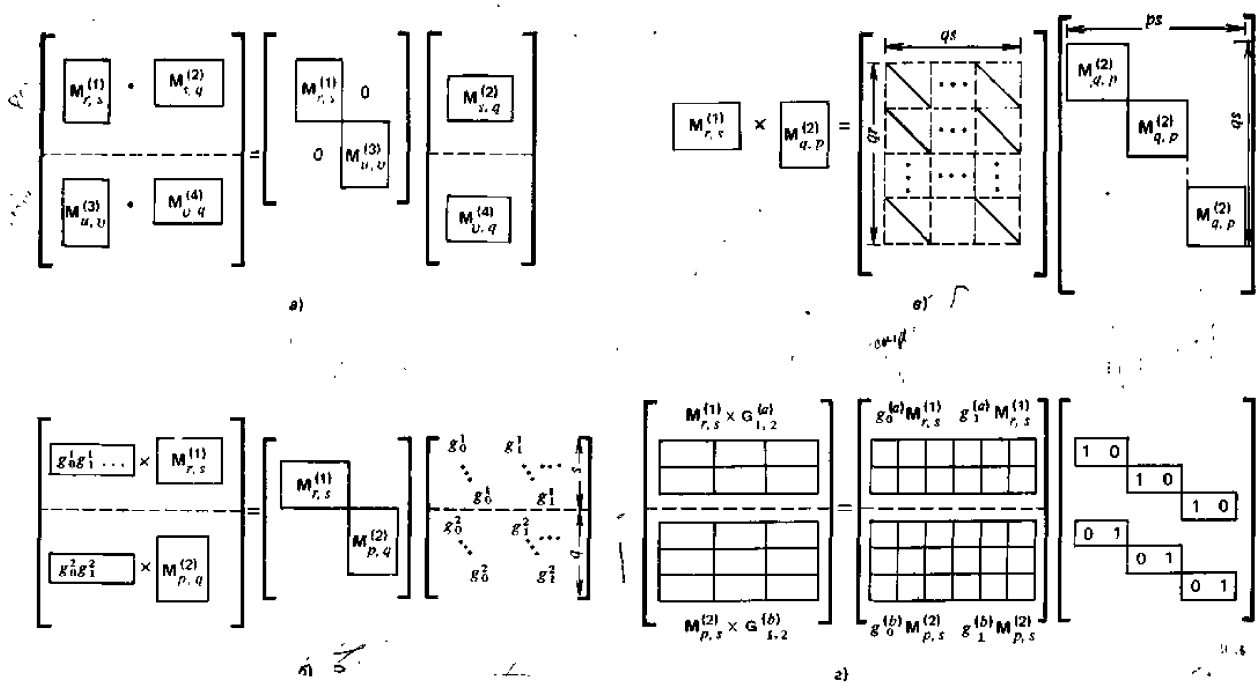


Рис. 5.1. К доказательству теорем факторизации

Тогда

$$\mathbf{M}_{r,s}\otimes \mathbf{N}_{p,q} = \left(\boxplus_{k=0}^{r-1} (\mathbf{I}_p \otimes \mathbf{G}_s^{(M),k})\right)\left(\boxplus_{n=0}^{p-1} (\mathbf{I}_s \otimes \mathbf{G}_q^{(N),n})\right).$$

Действительно,

$$\mathbf{M}_{r,s}\otimes \mathbf{N}_{p,q} = \left(\boxplus_{k=0}^{r-1} \mathbf{G}_s^{(M),k}\right)\otimes \left(\boxplus_{k=0}^{p-1} \mathbf{G}_q^{(N),n}\right) =$$

$$= \boxplus_{k=0}^{r-1}\left(\boxplus_{n=0}^{p-1} \mathbf{G}_s^{(M),k}\mathbf{G}_q^{(N),n}\right) = \boxplus_{k=0}^{r-1}\left(\bigotimes_{n=0}^{p-1} \mathbf{G}_s^{(M),k}\right)\left(\boxplus_{n=0}^{p-1} (\mathbf{I}_s \otimes \mathbf{G}_q^{(N),n})\right) =$$

$$= \left(\boxplus_{k=0}^{r-1} (\mathbf{I}_p \otimes \mathbf{G}_s^{(M),k})\right)\left(\boxplus_{n=0}^{p-1} (\mathbf{I}_s \otimes \mathbf{G}_q^{(N),n})\right).$$

Здесь при переходе от второго равенства к третьему использована теорема 1 для вертикальной суммы матриц по n, а при переходе от третьего равенства к четвертому– следствие 2 для вертикальной суммы матриц по k и очевидное соотношение для прямой суммы одинаковых матриц по n.

Для матриц-строк верны также:

Теорема 3.

$$\boxplus_{k=0}^{n-1} (\mathbf{G}_{q_k}^{(k)} \otimes \mathbf{M}_{r_k,\, s_k}^{(k)}) = \left(\bigoplus_{k=0}^{n-1} \mathbf{M}_{r_k,\, s_k}^{(k)}\right)\left(\boxplus_{k=0}^{n-1} (\mathbf{G}_{q_k}^{(k)} \otimes \mathbf{I}_{s_k})\right).$$

Доказательство этой теоремы иллюстрируется рис. 5.1, а.

Теоремы перестановок.

$$1.\ \mathbf{M}_{r,s}\otimes \mathbf{N}_{p,q} = \left(\boxplus_{n=0}^{r-1} (\mathbf{I}_p \otimes \mathbf{G}_r^{(n)})\right)(\mathbf{N}_{p,q}\otimes \mathbf{M}_{r,s})\left(\boxplus_{k=0}^{q-1} (\mathbf{I}_s \otimes \mathbf{G}_q^{(k)})\right),$$

где $\mathbf{G}_r^{(n)}$ –матрица-строка из r элементов $[\delta(r\text{-}n)]$.

$$2.\ \boxplus_{k=0}^{n-1} (\mathbf{M}_{r_k,s}^{(k)} \otimes \mathbf{G}_q^{(k)}) = \left(\boxplus_{k=0}^{n-1} \mathbf{G}_q^{(k)} \otimes \mathbf{M}_{r_k,\, s}^{(k)}\right)\left(\boxplus_{m=0}^{q-1} \mathbf{I}_s \otimes \mathbf{G}_q^{(m)}\right)$$

3. В частном случае n=2:

$$(\mathbf{M}_{r,s}^{(1)} \otimes \mathbf{G}_2^{(1)}) \boxplus (\mathbf{M}_{p,s}^{(2)} \otimes \mathbf{G}_2^{(2)}) = ((\mathbf{G}_2^{(1)} \otimes \mathbf{M}_{r,s}^{(1)}) \boxplus (\mathbf{G}_2^{(2)} \otimes \mathbf{M}_{p,s}^{(2)})) \times$$

$$\times ((\mathbf{I}_s \otimes \mathbf{G}_2^{0}) \boxplus (\mathbf{I}_s \otimes \mathbf{G}_2^{1})).$$

Доказательство этого тождества иллюстрируется рис. 5.1, г.

Теоремы 2–3. а также следствия 3, 4, 5 являются теоремами факторизации матриц в произведение слабо заполненных матриц. Например, по теореме 3 нетрудно найти, что количество действий сложения (вычитания) и умножения при умножении вектора на нефакторизованную матрицу в левой части равно $\sum_{k=0}^{n-1} s_k (q_k r_k)$, тогда как при последовательном умножении вектора на матрицы правой части требуется выполнить

$$\sum_{k=0}^{n-1} q_k s_k + \sum_{k=0}^{n-1} r_k s_k = \sum_{k=0}^{n-1} s_k (q_k + r_k)$$

операций, т.е. при $q_k, r_k > 2$ требуемое число операции уменьшается.

В дальнейшем нам понадобятся также следующие свойства прямой суммы, вертикальной суммы и прямого произведения матриц:

$$\left(\bigoplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r_k, s_k}\right)\left(\bigoplus_{k=0}^{n-1} \mathbf{N}^{(k)}_{s_k, q_k}\right) = \bigoplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r_k, s_k} \mathbf{N}^{(k)}_{s_k, q_k}; \tag{5.29}$$

$$\left(\prod_{k=0}^{n-1} \mathbf{M}^{(k)}\right) \otimes \left(\prod_{k=0}^{n-1} \mathbf{N}^{(k)}\right) = \prod_{k=0}^{n-1} \mathbf{M}^{(k)} \otimes \mathbf{N}^{(k)}; \tag{5.30}$$

$$\left(\bigotimes_{k=0}^{n-1} \mathbf{M}^{(k)}\right)\left(\bigotimes_{k=0}^{n-1} \mathbf{N}^{(k)}\right) = \bigotimes_{k=0}^{n-1} \mathbf{M}^{(k)} \mathbf{N}^{(k)}; \tag{5.31}$$

$$\boxplus_{k=0}^{n-1} (\mathbf{M}^{(k)}_{r_k, s_k} \otimes \mathbf{N}) = \left(\boxplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r_k, s_k}\right) \otimes \mathbf{N}; \tag{5.32}$$

$$\boxplus_{k=0}^{n-1} (\mathbf{G}_s \otimes \mathbf{M}^{(k)}_s) = \mathbf{G}_s \otimes \boxplus_{k=0}^{N-1} \mathbf{M}^{(k)}_{s,q}; \tag{5.33}$$

$$\boxplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r,s} \mathbf{N}_{s,p} = \left(\boxplus_{k=0}^{n-1} \mathbf{M}^{(k)}_{r,s}\right) \mathbf{N}_{s,p}; \tag{5.34}$$

$$\boxplus_{k=0}^{n-1} \mathbf{M}_{s,p} \mathbf{N}^{(k)}_{p,r} = (\mathbf{I}_n \otimes \mathbf{M}_{s,p}) \cdot \left(\boxplus_{k=0}^{n-1} \mathbf{N}^{(k)}_{p,r}\right). \tag{5.35}$$

5.3. АЛГОРИТМЫ БЫСТРОГО ПРЕОБРАЗОВАНИЯ ФУРЬЕ В МАТРИЧНОМ ПРЕДСТАВЛЕНИИ

Покажем, как алгоритмы быстрого преобразования Фурье (БПФ) вытекают из представления матрицы дискретного преобразования Фурье в виде поэтажно-кронекеровской (5.2) и из теорем факторизации.

Для упрощения записи дальнейших выкладок введем обозначение:

$$\overline{\mathbf{FOUR}}_{2^n} = 2^{n/2} \mathbf{M}^{\text{инв}}_{2^n} \mathbf{FOUR}_{2^n}.$$

Тогда (5.2) можно переписать в виде: •

$$\overline{\mathbf{FOUR}}_{2^n} = [\mathbf{G}^2_2 \otimes \overline{\mathbf{FOUR}}_{2^{n-1}}] \boxplus \left[\mathbf{G}^3_2 \otimes (\overline{\mathbf{FOUR}}_{2^{n-1}}) \left(\bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}}\right)\right].$$

Применив к этому выражению, например, теорему 3 из § 5.2, получим

$$\overline{\mathbf{FOUR}}_{2^n} = \left[\overline{\mathbf{FOUR}}_{2^{n-1}} \oplus \overline{\mathbf{FOUR}}_{2^{n-1}} \bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}}\right] ([\mathbf{G}^2_2 \otimes \mathbf{I}_{2^{n-1}}] \boxplus$$

$$\boxplus [\mathbf{G}^3_2 \otimes \mathbf{I}_{2^{n-1}}]),$$

или благодаря (5.22)

$$\overline{\mathbf{FOUR}}_{2^n} = (\mathbf{I}_2 \otimes \overline{\mathbf{FOUR}}_{2^{n-1}}) \left[\mathbf{I}_{2^{n-1}} \oplus \bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}}\right] \times$$

$$\times ([\mathbf{G}^2_2 \otimes \mathbf{I}_{2^{n-1}}] \boxplus [\mathbf{G}^3_2 \otimes \mathbf{I}_{2^{n-1}}]).$$

По определению прямого произведения матриц, подматрицу в последнем сомножителе можно вынести за знак матрицы:

$$\overline{\mathbf{FOUR}}_{2^n} = (\mathbf{I}_2 \otimes \mathbf{FOUR}_{2^{n-1}}) \left[\mathbf{I}_{2^{n-1}} \oplus \bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}} \right] (\mathbf{h}_2 \otimes \mathbf{I}_{2^{n-1}}).$$

Подставив теперь сюда вместо $\mathbf{FOUR}_{2n-1}$ аналогичную формулу с заменой n на $n-1$, получим:

$$\mathbf{FOUR}_{2^n} = \left\{ \mathbf{I}_2 \otimes (\mathbf{I}_2 \otimes \overline{\mathbf{FOUR}}_{2^{n-2}}) \left[\mathbf{I}_{2^{n-2}} \oplus \bigotimes_{s=0}^{n-3} \mathbf{d}_{2^{n-s-2}} \right] \times \right.$$
$$\left. \times (\mathbf{h}_2 \otimes \mathbf{I}_{2^{n-2}}) \right\} \times \left[\mathbf{I}_{2^{n-1}} \oplus \bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}} \right] (\mathbf{h}_2 \otimes \mathbf{I}_{2^{n-1}}),$$

или, представив первую слева матрицу $\mathbf{I}_2$ в кронекеровском произведении в фигурных скобках в виде $\mathbf{I}_2 = \mathbf{I}_2\mathbf{I}_2$ и воспользовавшись (5.23),

$$\overline{\mathbf{FOUR}}_{2^n} = (\mathbf{I}_2 \otimes \mathbf{I}_2 \otimes \overline{\mathbf{FOUR}}_{2^{n-2}}) \left(\mathbf{I}_2 \otimes \left[\mathbf{I}_{2^{n-2}} \oplus \bigotimes_{s=0}^{n-3} \mathbf{d}_{2^{n-s-2}} \right] \right) \times$$
$$\times (\mathbf{I}_2 \otimes \mathbf{h}_2 \otimes \mathbf{I}_{2^{n-2}}) \left(\mathbf{I}_{2^{n-1}} \oplus \bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}} \right) (\mathbf{h}_2 \otimes \mathbf{I}_{2^{n-1}}).$$

Продолжая подобные преобразования с матрицами рекурсивно, получим окончательно следующее факторизованное представление матрицы $\mathbf{FOUR}_{2n}$:

$$\mathbf{FOUR}_{2^n} = 2^{-n/2} \mathbf{M}_{2^n}^{\text{инв}} \prod_{r=0}^{n-1} \left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^r} \oplus \bigotimes_{s=0}^{r-1} \mathbf{d}_{2^{n-2-s}} \right] \right) \times$$
$$\times (\mathbf{I}_{2^{n-1-r}} \otimes \mathbf{h}_2 \otimes \mathbf{I}_{2^r}).$$

Эта формула является матричной записью одного из алгоритмов БПФ с двоичной инверсией результата преобразования. Выполняя над ней тождественные матричные преобразования, можно получить любые другие алгоритмы БПФ. Например, транспонировав ее и воспользовавшись симметричностью матрицы $\mathbf{FOUR}_{2n}$, получим алгоритм БПФ с двоичной инверсией последовательности отсчетов исходного сигнала:

$$\mathbf{FOUR}_{2^n} = 2^{-n/2} \left(\prod_{r=0}^{n-2} (\mathbf{I}_{2^r} \otimes \mathbf{h}_2 \otimes \mathbf{I}_{2^{n-1-r}} \right) \times$$
$$\times \left(\mathbf{I}_{2^r} \otimes \left[\mathbf{I}_{2^{n-1-r}} \oplus \bigotimes_{s=0}^{n-2-r} \mathbf{d}_{2^{n-s-2}} \right] \right) (\mathbf{I}_{2^{n-1}} \otimes \mathbf{h}_2)\, \mathbf{M}_{2^n}^{\text{инв}}.$$

Следует отметить, что операции, требуемые для ДПФ сигналов, являются операциями над комплексными числами. Между тем элементарными операциями используемых цнфрозых процессоров являются операции над вещественными числами. Поэтому для получения алгоритма БПФ в терминах элементарных операций над вещественными числами формулы (5.32) и (5.33) нужно преобразовать так, чтобы элементами матриц были вещественные числа. Способ преобразования зависит от формы представления комплексных чисел в цифровом процессоре. Покажем его для распространенного случая, когда вещественные и мнимые части отсчетов сигнала располагаются в последовательных ячейках памяти процессора друг за другом. Этому соответствует такое представление вектора отсчетов в матричной форме:

$$\mathbf{a} = (a_0^{re};\ a_0^{\text{Im}};\ a_1^{re};\ a_1^{\text{Im}};\ \ldots;\ a_{N-1}^{re};\ a_{N-1}^{\text{Im}}).$$

Размерность этого вектора вдвое больше числа отсчетов сигнала. Поэтому вдвое увеличиваются и размеры матриц, составляющих факторизованное представление матрицы $\mathbf{FOUR}_{2n}$, так что формула (5.32) переходит в

$$\mathbf{FOUR}_{2^n} = 2^{-n/2} (\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2) \prod_{r=0}^{n-1} \left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^{r+1}} \oplus \bigoplus_{s=0}^{2^r-1} \mathbf{four}_s^r \right] \right) \times$$
$$\times (\mathbf{I}_{2^{n-1-r}} \otimes \mathbf{h}_2 \otimes \mathbf{I}_{2^{r+1}}), \qquad (5.36a)$$

где

$$\mathbf{four}_s^r = \begin{bmatrix} \cos \varphi_s^r & -\sin \varphi_s^r \\ \sin \varphi_s^r & \cos \varphi_s^r \end{bmatrix}; \quad \varphi_s^r = \frac{2\pi s}{2^{r+1}}. \qquad (5.36б)$$

Нетрудно показать, что такой записи алгоритма соответствует следующее представление матрицы $\mathbf{FOUR}_{2n}$, аналогичное (5.2):

$$\mathbf{FOUR}_{2^n} = (2^{-n/2})\,(\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2)\,([\mathbf{G}_2^2 \otimes 2^{n/2}\,(\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2)\,\mathbf{FOUR}_{2^{n-1}}]\,|\Xi|$$

$$|\Xi|\left[\mathbf{G}_2^3 \otimes 2^{n/2}\,(\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2)\,\mathbf{FOUR}_{2^{n-1}}\left(\bigoplus_{s=0}^{2^{n-1}-1} \mathbf{four}_s^{n-1}\right)\right]\Big).$$

5.4. ОБЗОР БЫСТРЫХ АЛГОРИТМОВ ДРУГИХ ОРТОГОНАЛЬНЫХ ПРЕОБРАЗОВАНИЙ

Пользуясь описанным в § 5.2 матричным аппаратом и представлением матриц ортогональных преобразований как поэтажно-кронекеровских, можно аналогично тому, как это было сделано в § 5.3 для матрицы $\mathbf{FOUR}_{2n}$, получить факторизованное представление матриц и других ортогональных преобразований, описанных в § 5.1. Эти результаты для наиболее известных преобразований представлены ниже.

Преобразование Хаара:

$$\mathbf{HAR}_{2^n} = (2^{-n/2})\left(1 \oplus \bigoplus_{r=0}^{n-1} 2^{r/2}\mathbf{I}_{2^r}\right)\prod_{r=0}^{n-1}\left(\left[\begin{array}{c}\mathbf{I}_{2^r} \otimes \mathbf{G}_2^2 \\ \hline \mathbf{I}_{2^r} \otimes \mathbf{G}_2^3\end{array}\right] \oplus \mathbf{I}_{2^n-2^r+1}\right).$$

Модифицированное преобразование Адамара:

$$\mathbf{MHAD}_{2^n} = (2^{-n/2})\left(1 \oplus \bigoplus_{r=0}^{n-1} 2^{r/2}\mathbf{I}_{2^r}\right)\overline{\mathbf{MHAD}}_{2^n};$$

$$\overline{\mathbf{MHAD}}_{2^n} = \prod_{r=0}^{n-1}[(\mathbf{h}_2 \otimes \mathbf{I}_{2^r}) \oplus \mathbf{I}_{2^n-2^r+1}].$$

Преобразование Уолша – Адамара:

$$\overline{\mathbf{HAD}}_{2^n} = (2^{-n/2})\prod_{r=0}^{n-1}(\mathbf{I}_{2^{n-1-r}} \otimes \mathbf{h}_2 \otimes \mathbf{I}_{2^r}).$$

Преобразование Уолша – Пэли:

$$\mathbf{PAL}_{2^n} = (2^{-n/2})\prod_{r=0}^{n-1}\left(\mathbf{I}_{2^{n-r-1}} \otimes \left[\begin{array}{c}\mathbf{I}_{2^r} \otimes \mathbf{G}_2^2 \\ \hline \mathbf{I}_{2^r} \otimes \mathbf{G}_2^3\end{array}\right]\right).$$

Преобразование Уолша:

$$\mathbf{WAL}_{2^n} = (2^{-n/2})\prod_{r=0}^{n-2}\left(\mathbf{I}_{2^{n-2-r}} \otimes \left(\left[\begin{array}{c}\mathbf{I}_{2^r} \otimes \mathbf{G}_2^2 \\ \hline \overline{\mathbf{I}}_{2^r}(\mathbf{I}_{2^r} \otimes \mathbf{G}_2^3)\end{array}\right] \oplus \right.\right.$$

$$\left.\left. \oplus \overline{\mathbf{I}}_{2^{r+1}}\left[\begin{array}{c}\mathbf{I}_{2^r} \otimes \mathbf{G}_2^2 \\ \hline \overline{\mathbf{I}}_{2^r}(\mathbf{I}_{2^r} \otimes \mathbf{G}_2^3)\end{array}\right]\overline{\mathbf{I}}_{2^{r+1}}\right)\right)\left[\begin{array}{c}\mathbf{I}_{2^{n-1}} \otimes \mathbf{G}_2^2 \\ \hline \overline{\mathbf{I}}_{2^{n-1}}(\mathbf{I}_{2^{n-1}} \otimes \mathbf{G}_2^3)\end{array}\right]$$

Преобразование Адамара – Хаара:

$$\mathbf{HDHR}_{2^n}^m = (2^{-n/2})\left(\mathbf{I}_{2^m} \otimes \left[1 \oplus \bigoplus_{r=0}^{n-m+1} 2^{r/2}\mathbf{I}_{2^r}\right]\right)\prod_{r=0}^{m-1}(\mathbf{I}_{2^{m-r-1}} \otimes \mathbf{h}_2 \otimes$$

$$\otimes \mathbf{I}_{2^{n-m+r}})\prod_{r=m}^{n-1}\left(\mathbf{I}_{2^m} \otimes \left[\left[\begin{array}{c}\mathbf{I}_{2^{r-m}} \otimes \mathbf{G}_2^2 \\ \hline \mathbf{I}_{2^{r-m}} \otimes \mathbf{G}_2^3\end{array}\right] \oplus \mathbf{I}_{2^{n-2^r-m+1}}\right]\right).$$

Матрицы перехода между различными преобразованиями. При цифровой обработке сигналов иногда требуется, зная спектр сигнала по одному базису, найти его представление по другому базису. Для этого достаточно матрицу-столбец коэффициентов умножить на соответствующую матрицу перехода $a^{(2)} = M a^{(1)}$. С помощью представления матриц ортогональных преобразований в виде поэтажко-кронекеровскнх можно достаточно просто найти матрицы перехода между этими преобразованиями. Покажем это на примере связи с матрицей $\mathbf{MHAD}_{2n}$ модифицированного преобразования Адамара матрицы ДПФ $\mathbf{FOUR}_{2n}$.

Согласно (5.2)

$$\overline{\mathbf{FOUR}}_{2^n} = [\mathbf{G}_2^2 \otimes \overline{\mathbf{FOUR}}_{2^{n-1}}] \boxplus \left[\mathbf{G}_2^3 \otimes (\overline{\mathbf{FOUR}}_{2^{n-1}}) \left(\bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}} \right) \right],$$

откуда по теореме 3 § 5.2

$$\overline{\mathbf{FOUR}}_{2^n} = \left[\overline{\mathbf{FOUR}}_{2^{n-1}} \oplus (\overline{\mathbf{FOUR}}_{2^{n-1}}) \left(\bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-2}} \right) \right] (\mathbf{h}_2 \otimes \mathbf{I}_{2^{n-1}}).$$

Используя это выражение как рекуррентную формулу для $\overline{\mathbf{FOUR}_{2^n}}$, можно получить:

$$\overline{\mathbf{FOUR}}_{2^n} = \left(\mathbf{I}_2 \oplus \bigoplus_{r=1}^{n-1} (\overline{\mathbf{FOUR}}_{2^r}) \left(\bigotimes_{s=0}^{r-1} \mathbf{d}_{2^{r-s-2}} \right) \right) \times$$

$$\times \prod_{r=0}^{n-1} [(\mathbf{h}_2 \otimes \mathbf{I}_{2^r}) \oplus \mathbf{I}_{2^{n-2^r+1}}].$$

Но произведение $\prod_{r=0}^{n-1} [(\mathbf{h}_2 \otimes \mathbf{I}_{2^r}) \oplus \mathbf{I}_{2^{n-2^r+1}}]$ в правой части этого выражения является с точностью до диагональной матрицы $(2^{-n/2}) \left[1 \oplus \bigoplus_{r=0}^{n-1} 2^{r/2} \mathbf{I}_{2^r} \right]$ матрицей модифицированного преобразования Адамара. Следовательно, первая матрица является матрицей перехода от $\overline{\mathbf{MHAD}}_{2^n}$ к $\overline{\mathbf{FOUR}}_{2^n}$:

$$\overline{\mathbf{M}}_{2^n}^{\mathrm{MH-F}} = \left[\mathbf{I}_2 \oplus \bigoplus_{r=1}^{n-1} (\overline{\mathbf{FOUR}}_{2^r}) \left(\bigoplus_{s=0}^{r-1} \mathbf{d}_{2^{r-s-2}} \right) \right]. \quad (5.37)$$

Для того чтобы факторизовать эту матрицу, заметим, что из (5.37) вытекает:

$$\overline{\mathbf{M}}_{2^n}^{\mathrm{MH-F}} = \overline{\mathbf{M}}_{2^{n-1}}^{\mathrm{MH-F}} \oplus \overline{\mathbf{FOUR}}_{2^{n-1}} \left(\bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-3}} \right) = \overline{\mathbf{M}}_{2^{n-1}}^{\mathrm{MH-F}} \oplus$$

$$\oplus \overline{\mathbf{M}}_{2^{n-1}}^{\mathrm{MH-F}} \overline{\mathbf{MHAD}}_{2^{n-1}} \left(\bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-3}} \right) = (\mathbf{I}_2 \otimes \mathbf{M}_{2^{n-1}}^{\mathrm{MH-F}}) \times$$

$$\times (\mathbf{I}_{2^{n-1}} \oplus \overline{\mathbf{MHAD}}_{2^{n-1}}) \left(\mathbf{I}_{2^{n-1}} \oplus \bigotimes_{s=0}^{n-2} \mathbf{d}_{2^{n-s-3}} \right).$$

Используя эту формулу рекуррентно, получим

$$\overline{\mathbf{M}}_{2^n}^{\mathrm{MH-F}} = \prod_{r=1}^{n-1} (\mathbf{I}_{2^{n-1-r}} \otimes [\mathbf{I}_{2^r} \oplus \overline{\mathbf{MHAD}}_{2^r}]) \times$$

$$\times \left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^r} \oplus \bigotimes_{s=0}^{r-1} \mathbf{d}_{2^{r-s-2}} \right] \right).$$

Наконец, подставив сюда выражение для $\overline{\mathbf{MHAD}_{2^r}}$, после очевидных преобразований придем окончательно к выражению

$$\mathbf{M}_{2^n}^{\mathrm{MH-F}} = \mathbf{M}_{2^n}^{\text{инв}} \prod_{r=1}^{n-1} \left(\prod_{p=0}^{n-1} (\mathbf{I}_{2^{n-1-r}} \otimes [\mathbf{I}_{2^r} \oplus (\mathbf{h}_2 \otimes \mathbf{I}_{2^p}) \oplus \mathbf{I}_{2^r-2^p+1}]) \right) \times$$

$$\times \left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^r} \oplus \bigotimes_{s=0}^{r-1} \mathbf{d}_{2^{r-2-s}} \right] \right).$$

Обратная матрица перехода будет равна произведению этих же матриц, взятых в обратном порядке и с заменой $\mathbf{d}_s$ на $\mathbf{d}_{-s}$.

Переход от модифицированного преобразования Адамара к ДПФ с двоичной инверсией иллюстрируется графом на рис. 5.2. Для удобства в левой части рисунка показан граф, соответствующий модифицированному преобразованию Адамара [1]. Нетрудно видеть, что граф $\mathbf{MHAD}_{2^n}$ является усеченным графом БПФ. Отсюда, в частности, вытекает, что по количеству операций алгоритмы вычисления ДПФ через БПФ или через модифицированное преобразование Адамара с последующим переходом к ДПФ эквивалентны.

Ввиду аналогии между факторизованными представлениями матриц $\mathbf{HAD}_{2^n}$ и $\mathbf{FOUR}_{2^n}$ матрицы $\mathbf{M}_{2^n}^{\mathrm{MH-F}}$, $\mathbf{M}_{2^n}^{\mathrm{F-MH}}$ можно превратить в матрицы $\mathbf{M}_{2^n}^{\mathrm{MH-H}}$, $\mathbf{M}_{2^n}^{\mathrm{H-MH}}$ перехода от

модифицированного преобразования Аламава к преобразованию Уолша – Адамара, и наоборот, изъяв из $\mathbf{M}_{2^n}^{\text{МН}-\text{F}}$ и $\mathbf{M}_{2^n}^{\text{F}-\text{МН}}$ диагональные матрицы:

$$\mathbf{M}_{2^n}^{\text{МН}-\text{Н}} = \prod_{r=1}^{n-1}\prod_{p=0}^{r-1}(\mathbf{I}_{2^{n-1-r}} \otimes [\mathbf{I}_{2^r} \oplus (\mathbf{h}_2 \otimes \mathbf{I}_{2^p}) \oplus \mathbf{I}_{2^r-2^p+1}]).$$

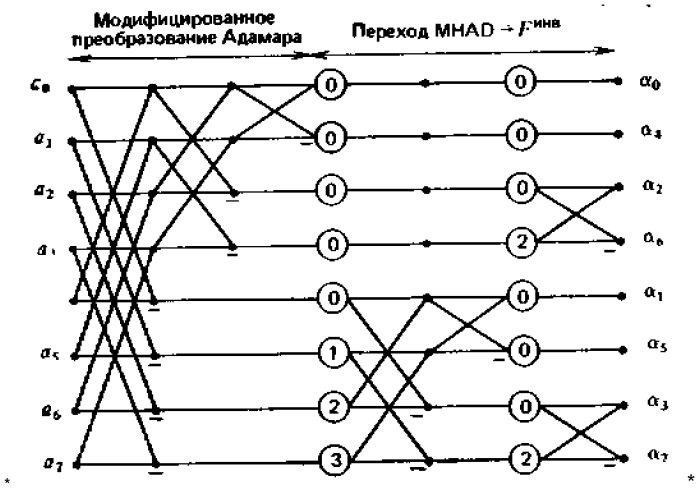


Рис. 5.2. Граф алгоритма перехода от модифицированного преобразования Адамара к ДПФ

Наконец, действуя так же, как для матрицы $M_{2n}^{\text{МН-Н}}$, нетрудно получить матрицу перехода от преобразования Хаара к преобразованию Пэли:

$$\mathbf{M}_{2^n}^{\text{HR}-\text{P}} = \prod_{r=1}^{n-1}\prod_{p=0}^{r-1}\left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^r} \oplus \left[\frac{\mathbf{I}_{2^p} \otimes \mathbf{G}_2^2}{\mathbf{I}_{2^p} \otimes \mathbf{G}_2^3}\right] \oplus \mathbf{I}_{2^r-2^p+1}\right]\right).$$

Матрицы перестановок. Если при выводе факторизованного представления матриц преобразований Уолша – Пэли и Уолша воспользоваться теоремой 3 перестановок (см. § 5.2), то можно получить [46] следующие факторизованные представления матриц перестановок по двоичной инверсии:

$$\mathbf{M}_{2^n}^{\text{инв}} = \prod_{r=0}^{n-1}\left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\frac{\mathbf{I}_{2^r} \otimes \mathbf{G}_2^0}{\mathbf{I}_{2^r} \otimes \mathbf{G}_2^1}\right]\right)$$

и перестановки из кода Грея в прямой двоичный код:

$$\mathbf{M}_{2^n}^{\text{гр/пр}} = \prod_{r=0}^{n-1}(\mathbf{I}_{2^r} \otimes [\mathbf{I}_{2^{n-1-r}} \oplus \overline{\mathbf{I}}_{2^{n-1-r}}]).$$

Обратив последнюю, получим матрицу перестановки из прямого двоичного кода в код Грея:

$$\mathbf{M}_{2^n}^{\text{пр/гр}} = \prod_{r=0}^{n-1}(\mathbf{I}_{2^{n-1-r}} \otimes [\mathbf{I}_{2^r} \oplus \overline{\mathbf{I}}_{2^r}]).$$

Алгоритмы транспонирования больших матриц. Рассмотрим задачу транспонирования больших матриц, хранящихся во внешних запоминающих устройствах ЭВМ (ВЗУ) так, что доступ к ним может осуществляться только по столбцам или строкам. Пусть, для определенности, внешнее запоминающее устройство допускает произвольный доступ к столбцам матрицы. Тогда матрицу $\mathbf{A}_{s,r}$ расположенную в таком запоминающем устройстве, удобно трактовать как вектор-столбец $\mathbf{V}_{s,r}$, элементы которого нумеруются по смешанному основанию (s, r), и связь между $\mathbf{A}_{s,r}$, и $\mathbf{V}_{s,r}$ можно выразить следующей формулой:

$$\mathbf{V}_{s,r} = \prod_{l=0}^{r-1} \mathbf{A}_{s,r}(\mathbf{G}_r^{(l)})^{\mathrm{T}}.$$

Транспонирование матрицы $\mathbf{A}_{s,r}$ означает инверсную перестановку элементов вектора $\mathbf{V}_{s,r}$, по смешанному основанию (r, s). Поэтому матрица транспонирования T(s,r/r,s) может быть записана следующим образом:

$$\mathbf{T}(s, r|r, s) = \sum_{i=0}^{s-1} \sum_{j=0}^{r-1} \mathbf{G}_r^{(j)} \otimes \mathbf{G}_s^{(i)} = \sum_{i=0}^{s-1} \mathbf{I}_r \otimes \mathbf{G}_s^{(i)}. \tag{5.38}$$

В более общем случае, если элементом матрицы является не одно число, а группа чисел, например, если элементы матрицы занимают несколько последовательных машинных слов, как в случае матриц из комплексных чисел или чисел в форме с плавающей точкой, следует считать, что вектор V имеет размерность не sr, а wsr, где w – размер группы (количество машинных слов на элемент матрицы). Тогда матрица транспонирования запишется как

$$\mathbf{T}(sw, r|r, sw) = \sum_{i=0}^{s-1} \sum_{j=0}^{r-1} \mathbf{G}_r^{(j)} \otimes \mathbf{G}_s^{(i)} \otimes \mathbf{I}_w. \tag{5.39}$$

Все алгоритмы транспонирования больших матриц основаны на разбиении исходной матрицы на фрагменты, которые можно вызвать из внешнего ЗУ в запоминающее устройство с произвольным доступом (ОЗУ). Эти фрагменты подвергаются в ОЗУ определенным перестановкам, после чего результаты записываются в выходное ВЗУ. Например, пусть s – составное число: $s=s_1s_2$. Заменим в (5.38) $i=ks_1+l$, $l=0, 1, \ldots, s_1-1$; $k=0, 1, \ldots, s_2-1$

Тогда $\mathbf{G}_s^{(i)} = \mathbf{G}_s^{(k)} \otimes \mathbf{G}_s^{(l)}$, и (5.38) перейдет в

$$\mathbf{T}(s, r|r, s) = \sum_{k=0}^{s_2-1} \sum_{l=0}^{s_1-1} \mathbf{I}_r \otimes \mathbf{G}_{s_2}^{(k)} \otimes \mathbf{G}_{s_1}^{(l)} \tag{5.40}$$

Записав в (5.40) $\mathbf{I}_r = \mathbf{I}_r\mathbf{I}_r$; $\mathbf{G}_{s_2}^{(k)} = \mathbf{I}_1\mathbf{G}_{s_2}^{(k)}$; $\mathbf{G}_{s_1}^{(l)} = \mathbf{G}_{s_1}^{(l)}\mathbf{I}_{s_1}$ и применив к полученному выражению тождество (5.30), придем к формуле

$$\mathbf{T}(s, r|r, s) = \sum_{k=0}^{s_2-1} \sum_{l=0}^{s_1-1} (\mathbf{I}_r \otimes \mathbf{G}_{s_1}^{(l)})(\mathbf{I}_r \otimes \mathbf{G}_{s_2}^{(k)} \otimes \mathbf{I}_{s_1}).$$

Далее, используя последовательно тождества (5.34) и (5.35), получаем

$$\mathbf{T}(s, r|r, s) = \sum_{k=0}^{s_2-1} \left(\sum_{l=0}^{s_1-1} \mathbf{I}_r \otimes \mathbf{G}_{s_1}^{(l)} \right) (\mathbf{I}_r \otimes \mathbf{G}_{s_2}^{(k)} \otimes \mathbf{I}_{s_1}) = \mathbf{T}_1\mathbf{T}_{1N} =$$

$$= \left(\mathbf{I}_{s_2} \otimes \sum_{l=0}^{s_1-1} \mathbf{I}_r \otimes \mathbf{G}_{s_1}^{(l)} \right) \left(\sum_{k=0}^{s_2-1} \mathbf{I}_r \otimes \mathbf{G}_{s_2}^{(k)} \otimes \mathbf{I}_{s_1} \right). \tag{5.41}$$

Последнее выражение соответствует алгоритму блочного транспонирования матриц.

Обсудим теперь наиболее эффективные алгоритмы с использованием разбиений на квадраты или прямоугольники. Сначала будем считать, что размеры матрицы являются составными числами.

Пусть $s = s_1s_2$ и $r=r_1r_2$. Тогда, учитывая матричные тождества (5.29) – (5.35) к (5.28) так же, как и при выводе формулы (5.41), матрицу $\mathbf{T}(s, r|r, s)$ из (5.38) можно факторизовать на два сомножителя:

$$\mathbf{T}(s, r|r, s) = \mathbf{T}_2\mathbf{T}_1 = \left(\sum_{k=0}^{s_2-1} \sum_{l=0}^{s_1-1} \mathbf{I}_{r_2} \otimes \mathbf{G}_{s_1}^{(l)} \otimes \mathbf{G}_{s_2}^{(k)} \otimes \mathbf{I}_{r_1} \right) \times$$

$$\times \left(\mathbf{I}_{r_2} \otimes \sum_{l=0}^{s_1-1} \sum_{j=0}^{s_2-1} \mathbf{I}_{r_1} \otimes \mathbf{G}_{s_2}^{(j)} \otimes \mathbf{G}_{s_1}^{(l)} \right). \tag{5.42}$$

Выражение (5.42) означает, в сущности, разбиение транспонируемой матрицы на подматрицы размером (s_1x r_1,), транспонирование этих подматриц (множитель $\mathbf{T}_1$) и последующее транспонирование матрицы, элементами которой являются подматрицы (множитель $\mathbf{T}_2$).

Факторизация (5.42) имеет достаточную общность, и при соответствующей интерпретации множителей $\mathbf{T}_1$, $\mathbf{T}_2$ и выборе s_1, s_2, r_1, r_2 описывает целый ряд алгоритмов транспонирования. В частности, она соответствует методу разбиения на прямоугольники или на квадраты (если $s_1=r_1$ и $s_2=r_2$) за два просмотра всей матрицы. Число просмотров в этих алгоритмах в основном зависит от количества столбцов матрицы, помещающихся в

ОЗУ одновременно. Для получения трехпроходового алгоритма необходимо факторизовать также матрицу T_2. Пусть $s_2=s_3s_4$, $r_2=r_3r_4$. Тогда множитель $\mathbf{T}_2$ можно представить как

$$\mathbf{T}_2 = \left(\sum_{k=0}^{s_4-1} \sum_{l=0}^{s_1 s_3-1} \mathbf{I}_{r_4} \otimes \mathbf{G}^{(l)}_{s_3 s_1} \otimes \mathbf{G}^{(k)}_{s_4} \otimes \mathbf{I}_{r_3 r_1}\right) \times$$

$$\times \left(\mathbf{I}_{r_4} \otimes \sum_{n=0}^{s_3-1} \sum_{m=0}^{s_1 s_4-1} \mathbf{I}_{r_3} \otimes \mathbf{G}^{(m)}_{s_1 s_4} \otimes \mathbf{G}^{(n)}_{s_3} \otimes \mathbf{I}_{r_1}\right).$$

Видно, что новые множители отличаются от множителей из (5.42) только размерами составляющих матриц. Поэтому особенности алгоритмов будем далее рассматривать на примере двухпроходового алгоритма, поскольку даже на мини-ЭВМ. все практически встречаемые матрицы можно транспонировать за два просмотра ВЗУ.

Матрица T_1, в (5.42) имеет блочно-диагональную структуру. Это значит, что если прочесть в ОЗУ первые п столбцов, все перестановки, определяемые выражением при вертикальной сумме, можно выполнить в этом фрагменте независимо от других столбцов. Следовательно, первый проход можно выполнить путем последовательного считывания фрагментов по r_1 столбцов, перестановок и записи результатов на то же место в ВЗУ. Вторая матрица $\mathbf{T}_2$ в выражении (5.42) не имеет диагональной структуры, и поэтому для получения выходного столбца необходимо в общем случае больше последовательных столбцов, чем может поместиться в ОЗУ. Ее факторизация

$$\mathbf{T}_2 = \mathbf{T}_{\mathrm{OUT}}\mathbf{T}^1_2\mathbf{T}_{\mathrm{IN}} = \left(\sum_{k=0}^{s_2-1} \mathbf{I}_{s_1} \otimes \mathbf{G}^{(k)}_{s_2} \otimes \mathbf{I}_{r_2 r_1}\right) \times$$

$$\times \left(\mathbf{I}_{s_1} \otimes \sum_{l=0}^{s_2-1} \mathbf{I}_{r_2} \otimes \mathbf{G}^{(l)}_{s_2} \otimes \mathbf{I}_{r_1}\right) \left(\sum_{l=0}^{s_1-1} \mathbf{I}_{r_2} \otimes \mathbf{G}^{(l)}_{s_1} \otimes \mathbf{I}_{s_2 r_1}\right) \qquad (5.43)$$

объясняет основную идею алгоритмов разбиения на прямоугольники или квадраты. Средняя матрица $\mathbf{T}^1_2$ является уже блочно-диагональной и, как и матрица $\mathbf{T}_1$, определяет перестановки в ОЗУ. Матрицы же $\mathbf{T}_{\mathrm{IN}}$,, $\mathbf{T}_{\mathrm{OUT}}$ определяют операции считывания и записи на ВЗУ соответственно. В таком представлении алгоритма транспонирования легко подсчитать необходимые затраты оперативной памяти ЭВМ и затраты времени на обмен данными в ВЗУ.

Необходимая емкость ОЗУ определяется размерами блоков матриц $\mathbf{T}_1$, $\mathbf{T}_2$ и равна $\max\{rs/r_2, rs/s_1\}$ Практически эта величина достаточна лишь для метода разбиения на квадраты, так как в этом случае матрицы $\mathbf{T}_1$ $\mathbf{T}_2$ симметричны, и перестановки в ОЗУ можно выполнять с замещением (с оставлением на месте). В случае разбиения на прямоугольники нужен дополнительный выходной буфер или необходимо усложнение алгоритма перестановок в ОЗУ, чтобы транспонировать прямоугольные субматрицы с оставлением на месте. Последнее нежелательно из-за усложнения программирования и увеличения времени на перестановки в памяти.

Отметим распространенный случай, когда $s_1=dg$ и $g>=r_2$, т.е. $(rs/r_2)>=d(rs/s_1)$. В этом случае при втором проходе используется только часть доступного ОЗУ, и можно применить более общую факторизацию матрицы $\mathbf{T}_2$:

$$\mathbf{T}_2 = \mathbf{T}_{\mathrm{OUT}}\mathbf{T}^2_2\mathbf{T}_{\mathrm{IN}} = \left(\sum_{k=0}^{s_2-1} \mathbf{I}_g \otimes \mathbf{G}^{(k)}_{s_2} \otimes \mathbf{I}_{rd}\right)\left(\mathbf{I}_g \otimes \sum_{j=0}^{s_2-1} \sum_{m=0}^{d-1} \mathbf{I}_{r_2} \otimes \mathbf{G}^{(m)}_{d} \otimes\right.$$

$$\left. \otimes \mathbf{G}^{j}_{s_2} \otimes \mathbf{I}_{r_1}\right)\left(\sum_{r=0}^{g-1} \mathbf{I}_{r_2} \otimes \mathbf{G}^{(n)}_{g} \otimes \mathbf{I}_{s_2 d r_1}\right). \qquad (5.44)$$

Из сопоставления выражений (5.42) и (5.44) видно, что блочно-диагональная матрица $\mathbf{T}^2_2$ отличается по структуре от $\mathbf{T}_1$ только наличием кронекеровского сомножителя $\mathbf{I}_{r1}$, который означает, что при транспонировании в ОЗУ в соответствии с (5.44) нужно переставлять не одно число, как в (5.42), а группу из r_1 последовательных чисел. Следовательно, матрицу $\mathbf{T}_1$ можно рассматривать как частный случай матрицы $\mathbf{T}^2_2$ и, значит, перестановки $\mathbf{T}_1$ и $\mathbf{T}^2_2$ в ОЗУ можно выполнять одной подпрограммой, меняя только ее параметры. Далее матрицу $\mathbf{T}^1_2$ из выражения (5.43) можно в том же смысле рассматривать как частный случай матрицы $\mathbf{T}^2_2$ и значит перестановку по $\mathbf{T}^1_2$ можно выполнить той же подпрограммой.

Скорость работы алгоритма транспонирования в основном определяют операции ввода-вывода на ВЗУ, так как перестановка данных в ОЗУ занимает гораздо меньше времени или может быть совмещена с выводом на ВЗУ. Пользуясь выражениями (5.42), (5.43), можно рассчитать затраты на обмен с внешним устройством. Выше отмечалось, что в первом проходе последовательно считываются из ВЗУ и после обработки записываются во ВЗУ фрагменты величиной $r_1 s$. Для этого прохода мы не выделяем отдельно матриц ввода и вывода, так как на первом проходе это единичные матрицы. Максимальная скорость обмена в ВЗУ может быть достигнута при последовательном считывании или записи. Следовательно, время, необходимое для первого прохода, приблизительно равно времени на считывание и запись транспонируемой матрицы с максимальной скоростью. Количество операций ввода-вывода для второго прохода определяют матрицы T_{1N}, T_{OUT}, которые представляют собой матрицы транспонирования [ср. (5.39)].

Рассмотрим подробнее алгоритм считывания и записи на ВЗУ на примере выражения (5.43). При операциях ввода (матрица $\mathbf{T}_{IN}$) транспонируемую матрицу $\mathbf{A}_{s,r}$ надо интерпретировать как матрицу $\mathbf{A}^{IN}{}_{s1,r}$ размером (s_1 x s_2), состоящую из элементов, занимающих в ОЗУ $s_1 r$, последовательных ячеек. Напомним, что в ВЗУ эта матрица располагается в лексикографическом порядке по столбцам. Для выполнения перестановок, определяемых вертикальной суммой в матрице $\mathbf{T}^1{}_2$. необходимо прочитать в ОЗУ rs_2 слов. Согласно виду матрицы $\mathbf{T}^{IN}$ для этого надо прочитать первую строку матрицы $\mathbf{A}^{IN}{}_{s1,r2}$, т.е. прочитать r_2 фрагментов длиной $r_1 s_2$, отстоящих друг от друга на s_1 таких же фрагментов. Выходная матрица $\mathbf{A}_{r,s}$ интерпретируется как матрица $\mathbf{A}^{OUT}{}_{s1,s2}$ размерами s_1 x s_2 и длиной элемента $r_1 r_2$, слов. Первой, аналогично вводу, записывается первая строка матрицы $\mathbf{A}^{OUT}{}_{s1,s2.}$ Следующей считывается вторая строка матрицы $\mathbf{A}^{IN}{}_{s1,r2}$ и после перестановок в ОЗУ записывается во вторую строку матрицы $\mathbf{A}^{OUT}{}_{s1,s2.}$ Всего для второго прохода согласно выражению (5.43) необходимо выполнить $r_2 s_1$ операций ввода с длиной записи $r_1 s_2$ и $s_1 s_2 = s$ операций вывода с длиной записи $r_2 r_1 = r$. Если при транспонировании используется больше чем два просмотра всей матрицы, структура матриц ввода или вывода только слегка изменяется (они могут быть представлены как блочно-диагональные), что не усложняет их интерпретации.

Отметим некоторую произвольность в интерпретации множителей матрицы транспонирования в (растеризованном виде. Так, в выражении (5.42) $\mathbf{T}_2$ можно рассматривать как матрицу вывода, и тогда оно описывает один из способов блочного транспонирования матриц. В формулах (5.43), (5.44) матрицу T_{IN} можно считать матрицей вывода для первого прохода. Тогда считывание для обоих проходов выполняется последовательно, а запись скачками. Такая реализация алгоритма транспонирования выгодна, когда выходные записи формируются в дополнительном буфере небольшого объема, и операцию вывода приходится выполнять небольшими фрагментами. Прочесть же из ВЗУ необходимое количество данных можно за одно обращение к ВЗУ. Это не только ускоряет алгоритм транспонирования, но и унифицирует выполнение разных проходов.

Все результаты, полученные для (5.38), можно легко перенести на транспонирование с перестановкой групп чисел по (5.39). Действительно, кронекеровский сомножитель $\mathbf{I}_w$ в (5.39) можно вынести за знак вертикальной суммы, факторизовать оставшуюся вертикальную сумму, совпадающую с (5.38), как было показано выше, и затем к каждому полученному сомножителю добавить кронекеровсква сомножитель $\mathbf{I}_w$, пользуясь свойствами кронекеровского и обычного произведения матрац и тривиальным тождеством $\mathbf{I}_w = \mathbf{I}_w \mathbf{I}_w \mathbf{I}_w$.

Из сопоставления матриц различных- преобразований в их поэтажно-кронекеровском (§ 5.1) и факторизованном представлении можно сделать следующие выводы о структуре ортогональных матриц преобразований, допускающих факторизацию на слабо заполненные матрицы, и их быстрых алгоритмов.

1. Возможно построение процессоров с универсальной структурой, осуществляющий произвольное ортогональное преобразование, простым изменением блоков, генерирующих элементарные матрицы, на базе которых построено данное преобразование.
2. Матрицы-сомножители, составляющие факторизованное представление матриц ортогональных преобразований, также являются ортогональными. Изменяя эти матрицы или комбинируя их в произвольном порядке, а также в сочетании с матрицами перестановок,

можно получать новые виды преобразований. Это еще один путь построения новых преобразований, гарантированно имеющих быстрые алгоритмы, в дополнение к тем, которые были указаны в § 5.1.

5.5. КВАНТОВАННОЕ ДИСКРЕТНОЕ ПРЕОБРАЗОВАНИЕ ФУРЬЕ И БЫСТРЫЙ АЛГОРИТМ

Метод построения ортогональных преобразований с гарантированным быстрым алгоритмом выполнения, заключающийся в модификации матриц-сомножителей в матричном представлении быстрых алгоритмов, можно использовать для построения преобразования, аппроксимирующего дискретное преобразование Фурье, но выполняемого без операций умножения и с уменьшенным числом сложений.

Рассмотрим алгоритм БПФ с двоичной инверсией результата преобразования [см. (5.29), (5.30)]. Заменим в (5.30) функции $\cos\varphi_s^r$ и $\sin\varphi_s^r$ функциями $\mathrm{qos}\,\varphi_s^r$ и $\mathrm{qin}\,\varphi_s^r$, значения которых получаются квантованием значений функций $\cos\varphi_s^r$ и $\sin\varphi_s^r$ соответственной Например, при трех уровнях квантования $\cos\varphi_s^r$ и $\mathrm{qin}\,\varphi_s^r$ принимают значения –1, О, 1, при пяти –1, –1/2, 0, 1/2, 1. Это соответствует замене матрицы $\mathbf{four}_s^r$ (5.366) в формуле (5.36а) матрицей

$$\mathbf{qour}_s^r = \begin{bmatrix} \mathrm{qos}\,\varphi_s^r & -\mathrm{qin}\,\varphi_s^r \\ \mathrm{qin}\,\varphi_s^r & \mathrm{qos}\,\varphi_s^r \end{bmatrix}. \tag{5.45}$$

В результате получаем следующую матрицу преобразования:

$$\mathbf{QOUR}_{2^n} = \mathbf{D}_q\,(\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2) \prod_{r=0}^{n-1} \left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^r+1} \oplus \bigoplus_{s=0}^{2^r-1} \mathbf{qour}_s^r \right] \right) \times$$
$$\times\,(\mathbf{I}_{2^{n-1-r}} \otimes \mathbf{h}_2 \otimes \mathbf{I}_{2^r+1}), \tag{5.46}$$

представляющую новое преобразование сразу в факторизованной форме. Здесь $\mathbf{D}_q$– диагональная нормирующая матрица, необходимая для коррекции нормировки, которая может быть нарушена в результате квантования, так как в отличие от суммы $\sin^2\varphi+\cos^2\varphi$ умма $\mathrm{qin}^2\varphi+\mathrm{qos}^2\varphi$ не обязательно равна единице. Назовем это преобразование квантованным ДПФ (КДПФ).

Нетрудно видеть, что это преобразование является ортогональным, так как матрицы (5.45), а значит, и все матрицы-сомножители в формуле (5.45) ортогональны.

При $\mathrm{qos}\,\varphi_s^r = 1;\ \ \mathrm{qin}\,\varphi_s^r = 0$ КДПФ переходит в преобразование Адамара в упорядочении Пэли. При трех уровнях квантования (–1, О, 1) для выполнения КДПФ не требуются операции умножения и выпадает часть операций сложения-вычитания. При пяти уровнях квантования (1, –1/2, 0, 1/2, 1) умножения заменяются сдвигом. Поэтому по сложности вычислений КДПФ находится

между быстрыми преобразованиями Хаара и Уолша – Адамара, т.е. требует намного меньше операций, чем БПФ.

Чтобы реализовать этот выигрыш, следует объединить произведение пар матриц в формуле (5.40). Воспользовавшись свойством (5.30), из (5.46) получим :

$$\mathbf{QOUR}_{2^n} = \mathbf{D}_q\,(\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2) \left(\prod_{r=0}^{n-1} \left(\mathbf{I}_{2^{n-1-r}} \otimes \left[\mathbf{I}_{2^r+1} \oplus \bigoplus_{s=0}^{2^r-1} \mathbf{qour}_s^r \right] \times \right.\right.$$
$$\left.\left. \times\,(\mathbf{h}_2 \otimes \mathbf{I}_{2^r+1}) \right) \right) = \mathbf{D}_q\,(\mathbf{M}_{2^n}^{\text{инв}} \otimes \mathbf{I}_2) \prod_{r=0}^{n-1} \left(\mathbf{I}_{2^{n-1-r}} \otimes \right.$$
$$\left. \otimes \left[\mathbf{I}_{2^r+1} \oplus \bigoplus_{s=0}^{2^r-1} \mathbf{qour}_s^r \right] \left[\bigoplus_{l=0}^{1} [1\ (-1)^l] \otimes \mathbf{I}_{2^r+1} \right] \right).$$

Применив теперь теорему 1 (§5.2), получим окончательно:

$$\mathrm{QOUR}_{2^n} = \mathrm{D}_q\,(\mathrm{M}^{\mathrm{inne}}_{2^n} \otimes \mathrm{I}_2) \prod_{r=0}^{n-1} \left(\mathrm{I}_{2^{n-1-r}} \otimes \left[(\mathrm{G}_2^2 \otimes \mathrm{I}_{2^{r+1}}) \boxplus \right.\right.$$
$$\left.\left. \boxplus \left(\mathrm{G}_2^3 \otimes \bigoplus_{s=0}^{2^r-1} \mathrm{qour}_s^r \right) \right] \right).$$

В силу своего происхождения КДПФ может служить удовлетворительной аппроксимацией ДПФ. Ошибку аппроксимации базисными функциями КДПФ экспоненциальных базисных функций ДПФ можно минимизировать оптимальным расположением моментов квантования значений синусов и косинусов и надлежащим выбором весовых поправок, учитывающих, что, вообще говоря, $\mathrm{qin}^2 \varphi_s^r + \mathrm{qos}^2 \varphi_s^r \neq 1$

Как показывают эксперименты по моделированию КДПФ[1], ошибка аппроксимации строк матрицы ДПФ строками матрицы КДПФ растет с ростом N, однако сравнительно медленно (см. табл. 5.1, где приведены минимальные среднеквадратические значения относительной ошибки ε при трех уровнях квантования для различных значений размерности матрицы N).

Таблица 5.1

ε	0,235	0,3	0,33	0,37	0,4
N	L6	32	64	128	256

Оптимальное правило квантования также слабо зависит от N, и для трех уровней квантования может быть приближенно описано следующим соотношением:

$$\mathrm{qos}\, 2\pi \frac{r}{N} = \begin{cases} 1, & 0 \leqslant r \leqslant N/8;\ \ 7N/8 \leqslant r < 1; \\ -1, & 3N/8 \leqslant r \leqslant 5N/8; \\ 0, & \text{для остальных } r; \end{cases}$$

$$\mathrm{qin}\, 2\pi \frac{r}{N} = \begin{cases} 1, & N/16 \leqslant r \leqslant 7N/16; \\ -1, & 9N/16 \leqslant r \leqslant 15N/16; \\ 0, & \text{для остальных } r. \end{cases}$$

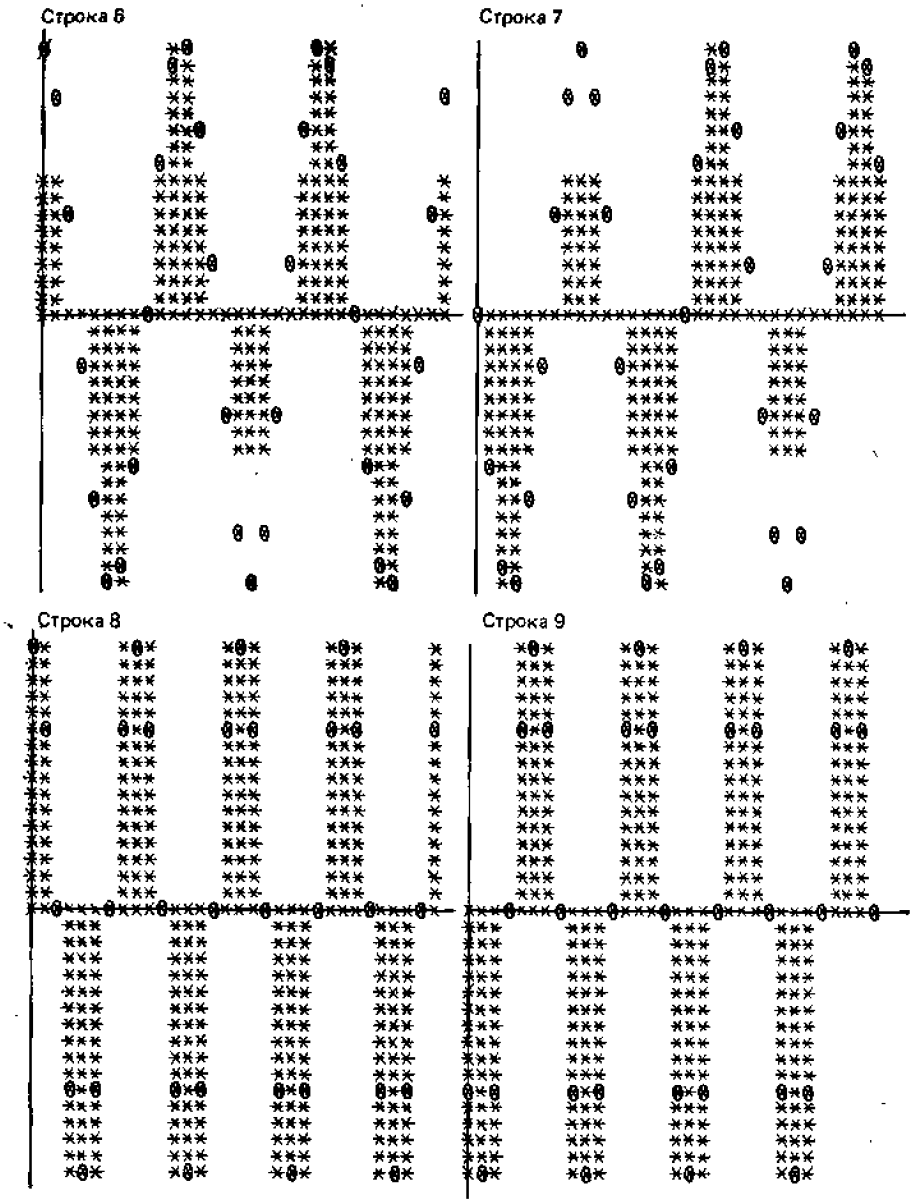


Рис. 5.3. Графики базисных функций преобразования $QOUR_2^n$ при $n=5$

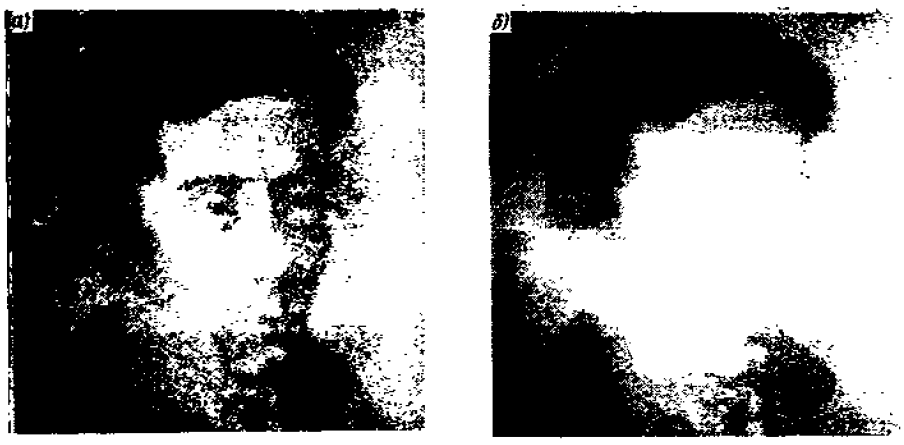


Рис 5.4. Изображения, восстановленные оптически с голограмм, рассчитанных с использованием КДПФ вместо ДПФ:
а — КДПФ с тремя уровнями квантования; *б* — КДПФ с пятью уровнями квантования

На рис. 5.3 показаны графики функций, описывающих строки матрицы $QOUR_{25}$ при оптимально подобранных моментах квантования. На том же рисунке крестиками в кружках показаны значения соответственно косинусондальных и синусоидальных функций для ДПФ.
Возможность использовать КДПФ в качестве быстро вычислимой аппроксимации ДПФ подтверждают также эксперименты с синтезом голограмм [49]. Они показывают, что квантование, присущее КДПФ, не приводит к разрушению изображения, восстанавливаемого в схеме Фурье с синтезированных с помощью КДПФ голограмм, хотя определенным образом и сказывается на его качестве, проявляясь в виде дополнительного шума. На рис. 5.4, а, б показаны восстановленные изображения при трех и пяти уровнях квантования соответственно.

Глава 6

ЦИФРОВОЕ СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ . И ИЗМЕРЕНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК

6.1. СТАТИСТИЧЕСКИЕ МОДЕЛИ СЛУЧАЙНЫХ ИЗОБРАЖЕНИЙ И ВОЛНОВЫХ ПОЛЕЙ

Для моделирования оптических сигналов и систем широко применяется цифровое моделирование как удобный и гибкий инструмент исследования и синтеза оптимальных систем [2, 49]. При моделировании оптических систем изучается, как правило, результат преобразования исследуемой оптической системой тех или иных оптических сигналов. В зависимости от того, какие характеристики систем и сигналов изучаются при моделировании, используют детерминированное и статистическое моделирование и соответственно детерминированные и статистические цифровые модели.

При детерминированном моделировании результат каждого модельного эксперимента рассматривается отдельно, как самостоятельный. При статистическом моделировании результат моделирования получается путем того или иного усреднения характеристик и данных, получаемых в наборе отдельных экспериментов. Поэтому статистические цифровые модели отличаются от детерминированных наличием датчиков случайных сигналов и программных блоков измерения статистических характеристик. В данном параграфе рассматриваются методы получения случайных изображений и объектов для статистического моделирования изображающих и других оптических и голографических систем.

Случайные изображения и объекты при статистическом моделировании генерируются в виде реализаций псевдослучайных двумерных последовательностей с заданными статистическими свойствами. Термин «псевдослучайный» означает, с одной стороны, что последовательности заданы только в статистическом смысле,

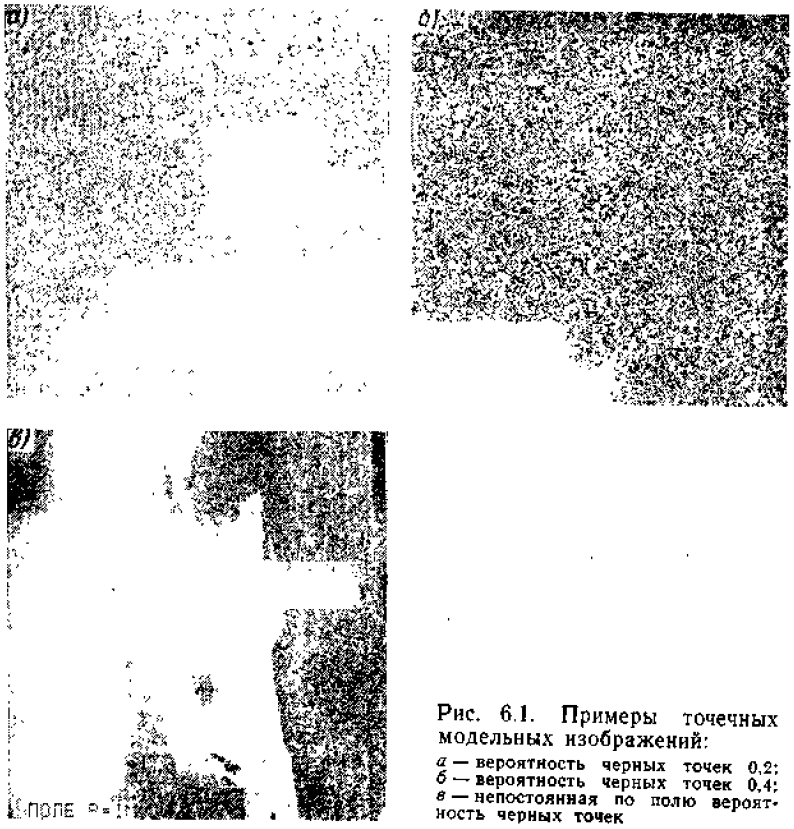


Рис. 6.1. Примеры точечных модельных изображений:
а — вероятность черных точек 0,2;
б — вероятность черных точек 0,4;
в — непостоянная по полю вероятность черных точек

т.е. заданы только средние по всем реализациям последовательности значения тех или иных (в зависимости от решаемой задачи моделирования) ее параметров или характеристик, а не сами последовательности. С другой стороны, эти последовательности не являются «истинно» случайными, так как при цифровом моделировании любая их реализация может быть воспроизведена и притом абсолютно точно.

В цифровом моделировании для описания изображений и волновых полей как случайных объектов обычно используют такие введенные в теории вероятностей и случайных процессов (см., например, [8]) характеристики, как функции распределения вероятностей, корреляционные функции, энергетические спектры, а также различные статистические модели случайных процессов и полей. Методы измерения статистических характеристик при цифровой обработке и методы генерирования псевдослучайных последовательностей с заданными характеристиками будут рассмотрены в § 6.2 и 6.3. Здесь же будут описаны несколько видов статистических моделей, используемых для цифрового статистического моделирования в цифровой оптике.

Статистические модели изображений по типу моделируемых изображений можно разделить на текстурные и детальные. Текстурные модели могут быть точечными, фигурными или площадными.

Точечные модели представляют собой изображения случайно расположенных на равномерном фоне точек (отсчетов) с заданными или случайными значениями сигнала (рис. 6.1). При моделировании пространственно-однородных полей густота точек (среднее число точек на единицу площади) постоянна по площади изображения. Она может также меняться по площади, как, например, на рис. 6.1,в, если модель пространственно-неоднородна.

Фигурные модели строятся на основе представлений о формировании изображения из случайно разбросанных фигур. При этом случайными являются координаты центров фигур, их размеры, ориентация, а параметрами модели – вероятность попадания ентра в данный элемент изображения, законы распределения размеров фигур и их наклона к выбранному направлению на растре. Простейшие фигурные модели – линейчатые, которые формируются из случайно разбросанных по площади

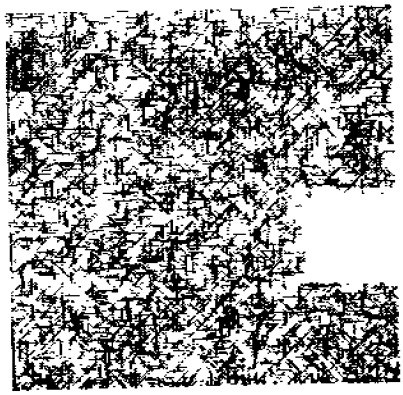


Рис. 6.2. Линейчатая модель изображений с вероятностью попадания центра отрезка в данный элемент изображения 0,3, равномерным распределением длины в диапазоне 0–25 элементов и равномерным распределением ориентации по четырем направлениям.

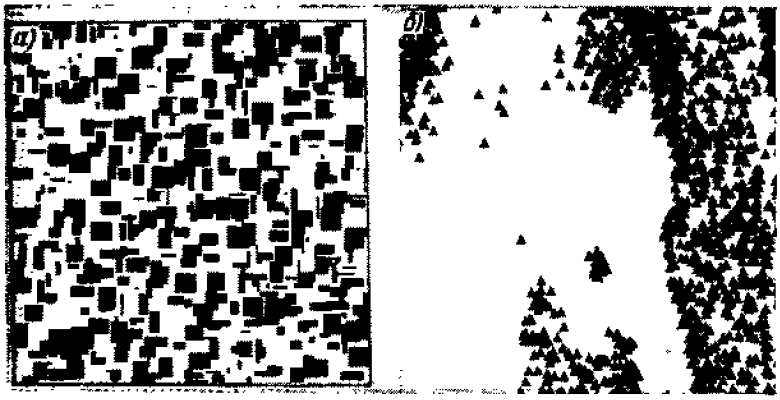


Рис. 6.3. Фигурные модельные изображения:
а – случайно разбросанные прямоугольники с вероятностью попадания центра прямоугольника в данный элемент изображения 0,01 и равномерным распределением ширины и высоты в диапазоне 0–25 элементов (значения ширины и высоты независимы); б–пространственно-неоднородное изображение случайно разбросанных треугольников постоянной формы и размеров

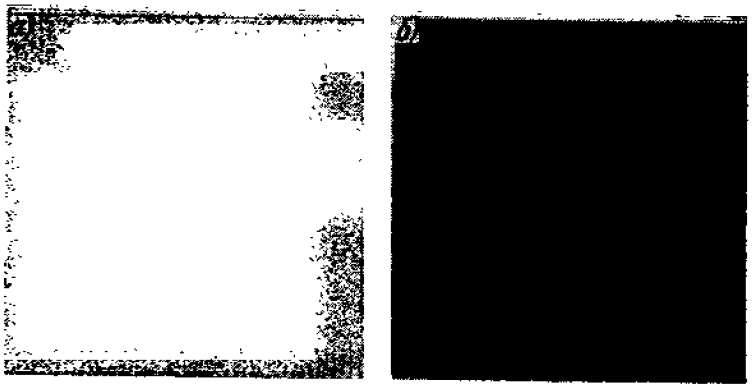


Рис. 6.4. Модельные изображения со статистически независимыми отсчетами: а – равномерным распределением значений от черного до белого; б –с усеченным гаус-совским распределением и тени же первым и вторым моментами, что и на рис. а.

изображений отрезков линий. Примеры линейчатых изображений показаны на рис. 6.2, примеры более сложных фигурных моделей – на рис. 6.3.

Простейшие площадные модели – это модели со статистическими независимыми отсчетами. Они описываются только распределениями вероятностей значений элементов изображения (рис. 6.4).

Площадные модели с коррелированными отсчетами можно разделить на одномерные и двумерные (рис. 6.5). В одномерных отсчеты коррелированы только в одном направлении. Эти модели, как правило, используются при моделировании систем передачи изображений, содержащих этап развертки изображений для преобразования их в одномерный видеосигнал.

В двумерных моделях отсчеты коррелированы в двух направлениях. Для статистического описания таких моделей используются, как правило, представления о гауссовских и марковских

случайных процессах и полях [29]. Марковские модели естественным образом строятся для одномерно коррелированных изображений. Известно также несколько вариантов обобщений марковских моделей на двумерный случай [64]. Как и все другие модели, описанные выше, площадные модели могут быть пространственно однородными и неоднородными.

Наконец, существует важный класс моделей случайных текстурных изображений и полей, которые можно было бы назвать функциональными (рис. 6.6). Они строятся как функциональные {обычно поэлементные нелинейные) преобразования более простых площадных моделей – моделей с независимыми и коррелированными отсчетами. Функциональные модели применяются для имитации изображений и полей, получаемых в физических экспериментах {например, как модели интерферограмм) [33, 34], или

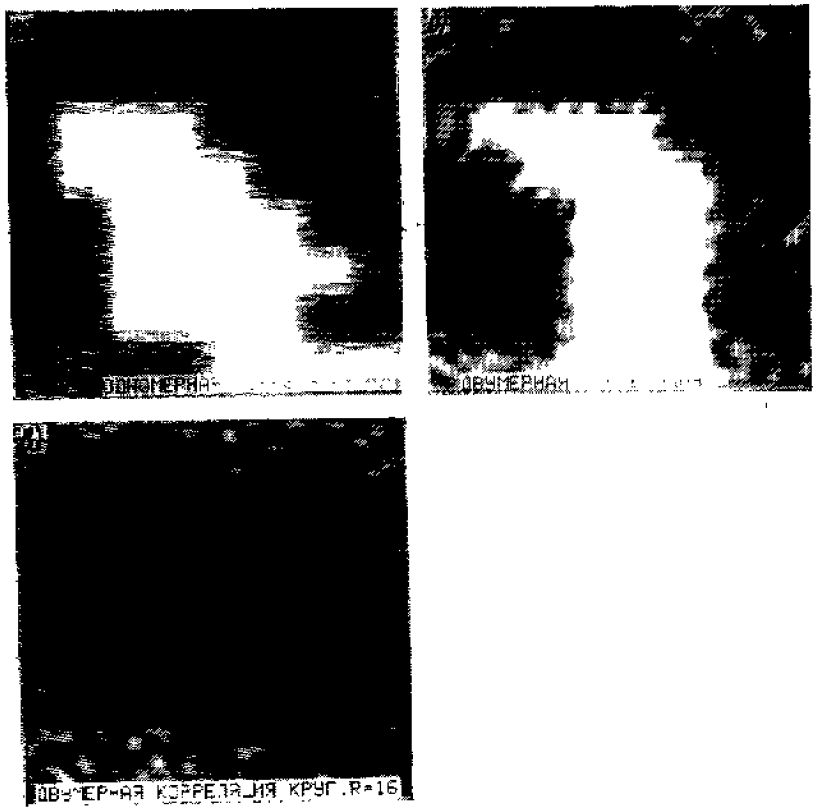


Рис. 6.5. Примеры модельных изображений с коррелированными отсчетами – реализаций гауссовских случайных полей:
а – одномерно коррелированное изображение с функцией корреляции; б – двумерно коррелированное изображение с разделимой функцией корреляции; в – двумерно коррелированное изображение с изотопно ограниченным энергетическим спектром, постоянным в пределах кругарадиусом в 16 элементов в плоскости пространственных частот

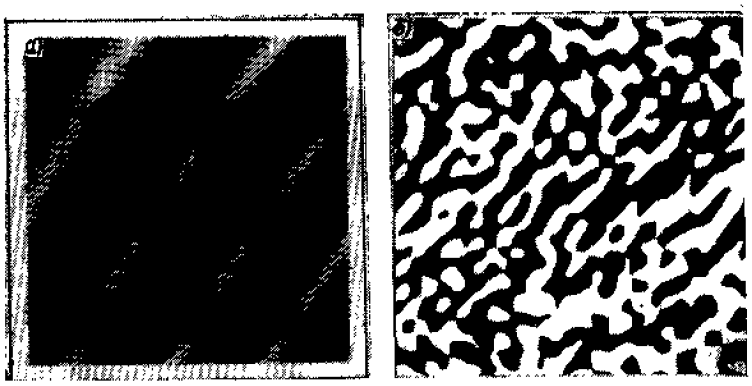


Рис. 6.6. Примеры изображений, полученных с помощью функциональных моделей:
а – модель интерферограммы, полученная путем преобразования коррелированного псевдослучайного поля a (k,l) по формуле $b(k,\ l) = \frac{1}{2} + \cos 2\pi\omega[(a(k,\ l) - a_{min})/(a_{min} - a_{max})]$, где ω – крутизна преобразования, определяющая пространственную частоту интерферограммы; б – фигурная модель, полученная пороговым преобразованием коррелированного псевдослучайного поля

для получения фигурных моделей из случайных фигур. В последнем случае для преобразования модельных изображений с коррелированными отсчетами используются пороговые функции. Детальные модели изображений–это модели, в которых изображения строятся как то или иное сочетание случайного фона и случайных деталей. Для создания случайного фона можно использовать площадные модели с коррелированными отсчетами, для создания случайных деталей в изображении – фигурные модели. Составное модельное изображение получается или как аддитивная смесь фона и деталей или путем «врезки» деталей в фоновое изображение, например, по формуле:

$$b(k, l) = \begin{cases} a_\Phi(k, l) \text{ при } a_\text{д}(k, l) < h; \\ a_\text{д}(k, l) \text{ при } a_\text{д}(k, l) \geq h, \end{cases} \quad (6.1)$$

где $a_\Phi(k, l)$ –модельное изображение для фоновой части;
$a_\text{д}(k, l)$ – модельное изображение для детальной компоненты; h – некоторое пороговое значение $a_\text{д}(k, l)$.
Иногда для создания детальных модельных изображений в качестве фоновой и/или детальной компоненты используют образцы натуральных изображений из заданного класса (аэрофотоснимки, изображения текста, фотографии объектов и т.д.).

6.2. ГЕНЕРИРОВАНИЕ ПСЕВДОСЛУЧАЙНЫХ ЧИСЕЛ ЗАДАННЫМИ СТАТИСТИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

Стандартный способ получения псевдослучайных последовательностей с заданными статистическими свойствами состоит в том, что сначала с помощью достаточно простых алгоритмов генерируют независимые псевдослучайные числа с равномерным распределением, а затем их подвергают линейным и нелинейным преобразованиям для получения заданных статистических свойств [24].

Чаще всего при генерировании независимых и равномерно распределенных псевдослучайных чисел используются рекуррентные алгоритмы типа [24]

$$\xi_k = (c_1 \xi_{k-1} + c_2) \bmod c_3, \quad (6.2)$$

где ξ_k – k-е псевдослучайное число в создаваемой последовательности; ξ_{k-1} – предыдущее число; c_1, c_2 и c_3 – некоторые константы. Начальное число ξ_0 обычно мало влияет на качество получаемой последовательности. Константа c_3 определяется длиной разрядной сетки применяемого цифрового процессора. От нее зависит период повторения последовательности, поэтому ее желательно выбирать максимально возможной. Константа c_2 незначительно влияет на качество последовательности и даже может выбираться равной нулю. Наиболее критичен выбор константы c_1 (см. [24]).

Для того чтобы проверить, насколько статистические свойства получающихся псевдослучайных чисел удовлетворяют заданным требованиям (например, можно ли считать закон их распределения равномерным или можно считать их независимыми или некоррелированными), применяют известные в статистике критерии. В задачах обработки изображений и моделирования систем преобразования изображений наиболее серьезные требования предъявляются к пространственной корреляции используемых псевдослучайных двумерных последовательностей. Для проверки независимости получаемых чисел весьма удобно воспользоваться свойством зрения обнаруживать на изображении регулярные структуры. Для этого значения элементов последовательности передают как яркости элементов изображения, и таким образом превращают поле псевдослучайных чисел с помощью фоторегистраторов или дисплеев в изображение. Если при рассматривании такого изображения на нем не обнаруживаются заметные структуры, псевдослучайные числа можно считать независимыми.

Наиболее часто при цифровом моделировании систем преобразования изображений и полей, а также при синтезе изображений и полей приходится генерировать псевдослучайные числа с гауссовским распределением вероятностей и с заданной ковариационной функцией.

Для того чтобы из независимых псевдослучайных чисел с равномерным распределением получить гауссовские числа, проще всего воспользоваться центральной предельной теоремой теории вероятностей, в соответствии с которой сумма достаточно большого количества независимых случайных величин имеет распределение, приближающееся (с ростом количества складываемых чисел) к гауссовскому. На этом основано большинство алгоритмов генерирования гауссовских чисел. Если одновременно требуется обеспечить заданную ковариационную функцию чисел, то суммирование можно производить с весом

$$\eta_k = \sum_n h_n \xi_{k-n}, \tag{6.3}$$

подбирая весовые коэффициенты $\{h_n\}$ для исходных чисел $\{\xi_k\}$ так, чтобы автосвертка последовательности $\{h_n\}$

$$R_k = \sum_n h_n h_{k-n}$$

дала отсчеты $\{R_k\}$ требуемой ковариационной функции.

При необходимости генерировать некоррелированные гауссовские числа вместо скользящего суммирования (6.3) приходится суммировать неперекрывающиеся группы исходных псевдослучайных чисел.

Если количество весовых коэффициентов в (6.3) невелико, то закон распределения чисел $\{\eta_k\}$ может оказаться недостаточно близким к гауссовскому. Если оно велико, то вычисление суммы (6.3) будет требовать больших затрат машинного времени. Поэтому во всех случаях для получения псевдослучайных чисел с законом распределения, возможно более точно соответствующим гауссовскому, целесообразно осуществлять нормализующее линейное преобразование с помощью дискретного преобразования Фурье, реализуемого алгоритмами БПФ [46]. Рассмотрим этот способ подробнее.

Пусть $\{\xi_k^{re}\}$, $\{\xi_k^{im}\}$ – два отрезка последовательности одинаково распределенных некоррелированных чисел, k = 0, 1, ..., N–1. Образуем из этих чисел последовательность комплексных чисел $\{\xi_k^{re} + i\,\xi_k^{im}\}$ и умножим каждое число этой последовательности на некоторый коэффициент h_k. Тогда в результате дискретного преобразования Фурье этой модифицированной последовательностиполучим комплексные числа

$$\eta_l = \eta_l^{re} + i\eta_l^{im} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} h_k (\xi_k^{re} + i\xi_k^{im}) \exp(i\,2\pi kl/N), \tag{6.4}$$

вещественная и мнимая части которых в силу центральной предельной теоремы имеют распределение, близкое к гауссовскому. Действительно, можно показать, что многомерная характеристическая функция, вычисленная для чисел $\{\eta_l^{re}\}$ и $\{\eta_l^{im}\}$, сходится к характеристической функции гауссовского распределения как $0(1/\sqrt{N})$. Таким образом, при больших N степень приближения закона распределения получаемых чисел к гауссовскому может быть очень высокой.

Найдем ковариационные функции полученных последовательностей. Для ковариационной функции вещественной части $\{\eta_l\}$ имеем

$$C_\eta^{re}(l_1, l_2) = \left\langle \left(\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} h_k (\xi_k^{re} \cos(2\pi kl_1/N) - \right.\right.$$
$$\left. - \xi_k^{im} \sin(2\pi kl_1/N) \right) \left(\frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} h_n (\xi_n^{re} \cos(2\pi nl_2/N) - \right.$$
$$\left.\left. - \xi_n^{im} \sin(2\pi nl_2/N) \right) \right\rangle = (\langle \xi^2 \rangle / N) \sum_{k=0}^{N-1} h_k^2 \cos[2\pi k(l_1 - l_2)/N],$$

где $\langle \xi^2 \rangle = \langle (\xi^{re})^2 \rangle = \langle (\xi^{im})^2 \rangle$ – второй момент псевдослучайных чисел исходной последовательности, а $\langle \xi_{k_1}^{re} \xi_{k_2}^{re} \rangle = \langle \xi_{k_1}^{im} \xi_{k_2}^{im} \rangle = \langle \xi_{k_3}^{re} \xi_{k_4}^{im} \rangle = 0$, так как числа исходной последовательности некоррелированы.

Аналогично найдем, что $\langle \eta_{l_1}^{im} \eta_{l_2}^{im} \rangle = C_\eta^{im}(l_1, l_2) = C_\eta^{re}(l_1, l_2)$ и

$$C_{\eta}^{re/im}(l_1,\ l_2) = <\eta_{l_1}^{re}\eta_{l_2}^{im}> = (<\xi^2>/N)\sum_{k=0}^{N-1}\eta_k^2 \sin\left[2\pi k\,(l_2 - l_1)/N\right]$$

Задавшись только первыми $(N/2)-1$ коэффициентами h_k (при N четном) и положив $h_k = h_{N-k}$, можно получить:

$$<\eta_{l_1}^{re}\eta_{l_2}^{re}> = <\eta_{l_1}^{im}\eta_{l_2}^{im}> = (<\xi^2>/\sqrt{N})\,(1/\sqrt{N})\sum_{k=0}^{N-1} h_k^2 \times$$

$$\times \exp\left[i\,2\pi k\,(l_2 - l_1)/N\right]$$

$$<\eta_{l_1}^{re}\eta_{l_2}^{im}> = 0,$$

т.е. что ковариационная функция вещественной и мнимой части преобразованной последовательности есть дискретное преобразование Фурье от набора коэффициентов $\{h_k^2\}$, k=0, 1, ..., N-1, и что между собой вещественная и мнимая части некоррелированы. Отсюда вытекает, что коэффициенты $\{h_k^2\}$ следует выбирать как отсчеты требуемого энергетического спектра последовательностей. Описанный способ позволяет получить из двух последовательностей некоррелированных одинаково распределенных чисел две некоррелированные между собой последовательности гауссовских чисел с заданным энергетическим спектром.

При использовании данного способа необходимо иметь в виду следующие нормировочные соотношения:

(6.5)

(6.6)

$$<(\eta_i^{re})^2> = <(\eta^{im})^2> = \frac{<\xi^2>}{N}\sum_{k=0}^{N-1} h_k^2;$$

$$<\eta^{re}> = \frac{1}{N}\sum_{l=0}^{N-1}\eta_l^{re} = \frac{h_0\xi_0^{re}}{\sqrt{N}};$$

$$<h^{im}> = \frac{1}{N}\sum_{l=0}^{N-1}\eta_l^{im} = \frac{h_0\xi_0^{im}}{\sqrt{N}},$$

(6.7)

определяющие связь между дисперсиями чисел исходной и преобразованной последовательностей и средние значения преобразованной гауссовской последовательности. В частности, для того-чтобы гауссовские последовательности имели нулевое среднее, достаточно положить

$$\xi_0^{re} = \xi_0^{im} = 0$$

(6.8)

Кроме выигрыша в быстродействии, обеспечиваемого данным способом получения гауссовских коррелированных последовательностей по сравнению с методом скользящего суммирования, он имеет еще одно важное достоинство. Этот способ экономнее использует исходную последовательность независимых чисел: на выработку одного гауссовского числа затрачивается не более одного числа исходной последовательности. Это особенно важна в моделировании изображающих и других оптических систем, когда приходится формировать массивы объемом в миллионы отсчетов, не допуская повторений.

6.3. ИЗМЕРЕНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК СИГНАЛОВ

Для статистического описания сигналов используют обычно следующие характеристики: функции распределения вероятностей, моменты функции распределения, ковариационные (корреляционные) функции, энергетические спектры.

Измерение функций распределения вероятностей. Функция распределения P(X) случайной величины x – это вероятность того, что случайная величина не превышает значение X. Производная P(X) по X

$$p(x) = dP(X)/dX \mid_{X=x}$$

называется плотностью распределения вероятностей случайной величины x. Цифровые последовательности, отсчеты которых принимают конечное множество значений и количество

отсчетов которых также конечно, характеризуются дискретными аналогами функции распределения и плотности распределения – относительной долей P(т) отсчетов, чье значение не превышает заданную величину т, и относительной долей h(m) отсчетов, имеющих заданное значение т. Последняя характеристика называется гистограммой распределения значений.

Гистограмма, описывающая частоту появления значений отдельных отсчетов сигнала независимо от значений других отсчетов, называется одномерной, или гистограммой одномерного распределения. Гистограмма, характеризующая частоту совместного появления значений нескольких отсчетов сигнала, называется многомерной, или гистограммой многомерного распределения.

Существует простой алгоритм определения n-мерной гистограммы распределения h(m), суть которого сводится к следующему. Для значений гистограммы отводится многомерный массив памяти процессора, размерность которого равна размерности гистограммы я, и в каждой ячейке массива подсчитывается относительное число совместных появлений значений *n* компонент отсчетов сигнала, равных координатам этой ячейки в массиве. Делается это на каждом шаге выборки заданной совокупности отсчетов путем добавления константы, равной обратной величине общего количества отсчетов сигнала, в ячейку массива, адрес которой A определяется значениями компонент наблюдаемых отсчетов:

$$A = \sum_{i=1}^{n-1} m_i \prod_{k=0}^{i-1} M_k + m_0 + A_0$$

где A_0 – начальный адрес массива гистограммы; m_i–квантованное значение i-й компоненты данного отсчета сигнала; M_k– количество уровней квантования k-й компоненты.

Математически эту процедуру можно описать как усреднение δ-функции Кронекера:

$$h(\mathbf{m}) = \frac{1}{N} \sum_{k=0}^{N-1} \delta(\mathbf{m} - \mathbf{m}_k) \qquad (6.9a)$$

где $\mathbf{m}_K$– вектор значений компонент *k*-го отсчета сигнала; **m** – вектор значений компонент аргумента гистограммы ($m_0, m_1, \ldots, m_{n-1}$) или для двумерного сигнала

$$h(\mathbf{m}) = \frac{1}{N_1 N_2} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} \delta(\mathbf{m} - \mathbf{m}_{k,l}). \qquad (6.9б)$$

Эти выражения можно рассматривать как цифровую свертку [см. (3.9)], а в двумерном случае – как разделимую двумерную цифровую свертку.

Гистограмма может использоваться как характеристика не только всего наблюдаемого изображения или всей реализации двумерного сигнала, но и отдельных их участков, или фрагментов. В этом случае она называется локальной.

Измерение локальных гистограмм применяется во многих адаптивных алгоритмах обработки изображений (см. гл. 9). Если требуется измерять локальные гистограммы перекрывающихся фрагментов изображения, для ускорения вычислений целесообразно воспользоваться тем, что формулы (6.9) могут быть записаны в виде рекурсивного соотношения между гистограммами соседних фрагментов. Действительно, если рассматривать (6.9) как выражение для локальной гистограммы (r, s)-го фрагмента:

$$h^{(r,s)}(\mathbf{m}) = (1/N_1 N_2) \sum_{k=rk_0}^{N_1+rk_0-1} \sum_{l=sl_0}^{N_2+sl_0-1} \delta(\mathbf{m} - \mathbf{m}_{k,l}),$$

где k_0, l_0 – интервал следования фрагментов по двум координатам, то оно может быть представлено через гистограмму, например, (r-l,s)-го фрагмента, как

$$h^{(r,s)}(\mathbf{m}) = h^{(r-1,s)}(\mathbf{m}) + (1/N_1 N_2) \sum_{k=(r-1)k_0}^{rk_0-1} \sum_{l=sl_0}^{N_2+sl_0-1} \delta(\mathbf{m} - \mathbf{m}_{k+N,l}) -$$

$$- (1/N_1 N_2) \sum_{k=(r-1)k_0}^{rk_0-1} \sum_{l=sl_0}^{N_2+sl_0-1} \delta(\mathbf{m} - \mathbf{m}_{k,l}).$$

Смысл этой формулы очевиден: гистограмма данного фрагмента может быть получена из гистограммы соседнего сдвинутого на один элемент фрагмента, если прибавить к ней разность гистограмм, вычисленных по тем участкам фрагментов, которые не принадлежат

одновременно им обеим.

Гистограммы, измеренные по небольшим реализациям сигнала (небольшим фрагментам изображений, полей), обычно довольно изрезаны. При увеличении количества измерений гистограмма сглаживается. Однако иногда необходимо получить сглаженную гистограмму при малом количестве измерений. Наиболее употребительны три метода сглаживания.

Ступенчатое сглаживание. Диапазон значений аргумента гистограммы разбивается на небольшое число интервалов. Значения гистограммы внутри каждого интервала заменяются средним значением по интервалу. Такая сглаженная гистограмма может быть построена сразу, если значения сигнала перед изменением гистограммы проквантовать с интервалом квантования, равным интервалу сглаживания.

Сглаживание скользящим суммированием. Значения сглаженной гистограммы $\hat{h}(\hat{\mathbf{m}})$ получаются из исходной гистограммы h(m) путем цифровой свертки ее с некоторой сглаживающей функцией $w(\mathbf{p})$:

$$\hat{h}(\mathbf{m}) = \sum_{\mathbf{p}} w(\mathbf{p})\, h(\mathbf{m} - \mathbf{p})$$

В качестве сглаживающих функций $w(\mathbf{p})$ выбирают более или менее быстро спадающие функции номера **p**. Простейшая сглаживающая функция – прямоугольное «окно»:

$$w(\mathbf{p}) = \prod_{i=0}^{n-1} \operatorname{rect}(p_i + N_i)/2N_i.$$

Сглаживание с помощью ортогональных преобразований. Вычисляются коэффициенты представления гистограммы h(m) понекоторому ортонормальному базису $\{\varphi_{\mathbf{s}}(\mathbf{m})\}$:

$$\eta(\mathbf{s}) = \sum_{m_{n-1}=0}^{M_{n-1}-1} \cdots \sum_{M_0=0}^{M_0-1} h(\mathbf{m})\, \varphi_{\mathbf{s}}(\mathbf{m}) \qquad (6.10)$$

Часть коэффициентов {обычно те из них, которые имеют малые значения) заменяются нулями. Сглаженная гистограмма получается в результате обратного преобразования коэффициентов $\{\eta(\mathbf{s})\}$ после такой отбраковки.

Сглаженную таким образом гистограмму можно получить и сразу в процессе измерения, если воспользоваться следующим приемом. Подставим в (6.10) выражение (6.9а). Тогда получим:

$$\eta_i(\mathbf{s}) = (1/N) \sum_{m_{n-1}=0}^{M_{n-1}-1} \cdots \sum_{m_0=0}^{M_0-1} \sum_{k=0}^{N-1} \delta(\mathbf{m} - \mathbf{m}_k)\, \varphi_s(\mathbf{m}) =$$

$$= (1/N) \sum_{m_{n-1}=0}^{M_{n-1}-1} \cdots \sum_{m_0=0}^{M_0-1} \varphi_s(\mathbf{m}_k).$$

Это значит, что коэффициенты $\eta(\mathbf{s})$разложения $h(\mathbf{m})$ по базису $\{\varphi_{\mathbf{s}}(\mathbf{m})\}$ могут быть найдены усреднением по всем отсчетам сигнала значений базисных функций, вычисляемых каждый раз по значениям $\{\mathbf{m}_k\}$ наблюдаемых отсчетов. Если в результате нужно иметь сглаженную гистограмму, то можно просто не вычислять значения $\eta(\mathbf{s})$ при тех **s**, которым соответствуют обнуляемые коэффициенты. Это экономит машинное время и память.

Оценка моментов распределений. Основными моментами распределений, используемыми в статистических измерениях, являются первый и второй, т.е. среднее значение

$$\bar{x} = \int_{\bar{X}} x p(x)\, dx,$$

и дисперсия

$$D = \int_{\bar{X}} x^2 p(x)\, dx,$$

а также стандартное отклонение (среднеквадратнческое значение)

$$\sigma = \left(\int_{\bar{X}} (x - \bar{x})^2 p(x)\, dx \right)^{1/2}.$$

Дискретными аналогами этих определений являются:

(6.11а)

(6.11б)

$$\bar{m} = \sum_{m=0}^{M-1} m h(m);$$

$$D_m = \sum_{m=0}^{M-1} m^2 h(m);$$

$$\sigma_m = \left(\sum_{m=0}^{M-1} (m - \bar{m})^2 h(m) \right)^{1/2}.$$

(б.11в)

Подставив в (6.11) выражение (6.9) для $h(m)$, получим, что первый и второй моменты цифрового сигнала могут быть найдены усреднением наблюдаемых значений сигнала и их квадратов:

$$\bar{m} = \frac{1}{N_1 N_2} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} m_{k,l};$$

(6.12а)

(6.12б)

$$D_m = \frac{1}{N_1 N_2} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} m_{k,l}^2;$$

$$\sigma_m^2 = \frac{1}{N_1 N_2} \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} (m_{k,l} - \bar{m})^2.$$

(6.12в)

Как и гистограмма распределения, моменты могут использоваться в качестве характеристик всего изображения или отдельных его фрагментов. Вычисление таких локальных моментов

$$\bar{m}(r,s) = \frac{1}{(2N_1+1)(2N_2+1)} \sum_{k=-N_1}^{N_1} \sum_{l=-N_2}^{N_2} m_{r-k,\,s-l}, \quad (6.13а)$$

$$\sigma_m(r,s) = \frac{1}{(2N_1+1)(2N_2+1)} \sum_{k=-N_1}^{N_1} \sum_{l=-N_2}^{N_2} (m_{r-k,\,s-l} - \bar{m}(r,s))^2$$

(6.13б)

по окрестности размером $(2N_1+1)(2N_2+1)$ элементов, очевидно, может быть организовано как рекурсивная фильтрация (см. §4.3).

Формулы (6.11) и (6.13) дают хорошую оценку среднего для текстурных моделей изображений. Если же изображения состоят из случайных деталей и фона, оценки, полученные по этим формулам, нельзя считать удовлетворительными, так как в них статистические характеристики деталей и фона смешиваются. Преодолеть это смешивание можно с помощью так называемых *робастных*, или *устойчивых оценок* параметров распределений. Термин «робастность» означает нечувствительность к малым отклонениям от предположений ([19]). По отношению к оценкам параметров распределений – это нечувствительность к малым посторонним примесям в распределениях. В известном смысле можно считать, что робастные оценки адаптируются к основному распределению.

Наиболее хорошо известны и исследованы следующие устойчивые оценки среднего: медиана выборки и α-усеченное среднее.

Медиана выборки – это такое значение MED элемента выборки, для которого количество элементов выборки, имеющих большее MED и меньшее MED значений, одинаково. Медиану можно найти, расположив элементы выборки в ряд по возрастанию, или в **вариационный ряд** и взяв средний элемент этого ряда. Медиану можно найти также из гистограммы распределения значений выборки, решив уравнение

$$\sum_{m=0}^{MED} h(m) = 0{,}5.$$

α-усеченное среднее – это среднее арифметическое значение элементов вариационного ряда, отстоящих не менее, чем на а элементов от его концов.

Медиана распределения и α-усеченное среднее устойчивы к примесям распределения, влияющим на его хвосты. Так, для изображений, содержащих небольшие детали на некотором фоне, они дают хорошую оценку среднего значения фона.

Дисперсия и стандартное отклонение – это параметры, характеризующие размах распределения. Из устойчивых к хвостам распределений оценок размаха упомянем α-усеченную дисперсию – среднее арифметическое квадратов значений элементов вариационного ряда, отстоящих не менее, чем на a элементов от его концов, и так называемый *квазиразмах* [29], который находится как разность значений элементов вариационного ряда, находящихся справа от медианы на месте R (или, как говорят, имеющих ранг R) и слева от медианы на месте L (имеющих ранг L).

Для быстрого вычисления локальных медианы, α-усеченного среднего, α-усеченной дисперсии и квазиразмаха можно использовать локальную гистограмму, которую, как было показано выше, можно находить рекурсивно. Кроме того, при обработке таких сигналов, как изображения, локальные статистические характеристики которых мало изменяются от элемента к элементу, для ускорения вычислений можно предсказывать значения характеристики по предыдущему элементу и далее находить только поправку.

Ранг заданного элемента выборки, т.е. номер этого элемента в вариационном ряду также является важной статистической характеристикой сигнала, связанной с гистограммой распределения его значений. Ранг показывает, сколько элементов выборки имеет значение, меньшее чем значение данного элемента. Если выборка образована фрагментом сигнала, например фрагментом изображения, при его сканировании, ранг (так же, как и медиана и другие рассмотренные выше характеристики) является локальным. Локальный ранг также можно вычислять рекурсивно, причем не только через локальную гистограмму, но и непосредственно. Для этого достаточно подсчитать разность между количеством вновь появившихся на данном шаге сканирования элементов фрагмента, меньших по значению данного, и таких же элементов, ушедших из фрагментов, и прибавить эту разность к величине ранга на предыдущем шаге.

Оценка корреляционных функций и спектров. Простейшими характеристиками сигналов, трактуемых как стационарные (пространственно однородные) случайные процессы и поля, являются корреляционные функции и энергетические спектры. Эти характеристики определяются следующим образом (см., например, [8]).

Взаимная корреляционная функция процессов *a(x) и b(x)*:

$$R_{a,b}(\xi) = \lim_{X\to\infty} \frac{1}{X} \int_{x}^{x+X} a(x)\, b^*(x+\xi)\, dx.$$

Автокорреляционная функция процесса

$$R_a(\xi) = \lim_{X\to\infty} \frac{1}{X} \int_{x}^{x+X} a(x)\, a^*(x+\xi)\, dx.$$

Энергетический спектр процесса *a(x)*:

$$A(f) = \int_{-\infty}^{\infty} R_a(\xi) \exp(i\, 2\pi f \xi)\, d\xi.$$

Применяя теорему отсчетов (см. § 2.2), можно получить соответствующие определения для дискретных процессов – результатов дискретизации непрерывных процессов:

$$R_{a,b}(n) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} a_k b^*_{k+n};$$

$$R_a(n) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} a_k a^*_{k+n};$$

$$A(r) = \frac{1}{\sqrt{M}} \sum_{k=0}^{M} R_a(k) \exp(i\, 2\pi k r/M),$$

где a_k, b_k, $R_a(n)$, $A(r)$ – отсчеты процессов, корреляционных функций и спектров соответственно, а M – количество отсчетов корреляционной функции.

Эти формулы предполагают, что анализируемые сигналы имеют бесконечную протяженность. Для цифровых же сигналов речь может идти лишь о том, что количество отсчетов N более или менее велико, но конечно, и отсчеты с номерами, большими N, не определены. Поэтому получающиеся при конечном N формулы для корреляционных функций

$$R_{a,b}(n) = \frac{1}{N}\sum_{k=0}^{N-1} a_k b^*_{k+n}, \qquad (6.14а)$$

$$R_a(n) = \frac{1}{N}\sum_{k=0}^{N-1} a_k a^*_{k+n} \qquad (6.14б)$$

рассматриваются как оценки корреляционных функций сигналов по их цифровому представлению, а энергетический спектр, получаемый с помощью дискретного преобразования Фурье оценок корреляционных функций, рассматривается как оценка энергетического спектра.

При расчетах по формулам (6.14) приходится сталкиваться с теми же краевыми эффектами, что и при цифровой фильтрации. Поэтому здесь применимы те же соображения о способах доопределения недостающих отсчетов сигналов, которые были высказаны в § 4.3.
Если заменять недостающие отсчеты нулями, то оценка корреляционных функций по формулам (6.14) оказывается смещенной (поскольку часть слагаемых – тем большая, чем больше п – в суммах (6.14) выпадает). В этом случае рекомендуется пользоваться модифицированной усеченной формулой вида

$$R_{a,b}(n) = \frac{1}{N-n}\sum_{k=0}^{N-n} a_k b^*_{k+n}, \qquad (6.15)$$

которая дает лучшую оценку корреляционной функции непрерывного сигнала для больших л при фиксированном N.

Формулы (6.14) и (6.15) для отсчетов корреляционной функции родственны формуле (3.9) цифровой свертки. Поэтому для вычисления корреляционных функций применяются те же алгоритмы, что и для вычисления свертки: прямое вычисление по (6.14) и (6.15); ускоренные алгоритмы с уменьшенным числом умножений, описанные в § 4.4; вычисление с помощью ДПФ и СДПФ (см. § 4.2) и такие же способы доопределения сигнала.

В качестве энергетического спектра изображений обычно используются квадрат модуля коэффициентов его ДПФ или СДПФ (если требуются значения спектра в произвольных точках). В соответствии с теоремой отсчетов разрешающая способность такого метода по частоте равна ширине полосы (для двумерных сигналов – площади пространственного спектра), поделенной на количество отсчетов последовательности, полученной в результате дискретизации сигнала.

Если анализируемые сигналы рассматриваются как реализации некоторого ансамбля случайных сигналов, то для получения спектра, характеризующего весь ансамбль сигналов в целом, необходимо сглаживать оценки спектров, найденные для отдельных реализаций. Для этого прибегают к методам, аналогичным методам сглаживания оценки распределений. Важнейшими из них являются [8]:

Усреднение периодограмм. Анализируемый процесс разбивается на фрагменты, размеры которых соответствуют требуемой разрешающей способности анализа по частоте. Сглаженная оценка спектра $\overline{A}(r)$ находится как среднее

$$\overline{A}(r) = \frac{1}{K}\sum_{i=1}^{K} A_i(r)$$

оценок спектра $A_i(r)$ для каждого фрагмента.

Маскирование анализируемого процесса гладкой функцией. Если размер реализации анализируемого процесса недостаточно велик, то, чтобы воспользоваться первым методом, ее умножают на некоторую гладкую функцию, более или менее плавно спадающую к краям (так

называемую функцию «окна»), и находят энергетический спектр такого маскированного процесса.

Прямое сглаживание спектра осуществляется путем свертки полученной оценки энергетического спектра с нормированной гладкой функцией, простирающейся на несколько отсчетов:

$$\overline{A}(r) = \sum_{k=-K}^{K} A(r+k)\,w(k); \qquad \sum_{k=-K}^{K} w(k) = 1.$$

Этот метод наименее эффективен в вычислительном отношении, но может оказаться полезным тогда, когда требуется иметь как сглаженную, так и несглаженную оценки спектра.

6.4. ИЗМЕРЕНИЕ ПАРАМЕТРОВ СЛУЧАЙНЫХ ПОМЕХ

Принципы оценки параметров помех. Задача измерения параметров случайных помех при обработке изображений, голограмм и интерферограмм обычно возникает при коррекции искажений этих сигналов в изображающих, голографических и интерферометрических системах. Знание этих параметров, как правило, необходимо для построения соответствующих корректирующих преобразований. Иногда нужные данные можно получить, зная конструктивные характеристики соответствующих систем, например отношение сигнал-шум в телевизионных и фототелевизионных изображающих системах, параметры зернистости фотоматериалов при известных условиях экспонирования и фотохимической обработки и т.п. Но чаще всего на практике такие данные отсутствуют, и параметры помех и искажений приходится определять непосредственно по наблюдаемому уже искаженному сигналу.

На первый взгляд может показаться, что эта задача внутренне противоречива: для того чтобы найти оценку параметров шума по наблюдаемой смеси, необходимо шум отделить от сигнала, а это можно сделать только, зная параметры шума. Выход из этого «порочного круга» заключается в том, чтобы не разделять сигнал и шум для определения статистических характеристик шума, а разделять их характеристики на основе измерений соответствующих характеристик наблюдаемого «зашумленного» сигнала [47].

Задача разделения характеристик сигнала и шума может решаться как детерминированная, если соответствующие характеристики искаженного сигнала известны точно, или как статистическая задача оценки параметров. В последнем случае анализируемые характеристики сигнала следует рассматривать как случайные величины, если это числа, или как случайные последовательности, а характеристики, найденные для наблюдаемого сигнала, как их реализации.

Построение оптимальных процедур оценки параметров помех в соответствии с этим подходом в принципе нужно основывать на статистических моделях анализируемых характеристик, которые необходимо строить и обосновывать конкретно для каждой избранной характеристики. Однако в большинстве практических случаев помехи ъ статистическом смысле являются очень простыми объектами, т.е. описываются небольшим числом параметров, поэтому такая редуцированная задача оценки параметров помех может быть решена сравнительно простыми средствами даже при весьма грубом априорном задании статистических характеристик измеряемых характеристик видеосигнала. Необходимо только из всех доступных измерению характеристик сигнала выбрать такие, в которых искажение сигнала шумом проявлялось бы в возможно более просто обнаруживаемом их аномальном поведении.

Опишем два просто реализуемых при цифровой обработке и в то же время достаточно универсальных метода обнаружения, основанных на общей априорной предпосылке о гладкости характеристик неискаженного сигнала: предсказание и «голосование» [47]. Эти методы родственны развиваемым в последнее время *стабильным (робастным) методам* оценки параметров (см., например, [19]).

Метод *предсказания* состоит в том, что для каждого данного элемента анализируемой последовательности находится отличие его значения от значения, предсказанного по предыдущим уже обследованным элементам. Если отличие превышает некоторый заданный

порог, принимается решение о наличии аномального выброса. Глубина предсказания, способ определения предсказанного значения и порог должны при этом задаваться априори для данного класса сигналов.

Простейший способ предсказания использует в качестве предсказанного значения элемента анализируемой последовательности значение предыдущего элемента последовательности. Более точное предсказание может быть осуществлено путем сложения с некоторыми весами нескольких предыдущих значений, причем оптимальные значения весовых коэффициентов могут находиться из условия повторения первой, второй и далее разностей анализируемой последовательности, либо из той или иной ее авторегрессионной модели [10].

Метод «голосования» является обобщением известного метода медианного сглаживания (см., например, [64]) и вариантом ранговых алгоритмов обнаружения (см. гл. 9). Он заключается в том, что каждый элемент анализируемой последовательности рассматривается одновременно с некоторым количеством 2n его ближайших соседних элементов. Эта выборка из 2n+1 значений упорядочивается по возрастанию или убыванию и проверяется, не попало ли значение данного элемента в заданное число *k* крайних (т.е. наибольших или наименьших) значений упорядоченной выборки. При положительном ответе принимается решение о наличии аномального большого (или, соответственно, малого) значения в данном элементе. Метод «голосования» основан на предположении, что «нормальная» анализируемая характеристика, как правило, локально монотонна, и отклонения от локальной монотонности, если они имеются, невелики. Величины *n* и *k* задаются априори из предположений о «нормальном» поведении исследуемой характеристики неискаженного сигнала.

Оценка параметров аддитивного независимого от сигнала флуктуационного широкополосного шума на изображении. Важнейший характеристикой аддитивного и статистически независимого от сигнала флуктуационного шума является его стандартное отклонение и корреляционная функция. Если, как это часто бывает, шум является некоррелированным или слабо коррелированным, для определения его дисперсии и корреляционной функции можно построить следующий простой алгоритм, основанный на измерении аномалий в ковариационной функции наблюдаемого изображения [46].

Благодаря аддитивности и .независимости шума ковариационная функция $C_н(r, s)$, измеренная по Q наблюдаемым изображениям размером $NixN_2$ элементов, является суммой ковариационной функции незашумленного изображения $C_0(r, s)$, ковариационной функции шума $C_ξ(r, s)$ и реализации некоторого случайного процесса $ε(r, s)$, характеризующего ошибку измерения ковариационной функции шума по его реализации конечных размеров:

$$C_н(r, s) = C_0(r, s) + C_ξ(r, s) + ε(r, s). \quad (6.16)$$

Как известно, дисперсия случайного процесса $ε(r, s)$ обратно пропорциональна объему выборки QNM, по которой производилось измерение $C_н(r, s)$. Этот объем обычно превышает сотни тысяч отсчетов. Поэтому случайная ошибка $ε(r, s)$ в (6.16) мала, так что $C_ξ(r, s)$ можно оценить как

$$C_ξ(r, s) = C_н(r, s) - C_0(r, s)$$

Рассмотрим сначала случай некоррелированного шума, когда $C_ξ(r, s) = σ_ξ^2 δ(r, s)$

где $σ_ξ^2$ – дисперсия шума; $δ(r, s)$ – дельта-функция (символ) Кронекера. В этом случае ковариационная функция наблюдаемого изображения отличается от ковариационной функции незашумленного изображения только в начале координат, и это отличие равно дисперсии шума:

$$σ_ξ^2 = C_н(0, 0) = C_0(0, 0) \quad (6.17)$$

а для всех остальных значений (r, s) величина $C_н(r, s)$ может служить оценкой $C_0(r, s)$: $C_н(r, s) = C_0(r, s)$

Как показывают измерения корреляционных функций изображений, вблизи начала координат (r= 0, s = 0) они являются весьма медленно меняющимися функциями г и s. Поэтому величину C_0 (0,0), необходимую для вычисления дисперсии некоррелированного шума по (6.17), можно с высокой точностью оценить экстраполяцией по значениям $C_0(r, s) = C_н(r, s)$ в точках r, s вблизи нуля. Таким образом, для определения дисперсии аддитивного некоррелированного шума на изображении достаточно измерить ковариационную функцию

$C_н(r, s)$ наблюдаемого изображения в малой окрестности вблизи точки (0, 0), найти экстраполяцией оценку C_0 (0, 0) величины C_0(0, 0) и применить в качестве оценки

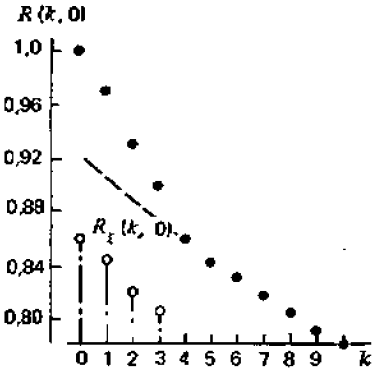


Рис. 6.7. Оценка ковариационной функции широкополосного шума на изображении

дисперсии величину $\sigma_ξ^2 = C_н(0, 0) - \hat{C}_0(0,0)$. Эксперименты показывают, что хорошая оценка получается даже при экстраполяции по одномерным сечениям ковариационной функции [46].

Подобный подход можно использовать и для оценки дисперсии и ковариационной функции слабо коррелированного шума, т.е. шума с функцией ковариации $C_ξ(r, s)$, отличной от нуля лишь в небольшой области вблизи начала координат, где значения ковариационной функции незашумленного изображения можно удовлетворительно экстраполировать по значениям $C_н(r, s)$ в тех точках, где $C_ξ(r, s)$ заведомо равна нулю. Примерные размеры области, в пределах которой сосредоточены ненулевые значения $C_ξ(r, s)$, и гладкость $C_н(r, s)$ в окрестности этой области постулируются априорно.

Для иллюстрации на рис. 6.7 показана в полулогарифмическом масштабе ковариационная функция реального изображения. На графике хорошо заметен излом ковариационной функции и штриховой линией даны значения ковариационной функции изображения вблизи нуля, полученные путем экстраполяции по остальным точкам. Внизу на этом рисунке приведена разность исходной и экстраполированной функций, служащая оценкой ковариационной функции шума на изображении.

Оценка параметров аддитивного широкополосного шума на интерферограммах с пространственной несущей. Для измерения интенсивности широкополосного шума на интерферограммах с пространственной несущей (как, например, на рис. 6.8, а) можно в принципе использовать описанный выше метод измерения шума на изображениях по корреляционной функции наблюдаемого сигнала. Однако поскольку фильтрацию шума удобнее выполнять обработкой сигнала в спектральной области, оценку параметров сигнала и шума, необходимых для построения фильтра, также лучше производить по спектру сигнала.

Идеальная, т.е. нез а шум ленная интерферограмма, представляет собой двумерный синусоидальный сигнал. Как вытекает из свойств дискретного преобразования Фурье, в энергетическом спектре двумерного сигнала имеется острый пик вблизи значения средней пространственной частоты интерферограммы (см. рис. 6.8, б). Если интерферограмма содержит аддитивный шум, как в случае, показанном на рис. 6.8, а, пик в спектре наблюдается также на фоне шума (рис. 6.8, а). Задача оценки параметров сигнала и шума по их наблюдаемому энергетическому спектру

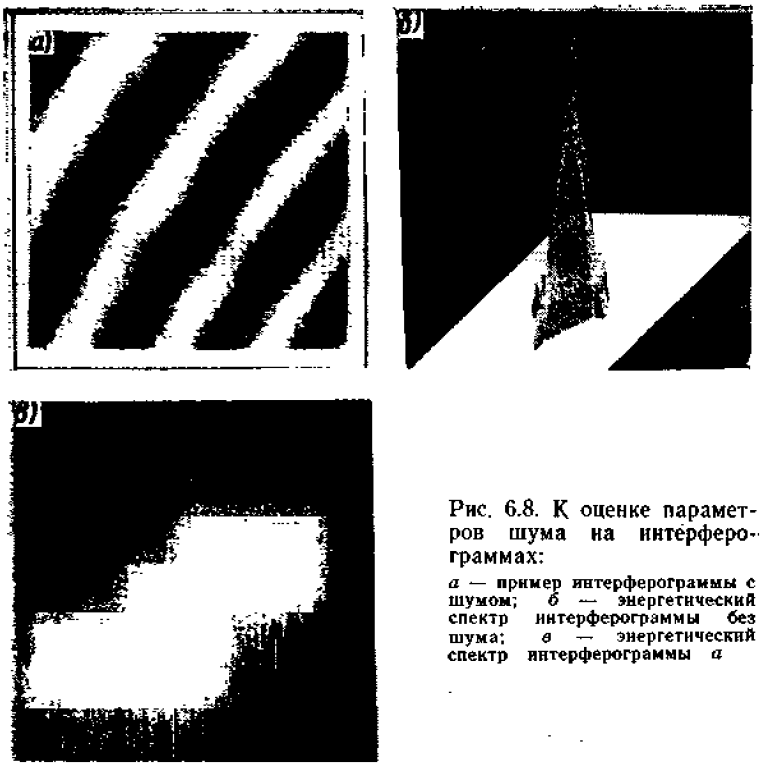


Рис. 6.8. К оценке параметров шума на интерферограммах:
а — пример интерферограммы с шумом; *б* — энергетический спектр интерферограммы без шума; *в* — энергетический спектр интерферограммы *а*

сводится, очевидно, к задаче обнаружения пика сигнала в его спектре на фоне шума и выделения того участка спектральной плоскости, где интенсивность спектральных компонент сигнала существенно отлична от нуля.

Найти границы этого участка можно на основании априорных данных о средней пространственной частоте интерферограммы, которая определяется схемой интерферометра, и о максимальной площади пространственного спектра, определяемой априорными сведениями об объекте интерферометрических измерений.

Достаточно хорошие оценки спектральной плотности мощности шума получаются простым усреднением спектра зашумленной интерферограммы по априори известным периферийным участкам спектральной плоскости, не занятым спектром сигнала. Отметим, что для лучшей очистки периферии спектра от хвостов пика спектра сигнала интерферограммы следует пользоваться описанными в гл. 4 методами борьбы с краевыми эффектами при спектральном анализе.

Возможно и более полное использование априорных данных об идеальной интерферограмме и аддитивном шуме для определения параметров шума. Так, в [33, 34] предложен алгоритм, который' для каждой спектральной компоненты по отношению правдоподобия принимает решение о том, принадлежит ли она сигналу или шуму. Алгоритм использует априори известный закон распределения интенсивности спектральных компонент аддитивного шума – закон распределения Рэлея (это распределение благодаря нормализующему действию преобразования Фурье имеет место для шума с достаточно произвольным распределением) и простейшую гипотезу о равномерном распределении амплитуды спектральных компонент сигнала. Такая гипотеза соответствует предположению о треугольной форме пика сигнала в частотной области.

Эксперименты (см. [33, 34]) показали, что, используя такую методику определения области спектральной плоскости, не занятой сигналом интерферограммы, можно добиться высокой степени подавления шума на ннтерферограммах.

Оценка интенсивности и частоты гармонических составляющих периодических и других помех с узким спектральным составом. Периодические (муаровые) помехи возникают чаще всего в теле-

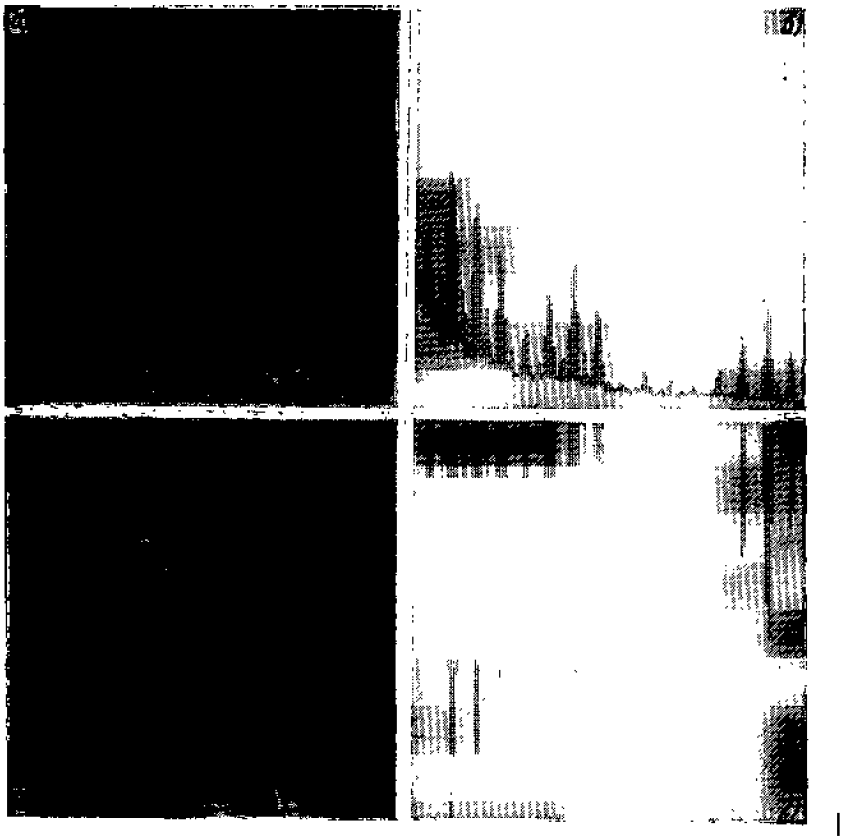


Рис. 6.9. Фильтрация периодических помех:
а – исходное зашумленное изображение; б – усредненный спектр Уолша вдоль строк; в – результат фильтрации; г – характеристика фильтра-маски (график значений отсчетов частотной характеристики фильтра)

визионных и фототелевизионных системах с передачей видеосигнала по радиоканалу. Иногда они появляются вследствие дискретизации изображений, содержащих детали с высокочастотными периодическими структурами (см. гл.2), или как интерференционные эффекты для изображений, получаемых в когерентно-оптических изображающих системах.

Характерной особенностью таких помех является то, что их спектр в базисе Фурье содержит только небольшое число заметно отличных от нуля компонент. К помехам этого класса можно отнести также помехи, имеющие небольшое число компонент в других базисах (рис. 6.9,а, б). В то же время пространственный спектр «незашумленных» изображений в базисе Фурье и ряде других базисов (например, в базисах Уолша, косинусного преобразования) является обычно более или менее гладкой и монотонной функцией. Поэтому наличие узкополосного шума проявляется в виде аномально больших и локализованных отклонений, или выбросов, в спектре искаженных изображений. Расположение этих выбросов в отличие от рассмотренного выше случая флуктуационного шума, дававшего выбросы корреляционной функции в нуле, в данном случае обычно не известно. Локализовать выбросы можно, применяя описанные выше методы предсказания или «голосования».

Для этого усреднением по всем наблюдаемым изображениям с однотипными периодическими помехами находится среднее значение квадрата модуля спектральных компонент «зашумленного» сигнала $\langle|\overline{\beta_{r,s}}|^2\rangle$ по избранному базису, вычисленных с использованием соответствующих быстрых алгоритмов. Если производится одномерная фильтрация, например вдоль строк изображения, то и усреднение может производиться по всем строкам изображений, подлежащих фильтрации. Затем методами «голосования» или предсказания производится обнаружение локализованных компонент шума, т.е. отмечаются те спектральные компоненты $\langle|\overline{\beta_{r,s}}|^2\rangle$ наблюдаемого сигнала, которые искажены шумом.

В силу аддитивности шума величины $\langle|\overline{\beta_{r,s}}|^2\rangle$, очевидно, равны сумме интенсивности спектральных компонент «незашумленного» сигнала $\langle|\overline{\alpha_{r,s}}|^2\rangle$ и шума $|\overline{x_{r,s}}|^2$ (см. гл. 7). Следовательно, $|\overline{x_{r,s}}|^2 = \langle|\overline{\beta_{r,s}}|^2\rangle - \langle|\alpha_{r,s}|^2\rangle$. Используя априорную гладкость

спектра «незашумленного» сигнала, необходимые значения $<|\alpha_{r,s}|^2>$ можно найти интерполяцией по ближайшим к точке $<\overline{|\beta_{r,s}|^2}>$ (r, s) не отмеченным как искаженные шумом значениям

Оценка параметров импульсного шума, шума квантования, помех типа полосатости на изображениях. Основной статистической характеристикой импульсного шума является вероятность искажения отсчетов сигнала. Эту вероятность можно оценить, подсчитав относительное количество элементов изображения, искаженных выбросами шума. Для обнаружения выбросов шума могут быть использованы описанные методы предсказания и «голосования», если применять их для анализа последовательности значений са-.мого «зашум лен наго» видеосигнала в небольшой окрестности данного изображения. Таким образом, оценка параметров импульсного шума совмещается с процедурой его фильтрации. Подробнее
о диагностике и фильтрации импульсного шума см. в гл. 8, 9.

Шум квантования определяется числом уровней квантования
сигнала. Для определения числа уровней квантования сигнала достаточно построить гистограмму распределения его значений и подсчитать количество значений сигнала, для которых гистограмма отлична от нуля. Таким же образом, обнаруживая методами предсказания или «голосования» провалы в гистограмме, построенной по достаточно большому набору изображений, и измеряя их глубину, можно оценить вероятность характерных для некоторых цифровых изображающих систем сбоев отдельных уровней квантования.

Помехи типа полос на изображениях представляют собой обыч-«о случайные выбросы среднего значения видосигнала, вычисленного в направлении полос. Помехи такого типа характерны для многих видов фото телевизионных изображающих систем со сканированием. Естественной характеристикой изображений, по аномальным выбросам которой можно измерить параметры такого шума, является последовательность средних значений видеосигнала в направлении вдоль полос, например последовательность средних значений вдоль строк развертки, найденных для каждой строки, если полосы расположены вдоль строк. Обнаружить и измерить аномальные выбросы можно теми же методами предсказания и «голосования», опираясь на предположение о том, что на неискаженном изображении средние значения видеосигнала вдоль строк от строки к строке не могут претерпевать значительных изменений. Примеры фильтрации помех типа полосатости, основанной на таком способе диагностики, описаны в гл. 7.

В заключение отметим, что описанные методы диагностики искажений сигналов опираются на априорные предположения о «нормальном» поведении тех или иных характеристик сигнала. Это поведение необходимо изучать отдельно для разных видов сигналов. В настоящее время количественные данные о «нормальных» характеристиках изображений, голограмм и интерферотрамм как сигналов еще недостаточны.

6.5. ПРИМЕРЫ МОДЕЛИРОВАНИЯ ОПТИЧЕСКИХ И ГОЛОГРАФИЧЕСКИХ СИСТЕМ

Цифровая модель оптической системы – это комплекс программ, которые реализуют цифровые преобразования над цифровыми сигналами, соответствующие (в смысле принципа соответствия, сформулированного в гл. 3) непрерывным преобразованиям непрерывных сигналов в моделируемой системе.

С точки зрения программной реализации цифровых моделей можно различать *формальные и блочные модели*. Формульное моделирование – простейший вид моделирования, который состоит в том, что создается программа вычислений требуемых характеристик объекта по аналитическим выражениям, т.е. по аналитическим моделям, описывающим эти характеристики. При таком моделировании цифровая модель, представляющая собой программу вычислений по заданной формуле, является узко специализированной, так что изменение цели моделирования требует, как правило, полной перестройки модели – написания и отладки новой программы.

Блочное моделирование – более совершенная форма моделирования. При блочном моделировании цифровая модель (программа) состоит из отдельных блоков (подпрограмм), обладающих определенной самостоятельностью, благодаря которой программы, реализующие разные цифровые модели, могут набираться из этих блоков, как приборы из типовых деталей. Каждый блок осуществляет некоторую элементарную операцию и является, как правило, формульной моделью того или иного блока реальной системы. Основой построения блочных цифровых моделей являются понятия теории сигналов и систем, рассмотренные в § 1.3, – понятия однородных и неоднородных поэлементных и линейных преобразований и статистические модели случайных объектов. Методы цифрового представления и реализации этих преобразований описаны в гл. 3–5, методы построения цифровых статистических моделей случайных изображений и объектов – в § 6.1, 6.2. Приведем два примера моделирования оптических и голографических систем.

Цифровое моделирование искажений при записи и реконструкции голограмм. Свойство тел диффузно рассеивать падающее на них излучение имеет очень большое значение для возможности наблюдать их, регистрируя это излучение. Его важно учитывать к в физической голографии и при синтезе голограмм. Благодаря диффузности объектов их голограммы оказываются устойчивыми к искажениям, способны восстанавливать изображения объекта с любой своей части и передавать большие перепады яркости объекта при ограниченном динамическом диапазоне материала, на котором зарегистрирована голограмма.

Однако наличие диффузности имеет и отрицательные последствия: при наблюдении диффузных объектов в когерентном свете макроскопические свойства объекта, определяемые разрешающей способностью наблюдателя, маскируются шумом диффузности, или, как его иногда называют, спекл-шумом. Такой же шум, естественно, наблюдается и при восстановлении объектов с их голограмм. Чем больше искажено поле при его регистрации в виде голограммы или при восстановлении голограммы, тем больше шум диффузности.

Можно указать несколько основных факторов, определяющих искажения голограммы при ее записи:

1) ограничение размеров голограммы, т.е. запись не всего Лол я, рассеиваемого объектом, а только части волнового фронта;
2) нелинейные искажения при записи интерференции регистрируемого волнового фронта и опорного пучка, в том числе ограничение динамического диапазона записываемого сигнала, нелинейность при записи, квантование (например, при синтезе голограмм);
3) дискретизация голограмм (при цифровом синтезе);
4) фазовые искажения волнового фронта при восстановлении голограмм.

Другие факторы определяют искажения, возникающие в процессе наблюдения восстанавливаемого с голограмм изображения.ч Это прежде всего конечная разрешающая способность наблюдателя и способность накопления сигнала в пределах элемента разрешения (интегрирование интенсивности или комплексной амплитуды поля).

Аналитическое описание действия этих факторов, и в особенности их совместного действия, является очень трудной и громоздкой задачей. По сути, в настоящее время поддаются расчету лишь влияние усреднения интенсивности в плоскости восстановленного изображе'ния на форму корреляционной функции шума диффузности и отношение сигнал-шум на восстановленном изображении [56].

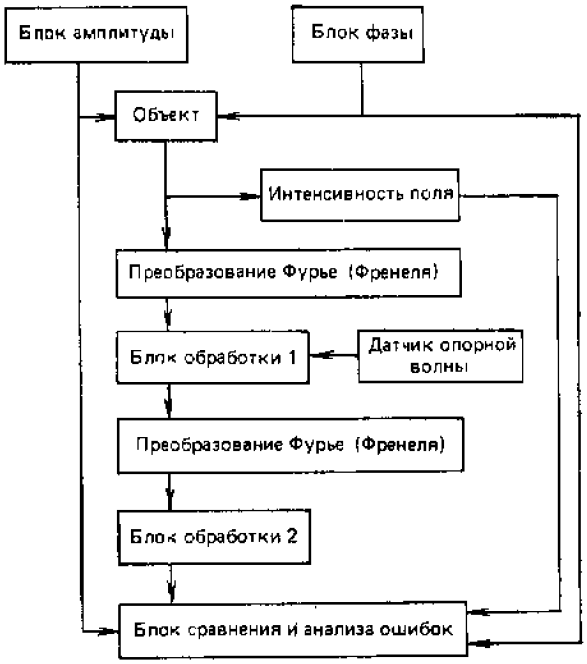


Рис. 6.10. Схема цифровой модели процесса записи и реконструкции голограмм Фурье и Френеля

Использование цифрового моделирования позволяет получить как качественную информацию, необходимую для разработки приближенных аналитических методов расчета, так и конкретную количественную информацию об искажениях и шумах на восстановленном изображении [49].

Структурная схема цифровой модели процесса записи и реконструкции голограмм Фурье и Френеля показана на рис. 6.10. В этой модели объект задается двумя последовательностями чисел, описывающих амплитуду и фазу поля соответственно. Эти последовательности генерируются либо детерминированной функцией, либо в виде последовательности псевдослучайных чисел с гауссовским распределением и заданным энергетическим спектром.

При необходимости моделировать одновременно медленные (макроскопические) и быстрые (микроскопические) изменения фазы поля на объекте последовательность фаз может получаться в виде поэлементной суммы двух отдельных последовательностей. .Далее по этим последовательностям в блоке «Объект» вычисляются последовательности отсчетов вещественной и мнимой частей комплексной амплитуды поля на объекте, а также последовательность отсчетов интенсивности поля, необходимая в дальнейшем .для сравнения с результатом восстановления.

Сформированный массив чисел подвергается дискретному преобразованию Фурье или Френеля (вид преобразования выбирается в зависимости от поставленной задачи), в результате чего образуется массив математической голограммы. Он поступает на «Блок обработки 1», в котором может быть подвергнут преобразованиям, моделирующим процессы регистрации голограмм. Этот блок состоит из следующих подблоков:

- формирования «физической голограммы» путем поэлементного суммирования массива «математической голограммы» с отсчетами опорной волны;
- ограничения сигнала, в котором все значения отсчетов поступающего сигнала, превышающие по модулю задаваемый порог, заменяются значением порога (с сохранением знака);
- квантования, осуществляющего равномерное квантование поступающих отсчетов на задаваемое число уровней в пределах задаваемого диапазона значений;
- амплитудного маскирования голограммы, в котором исходная последовательность умножается на последовательность положительных чисел, определяющих форму маскирующей функции; этот подблок позволяет моделировать аподизацию и ограничение размеров голограмм:

- фазового маскирования, в котором исходная последовательность умножается на заданную последовательность комплексных чисел с единичным модулем, моделирующую фазовые шумы при регистрации голограмм, изгибы фотопленки, и другие эффекты;
- сглаживания, в котором производится скользящее суммирование поступающей последовательности чисел по заданному количеству отсчетов; этот подблок имитирует конечную разрешающую» способность среды, используемой для регистрации голограммы.

Подблоки могут быть включены в произвольной последовательности. Далее следует обратное преобразование Фурье или Френеля, восстанавливающее объект. Результат восстановления может быть подвергнут в «Блоке обработки 2» преобразованиям, моделирующим конечную разрешающую способность устройства-наблюдения голограмм путем скользящего суммирования получающейся последовательности отсчетов комплексной амплитуды$_ч$ или интенсивности.

Результат восстановления сравнивается с исходным объектом в блоке сравнения и анализа ошибок. Здесь находятся дисперсия и корреляционная функция отсчетов вещественной, мнимой частей и квадрата модуля полученных отсчетов, дисперсия и корреляционная функция разностей вещественных, мнимых частей и квадрата модуля исходной и восстановленной последовательностей, а также отношение дисперсии интенсивности восстановленной последовательности к ее среднему значению (спекл-контраст). Результаты сравнения выдаются в виде таблиц и графиков.

Модели оптических систем пространственной фильтрации изображений. Как известно, простейшим методом коррекции искажений в линейных изображающих системах является линейная фильтрация. Ее можно реализовать как средствами цифровой обработки изображений (см. гл. 7), так и средствами аналоговой когерентно-оптической обработки [13], причем последние намного превосходят цифровые средства в потенциальном быстродействии.

Известно также, что применение линейных методов коррекции линейных искажений изображений ограничено, так как в результате коррекции неизбежно возрастает шум, всегда присутствующий в корректируемом сигнале. Поэтому практический интерес представляет выяснение предельных возможностей линейных методов коррекции линейных искажений, в частности таких искажений, как дефокусировка, т.е. определение обменных соотношений между степенью дефокусировки и уровнем аддитивного шума, при котором данную дефокусировку еще можно скорректировать линейной фильтрацией. Поскольку в большинстве применений скорректированное изображение предъявляется для интерпретации человеку-оператору, определить эти соотношения можно только с помощью постановки субъективной экспертизы.

Для проведения субъективной экспертизы необходимо генерировать набор тестовых изображений, скорректированных при разной степени дефокусировки и разном уровне шума, и в процессе экспертизы оценивать по тому или иному критерию качество скорректированных изображений. Создать требуемый набор тестовых изображений можно с помощью цифрового моделирования.

Когерентно-оптические системы пространственной фильтрации, которые можно использовать для коррекции искаженных изображений, осуществляют преобразование изображений в три этапа, которые математически описываются как преобразование Фурье изображения, умножение спектра Фурье изображения на оптическую передаточную (частотную) характеристику корректирующего фильтра и обратное преобразование Фурье, в результате которого восстанавливается скорректированное изображение (см., например, [13]). Структурная схема цифровой модели такой системы показана на рис. 6.11 [21].

Поскольку данная цифровая модель предназначена для изучения влияния уровня аддитивного гауссовского шума в искаженном изображении на возможность коррекции дефокусировки, она построена как двухканальная. В одном канале обрабатывается изображение без шума, в другом только шум. Благодаря этому сокращается количество полных прогонов модели для синтеза тестовых изображений, так как одну и ту же реализацию шума можно использовать для получения разных вариантов скорректированных изображений при разных

отношениях сигнал-шум, меняя только вес шума при смешивании изображения и шума в последнем блоке модели.

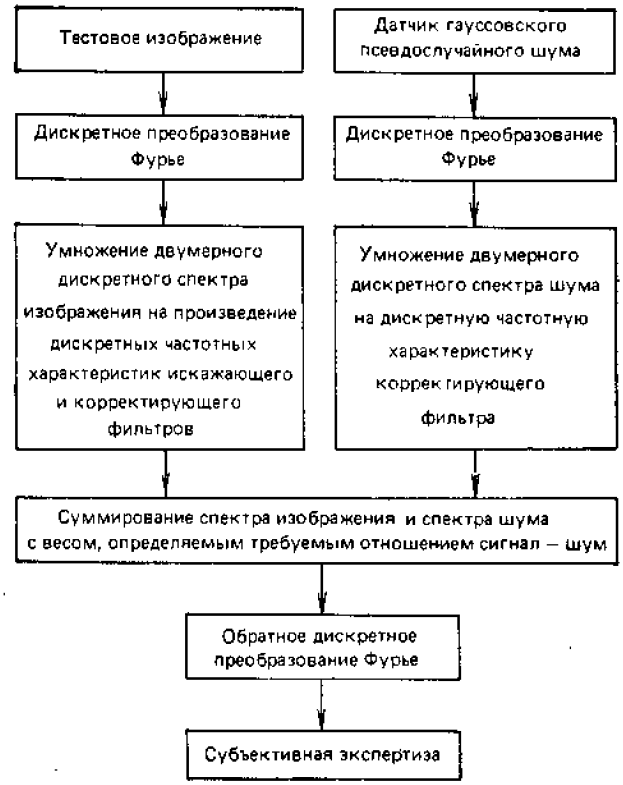
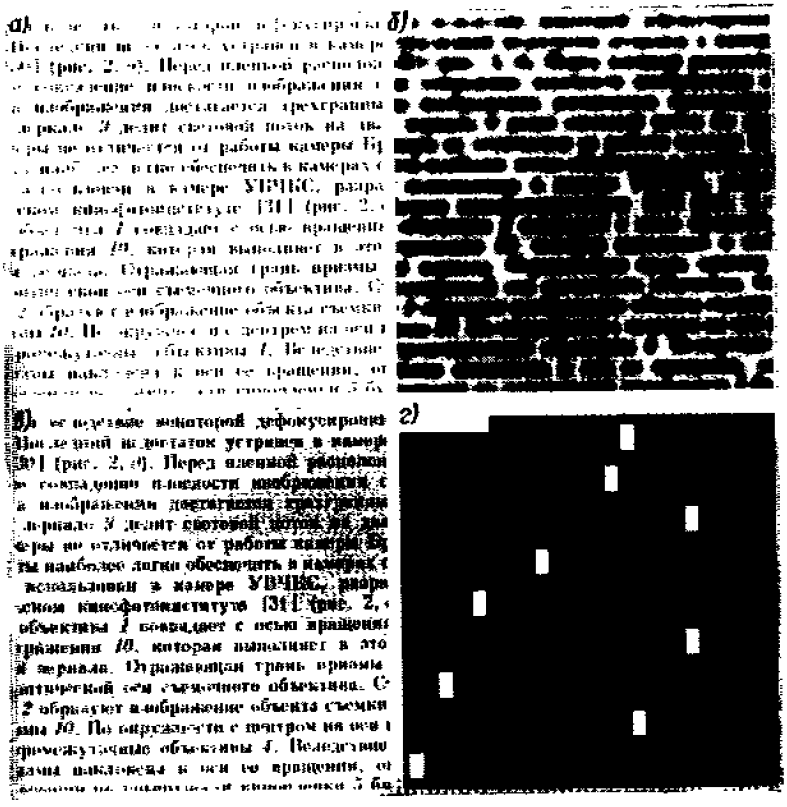


Рис. 6.11. Схема цифровой модели оптической системы коррекции дефокусированных изображений



Рас. 6.12. Изображения, использовавшиеся при моделировании оптических систем коррекций дефокусированных изображений:
а. – кслодное тестовое изображение; б – изображение после дефокусировки с кружком рассеивания, днамгтр которого равен двойной ширине буквы; в – скорректированное изображение при отношении сигнал-шум, равном 20; е–вариант маски для субъективной экспертизы

В качестве тестового изображения выбрано изображение печатного текста (рис. 6.12). Для такого изображения можно сформулировать естественный количественный критерий

качества коррекции – вероятность правильного узнавания букв, и тем самым облегчить постановку субъективной экспертизы. Для того чтобы при экспертизе исключить влияние семантики текста на результат узнавания, наблюдателю предъявлялось не непосредственно само скорректированное изображение, а изображение, закрытое маской со случайно расположенными отверстиями размером примерно в одну букву.

При субъективной экспертизе подсчитывалось относительное количество неверно узнанных букв. После усреднения этих данных для нескольких наблюдателей оказалось возможным построить обменные соотношения между степенью дефокусировки и допустимым уровнем аддитивного шума для разных значений часто ты ошибок узнавания (рис. 6.13). Полученные данные позволяют сделать важный практический вывод о том, что потенциальные возможности линейных методов коррекции дефокусировки изображений намного превышают требования, возникающие в реальных задачах.

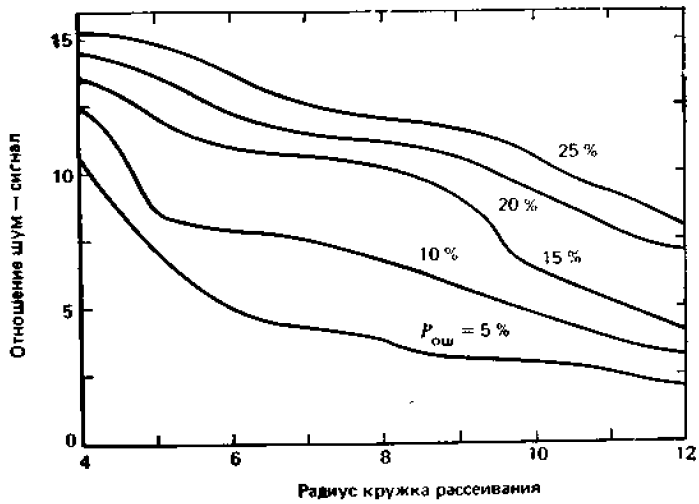


Рис. 6.13. Зависимость допустимого уровня аддитивного шума от степени геометрической дефокусировки для различной частоты ошибочного узнавания букв на скорректированном после дефокусировки изображении.
По оси абсцисс отложен радиус кружка рассеивания в элементах изображения (для сравнения укажем, что горизонтальный размер букв на тестовом изображении равнялся 5 элементам); по оси ординат – отношение среднеквадратического значения шума к среднеквадратическом у значению сигнала в процентах

Глава 7

АЛГОРИТМЫ ЛИНЕЙНОЙ ФИЛЬТРАЦИИ ДЛЯ КОРРЕКЦИИ И ПРЕПАРИРОВАНИЯ ИЗОБРАЖЕНИЙ

7.1. ПОНЯТИЕ ОБ ОПТИМАЛЬНЫХ АДАПТИВНЫХ ЛИНЕЙНЫХ ФИЛЬТРАХ

В соответствии с фундаментальными положениями теории информации и теории оптимального приема сигналов [42] сигналы можно рассматривать как элементы статистического ансамбля, определяемого ансамблями переносимых сигналами сообщений и случайных искажений и помех, действующих на сигналы, а качество обработки сигналов определяется в среднем по этим ансамблям. Для конкретизации смысла усреднения по ансамблю сигналов необходимо иметь описание сигналов как элементов статистического ансамбля.

В задачах обработки изображений сообщениями являются неизвестные (случайные) параметры изображений, определение которых и является конечной целью интерпретации изображения. Это могут быть, например, размер, форма, ориентация, относительное расположение деталей изображений, параметры, определяющие вид текстуры изображений, и т.п. Поэтому при формулировке статистического описания изображений с позиций того, какая информация должна извлекаться из изображений, следует различать два типа изображений [12]: детальные и текстурные.

Изображения относятся к детальному типу, если их можно рассматривать как совокупность объектов интерпретации и случайного фона. Объекты интерпретации – это детали изображения, случайные параметры которых, например их взаимное расположение, форма, ориентация, число, являются теми сообщениями, которые должны быть получены в результате интерпретации изображений. Фон–это остальная часть изображения, не содержащая информативных для данной задачи интерпретации параметров. В случае текстурных изображений информационными являются параметры всего изображения в целом, и изображение нельзя разбить на объекты интерпретации и фон.

Для статистического описания изображений как текстур можно использовать известные классические методы и модели теории случайных процессов и полей.

Статистическое описание изображений детального типа сложнее. Его нужно основывать на раздельном статистическом описании объектов интерпретации и фона, а также их связи. Это приводит к тому, в частности, что при нахождении среднего эффекта обработки результат обработки следует усреднять раздельно по случайным параметрам объектов интерпретации и случайному фону. Если выполнить оба этих усреднения при оптимизации процедуры обработки, найденное преобразование будет оптимальным именно в среднем. Однако, как правило, желательно, чтобы искомое преобразование было наилучшим не в среднем, а для данного обрабатываемого изображения. С точки зрения описания изображений как детально-информационных, это значит, что нужно отыскивать условно-оптимальное преобразование при фиксированной фоновой части изображения, т.е. не производить усреднение результата обработки по случайным параметрам фона. Получаемые при этом алгоритмы обработки будут зависеть от фоновой части изображения, В этом смысле они будут адаптивными. Именно такие алгоритмы рассматриваются ниже.

Одним из наиболее известных способов обработки является линейная фильтрация сигналов. Она находит широкое применение как в когерентно-оптических системах обработки информации (си., например, [13]), так и при цифровой обработке сигналов, где она основывается на использовании быстрых алгоритмов свертки и спектрального анализа, описанных в гл. 3–5. Параметры требуемых фильтров обычно находят, пользуясь принципами оптимальной (винеровской) фильтрации [10], разработанной для непрерывных сигналов и

среднеквадратичного критерия качества фильтрации. При этом фильтры являются оптимальными в среднем по всем возможным изображениям и, следовательно, не адаптивными. Ниже описаны характеристики оптимальных адаптивных фильтров для решения задач коррекции искажений и препарирования изображений. Сформулируем критерий качества фильтрации, базируясь на описанном представлении о детально-информационных изображениях. При этом будем считать, что случайная природа обрабатываемых изображений определяется тремя факторами: шумом датчика видеосигнала, случайной фоновой частью и случайными параметрами объектов интерпретации. Для простоты и определенности в качестве меры отличия результата обработки от «идеального» изображения будем использовать квадратичную меру. С целью упрощения записи воспользуемся одномерными обозначениями. Для перехода к двум переменным достаточно считать индексы двухкомпонентными.

Пусть $\mathbf{b}=\{b_k\}$, k = 0, 1, ..., N–1, – N-мерный вектор отсчетов обрабатываемого сигнала, $\mathbf{a}=\{a_k\}$ –N-мерный вектор отсчетов сигнала «идеального» изображения, $\hat{\mathbf{a}}=\{\hat{a}_k\}$ –W-мерный вектор отсчетов сигнала, полученного в результате обработки. Потребуем, чтобы значение квадратов модулей поэлементных разностей «идеального» и обработанного изображений

$$\langle\overline{[|\varepsilon|^2]}\rangle=\left\langle\left[\overline{\sum_{k=0}^{N-1}|a_k-\hat{a}_k|^2}\right]\right\rangle, \tag{7.1}$$

усредненных по ансамблю шума датчика видеосигнала (горизонтальная черта), случайному фону (квадратные скобки) и случайным параметрам объектов интерпретации (уголковые скобки) в расчете на один отсчет сигнала (суммирование по k) было минимальным. Этот критерий будем называть *критерием МСКО.*

При линейной фильтрации в общем случае над обрабатываемым сигналом $\mathbf{b}=\{b_k\}$ выполняется преобразование вида (см. гл. 3):

$$\hat{a}_k=\sum_{n=0}^{N-1}h_{k,n}b_n, \tag{7.2}$$

где $\{h_{k,n}\}$ –матрица NxN коэффициентов, определяющих линейный фильтр. Линейная фильтрация в таком виде требует N^2 операций на N отсчетов обработанного сигнала, что неэффективно в вычислительном отношении. Как уже указывалось в гл. 4.5, для сокращения количества вычислительных операций целесообразно пользоваться линейными преобразованиями, имеющими быстрые алгоритмы. Это означает, что линейная фильтрация проводится в три этапа: определение спектра обрабатываемого сигнала путем преобразования сигнала по некоторому базису, для которого существует быстрый алгоритм преобразования, поэлементное умножение отсчетов спектра на коэффициенты, определяющие скалярный линейный фильтр, и обратное преобразование по тому же базису для получения отсчетов обработанного сигнала. В матричной форме это можно записать, как

$$\hat{\mathbf{a}}=\mathbf{T}^{-1}\mathbf{H}_d\mathbf{T}\mathbf{b},$$

где T и T^{-1} – матрицы прямого и обратного преобразований, имеющих быстрый алгоритм, а H_d– диагональная матрица, описывающая скалярный линейный фильтр.

С учетом этих особенностей вычислительной реализации линейной фильтрации критерий качества фильтрации (7.1) можно выразить через отсчеты спектров $\{\hat{\alpha}_r\}$, $\{\alpha_r\}$ обработанного и «идеального» сигналов по выбранному базису:

$$\langle\overline{[|\varepsilon|^2]}\rangle=\left\langle\left[\overline{\sum_{r=0}^{N-1}|\alpha_r-\hat{\alpha}_r|^2}\right]\right\rangle \tag{7.3}$$

воспользовавшись тем, что в соответствии с соотношением Парсеваля норма вектора не зависит от выбора ортогонального базиса.Для скалярного линейного фильтра отсчеты спектра сигнала $\{\hat{\alpha}_r\}$ на выходе фильтра связаны с отсчетами спектра сигнала $\{\beta_r\}$ на входе фильтра соотношением

$$\hat{\alpha}_r=\eta_r\beta_r, \tag{7.4}$$

где $\{\eta_r\}$ – коэффициенты, описывающие скалярный фильтр. В частном случае базиса ДПФ $\{\eta_r\}$– отсчеты дискретной частотной характеристики линейного фильтра. Подставив (7.4) в (7.3), найдем

$$\langle\overline{[|\varepsilon|^2]}\rangle = \left\langle\overline{\left[\sum_{r=0}^{N-1}|\alpha_r - \eta_r\beta_r|^2\right]}\right\rangle$$

или

$$\langle\overline{[|\varepsilon|^2]}\rangle = \overline{\sum_{r=0}^{N-1}\langle[|\alpha_r - \eta_r\beta_r|^2]\rangle}, \tag{7.5}$$

так как операции усреднения и суммирования можно менять местами. Коэффициенты $\{\alpha_r\}$, $\{\beta_r\}$, $\{\eta_r\}$ – это, вообще говоря, комплексные числа. Поэтому будем отыскивать минимальное значение средней ошибки фильтрации, приравнивая нулю частные производные правой части (7.5) по вещественной η_r^{re} и мнимой η_r^{Im} частям η_r. В результате получим:

$$-\langle\overline{[\alpha_r\beta_r^*]}\rangle - \langle\overline{[\alpha_r^*\beta_r]}\rangle + \langle\overline{[2\eta_r^{re}|\beta_r|^2]}\rangle = 0;$$

$$i\langle\overline{[\alpha_r\beta_r]}\rangle - i\langle\overline{[\alpha_r^*\beta_r]}\rangle + \langle\overline{2\eta_r^{Im}|\beta_r|^2}\rangle = 0.$$

Разделив второе уравнение на (–i) и сложив его с первым, придем к следующему соотношению для η_r:

$$\langle 2\eta_r[|\beta_r|^2]\rangle - 2\langle[\alpha_r\beta_r^*]\rangle = 0,$$

из которого следует, что оптимальные значения коэффициентов скалярного линейного фильтра, минимизирующего среднеквадра-тическую ошибку фильтрации (7.5), определяются выражением:

$$\eta_{r,\,opt} = \langle\overline{[\alpha_r\beta_r^*]}\rangle / \langle\overline{[|\beta_r|^2]}\rangle$$

В эту формулу вошли все усреднения по случайным факторам, заложенные в критерий фильтрации (7.3): усреднение по случайному шуму датчика видеосигнала, по случайным параметрам объектов интерпретации и случайному фону.

Если в критерии (7.3) отказаться от усреднения по случайному фону, получим, очевидно,

$$\eta_{r,\,opt} = \langle\overline{\alpha_r\beta_r^*}\rangle / \langle\overline{|\beta_r|^2}\rangle. \tag{7.6}$$

Другим полезным критерием оптимизации параметров линейных фильтров является критерий минимума среднеквадратической ошибки восстановления энергетического спектра сигнала в заданном базисе. Будем называть этот критерий ВСС-критерием. Критерий такого рода описан в [64].

Оптимальный по этому критерию корректирующий скалярный фильтр будет определяться соотношением:

$$\eta_{r,\,opt} = (\langle|\alpha_r|^2\rangle / |\beta_r|^2)^{1/2} \tag{7.7}$$

Действие такого фильтра, очевидно, состоит в том, что он заменяет модуль спектра обрабатываемого изображения модулем спектра «идеального» изображения, усредненного по параметрам объектов интерпретации. Особенностью фильтров этого типа является то, что их характеристики не зависят от характеристик изображающей системы, влияющих на энергетический спектр видеосигнала.

Характеристики фильтров (7.6) и (7.7) зависят от спектров обрабатываемых изображений. Поэтому эти фильтры являются адаптивными. В зависимости от глубины усреднения ошибки фильтрации в критерии (7.3) фильтры будут глобально- или локально-адаптивными.

Если ошибка усредняется по всему обрабатываемому изображению, в формулы (7.6) и (7.7) войдет энергетический спектр всего изображения, и фильтры будут глобально-адаптивными. Если ошибка усредняется в пределах отдельных фрагментов изображения, в формулы войдут энергетические спектры соответствующих фрагментов, и фильтры будут локально-адаптивными. Реализовать локально-адаптивные линейные фильтры можно, воспользовавшись методом параллельной рекурсивной фильтрации, описанным в § 4.3 для случая представления импульсных реакций фильтров в базисе Фурье.

Отметим также, что фильтры типа (7.6), (7.7) могут быть реализованы в адаптивных когерентно-оптических системах пространственной фильтрации с нелинейной средой в фурье-пдоско-сти [481.

Адаптация к фоновой части изображений – это только один аспект адаптации, желательной при обработке изображений. Другой аспект – адаптация к неизвестным параметрам помех и искажений сигналов. Она предполагает автоматическое определение параметров помех и искажений непосредственно по обрабатываемому изображению (см. § 6.4).

Формулы (7.6) и (7.7) являются исходными для определения линейных фильтров, адаптирующихся к фону и помехам на изображениях в конкретных задачах обработки изображений, рассмотренных ниже.

7.2. АДАПТИВНЫЕ ЛИНЕЙНЫЕ ФИЛЬТРЫ ДЛЯ ПОДАВЛЕНИЯ АДДИТИВНОГО НЕЗАВИСИМОГО ШУМА

Аддитивный независимый шум является обычно одним из главных видов искажений изображений, возникающих в изображающих системах. Линейная фильтрация зашумленного сигнала является наиболее известным способом подавления такого шума.

В обозначениях § 7.1 действие на сигнал аддитивного независимого шума можно выразить соотношением

$$\beta_r = a_r + x_r \qquad (7.8)$$

где у.г–отсчеты спектра реализации шума по тому же базису, для которого находятся спектры неискаженного $\{\alpha_r\}$ и обрабатываемого $\{\beta_r\}$ сигналов. Подставив (7.8) в (7.6), получим, что характеристика оптимального адаптивного линейного фильтра, обеспечивающего минимум среднеквадратической в смысле (7.3) ошибки между сигналом $\hat{a}$ на выходе фильтра и неискаженным сигналом а определяется соотношением:

$$\varkappa_{r,\,opt} = (\langle\overline{|\beta_r|^2}\rangle - \overline{|x_r|^2})/\langle\overline{|\beta_r|^2}\rangle \qquad (7.9)$$

В этой формуле в числителе усреднение спектральных коэффициентов шума по параметрам объектов интерпретации на изображениях не производится, так как по условию шум не зависит от сигнала изображения. Величины $\overline{|x_r|^2}$ являются, очевидно, отсчетами спектральной плотности мощности аддитивного шума. Формула (7.9) записана в виде, предполагающем, что спектральная плотность мощности шума известна. Она может быть известна из конструктивных характеристик изображающей системы, например из величины отношения сигнал-шум в канале передачи изображений или в телевизионной системе, где эти параметры можно измерить, пользуясь специальными тестовыми сигналами. В тех случаях, когда она не известна, ее можно оценить, пользуясь методами, описанными в § 6.4.

Для того чтобы определить характеристику оптимального фильтра (7.9), необходимо также определить спектр $\langle\overline{|\beta_r|^2}\rangle$ обрабатываемого изображения (или его фрагментов при пофрагадентной обработке для локальной адаптации) по выбранному базису, усредненный по случайным параметрам объектов интерпретации, а также по реализациям шума. Усреднение по параметрам объектов интерпретации мало сказывается на энергетическом спектре наблюдаемого изображения, поскольку деление изображения на фон и объекты интерпретации имеет смысл тогда, когда размеры объектов по площади малы по сравнению с площадью всего обрабатываемого изображения. Поэтому можно приближенно считать, что $\langle\overline{|\beta_r|^2}\rangle \approx \overline{|\beta_r|^2}$, или, более точно, $\langle\overline{|\beta_r|^2}\rangle \approx$

$\approx \overline{|\beta_r|^2} + \varepsilon^2$, где ε^2 – величина, равная отношению площади объектов к площади фонового изображения. С учетом этого упрощения характеристика оптимального в смысле МСКО-критерия фильтра определится выражением

$$\varkappa_{r,\,opt} \approx (\overline{|\beta_r|^2} + \varepsilon^2 - \overline{|x_r|^2})/(\overline{|\beta_r|^2} + \varepsilon^2) \qquad (7.10)$$

Усреднение спектра $|\beta_r|^2$ по реализациям шума можно выполнить двояко: либо прямо найдя средний спектр одного изображения или его фрагмента для разных реализаций шума, либо, если это невозможно, сглаживая спектр наблюдаемого изображения, например одним из методов, описанных в § 6.3.

Одним из практически важных случаев, когда способ линейной фильтрации, вытекающий из (7.10), дает хорошие результаты, является фильтрация сильно коррелированного аддитивного шума (узкополосного шума), в энергетическом спектре которого $\{\overline{|x_r|^2}\}$ имеется только небольшое количество заметно отличных от нуля компонент, или фильтрация такого же узкополосного сигнала на фоне шума с широким спектром. Примером узкополосного шума могут служить периодические помехи, характерные для некоторых систем передачи изображений, примером фильтрации узкополосного сигнала на фоне широкополосного шума – подавление аддитивного шума на одномерных интерферограммах.

На рис. 6.9 представлен результат фильтрации периодических помех с помощью фильтра (7.10). В данном случае производилась одномерная фильтрация в базисе функций Уолша, причем в качестве оценки усредненного спектра изображения $<\overline{|\beta_r|^2}>$ использовался квадрат модуля спектра строк изображения, усредненный по всем строкам, а оценка спектра шума производилась методом предсказания, описанным в § 6.4.

Фильтр (7.10), построенный для подавления узкополосной помехи, будет пропускать без ослабления те спектральные компоненты видеосигнала, где интенсивность шума равна нулю, и значительно ослаблять те компоненты, на которых интенсивность шума велика. При большой интенсивности отдельных компонент шума $\overline{|x_r|^2}$ по сравнению с сигналом фильтр (7.10) хорошо аппроксимируется так называемым режекторным фильтром, полностью подавляющим спектральные компоненты сигнала, искаженные интенсивными компонентами шума:

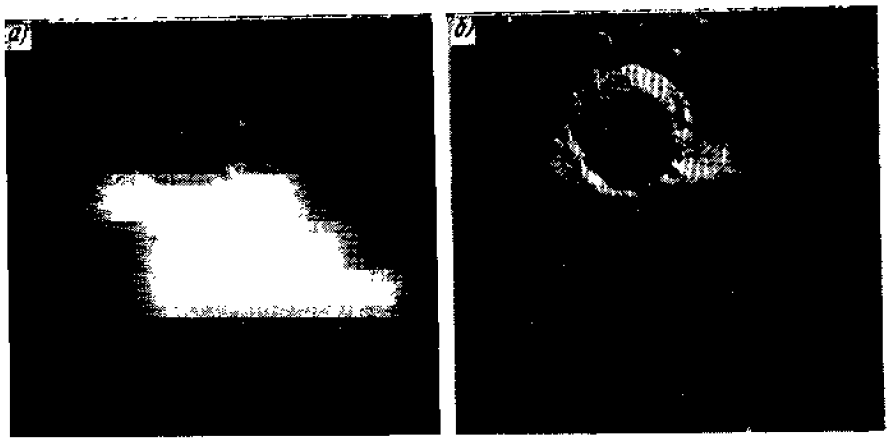


Рис. 7.1. Пример фильтрации помех типа полос и неравномерности фона

$$\eta_r = \begin{cases} 1, & \overline{|x_r|^2} = 0; \\ 0, & \overline{|x_r|^2} \approx <\overline{|\beta_r|^2}>. \end{cases}$$

Режекторный фильтр еще проще в вычислительном отношении, чем скалярный. При режекторной фильтрации спектральные компоненты сигнала с «отбракованными» на этапе обнаружения номерами r просто приравниваются нулю. Другой вариант режекторной фильтрации состоит в вычислении отдельных отмеченных при обнаружении спектральных компонент фильтруемого сигнала и вычитании их из сигнала. Если количество этих компонент невелико, такой способ может оказаться быстрее способа, предусматривающего прямое и обратное спектральное преобразование сигнала даже при использовании для выполнения последних быстрых алгоритмов.

Эксперименты с фильтрацией периодических помех на изображении показывают, что режекторная фильтрация дает результаты, визуально не отличимые от фильтрации фильтром (7.10), но при режекторной фильтрации следует более тщательно заботиться о надлежащем продолжении сигнала для борьбы с краевыми эффектами. Кроме того, наличие на изображении очень контрастных деталей с резкими границами (как, например, граница планета – космос в космической съемке, глубокие тени и т.д.) при режекторной фильтрации может привести, наоборот, к появлению периодических искажений в районе этих границ, так как для таких деталей не выполняется предположение о гладкости спектра. В этих случаях приходится прибегать к специальным мерам по удалению этих деталей из снимков перед фильтрацией и восстановлению их после фильтрации [46].

Другим примером аддитивных помех с сосредоточенным спектром, хорошо поддающихся линейной фильтрации, являются помехи типа полос и неравномерности фона [4] (рис. 7.1,а). Их дискретный спектр Фурье сосредоточен в области очень низких пространственных частот. Подавление таких помех целесообразно (с точки зрения простоты вычислений) производить обработкой не в спектральной области, а с помощью двукратной цифровой фильтрации сигнала одномерными рекурсивными фильтрами типа (3.16) по следующим формулам:

$$\hat{a}^{(1)}_{k,l} = \left[a_{k,l} - \frac{1}{2N_1+1} \sum_{m=-N_1}^{N_1} a_{k+m,l} \right] + \bar{a};$$

$$\hat{a}^{(2)}_{k,l} = \left[\hat{a}^{(1)}_{k,l} - \frac{1}{2N_2+1} \sum_{n=-N_2}^{N_2} \hat{a}^{(1)}_{k,l+n} \right] + \bar{a}, \qquad (7.11)$$

где $\bar{a}$ – константа, равная половине максимального значения видеосигнала, использовавшаяся в качестве оценки неизвестного среднего по кадру значения видеосигнала $\alpha_{0,0.}$ Каждый из одномерных фильтров подавляет полосы в соответствующем направлении. При этом параметры N_1 и N_2, сообразуясь с размерами пятна фона, выбирались в данной задаче равными 256 (в отдельных случаях 128) элементов при общих размерах кадра 1024x1024 элемента. В спектральной области эти два одномерных фильтра полностью подавляют компоненту $\alpha_{0,0}$ заменяя ее величиной, пропорциональной $\bar{a}$, и ослабляют низкочастотные компоненты спектра. Результат обработки таким фильтром снимка, показанного на рис. 7.1, о, приведен на рис. 7.1,6. Достоинством фильтра (7.11) является его высокое быстродействие при цифровой реализации. Так, для изображений размером 1024x1024 элемента выигрыш в быстродействии по сравнению с фильтрацией в частотной области с использованием алгоритма БПФ достигал примерно 10 раз. Его недостаток по сравнению с частотной фильтрацией – трудность автоматического определения параметров N_1 и N_2.

Если требуется подавлять только полосы, целесообразно применять другой метод фильтрации в пространственной области, допускающий автоматическую перестройку параметров описанными в § 6.4 способами:

$$\hat{a}_{k,l} = a_{k,l} - (\bar{a}_k - \hat{\bar{a}}_k),$$

где $\bar{a}_k$ – среднее значение сигнала вдоль строки k (направление строк совпадает с направлением полос); $\hat{\bar{a}}_k$–оценка этого среднего значения. На тех строках k, где методом предсказания или голосования обнаружен выброс,

$$\hat{\bar{a}}_k = (\bar{a}_{k-1} + \bar{a}_{k+1}/2);$$

на тех строках, где выброс не обнаружен,

$$\hat{\bar{a}}_k = \bar{a}_k.$$

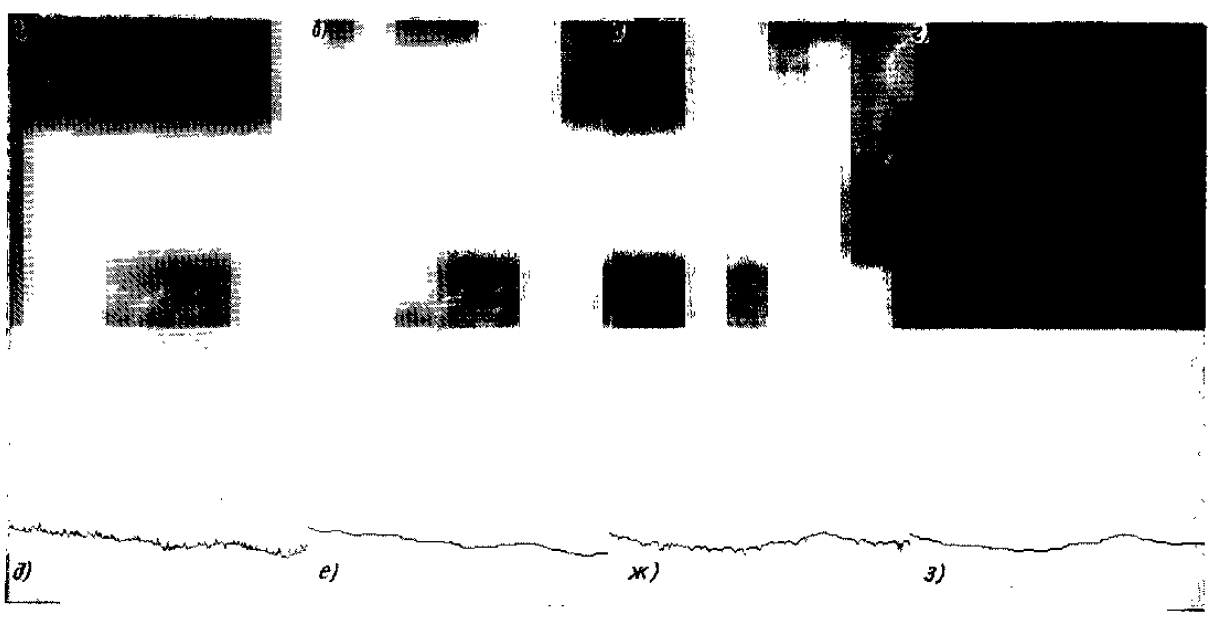


Рис. 7.2. АНП-фильтрация помех типа полос:
а — зашумленное изображение; *б* — результат фильтрации по горизонтали; *в* — то же после транспонирования; *г* — результат фильтрации по вертикали; *д*, *ж* — средние значения соответственно вдоль строк и столбцов изображения *а*; *е*, *з* — то же после обнаружения выбросов и интерполяции

Для подавления вертикальных и горизонтальных полос фильтр применяется дважды: по вертикали и по горизонтали. Результат работы фильтра показан на рис. 7.2.

7.3. КОРРЕКЦИЯ ЛИНЕЙНЫХ ИСКАЖЕНИЙ В ИЗОБРАЖАЮЩИХ И ГОЛОГРАФИЧЕСКИХ СИСТЕМАХ

Линейные искажения – это искажения, вносимые линейными блоками изображающих и голографических систем. Они определяются отличием импульсных реакций этих блоков от дельта-функции. В настоящее время развивается несколько подходов к решению задачи оптимального корректирования линейных искажений (см., например, [13, 38, 53, 64, 72]). Ниже будут рассмотрены методы коррекции, вытекающие из сформулированного в § 7.1 адаптивного подхода, особенности реализации коррекции в цифровых системах обработки изображений, а также вопрос о коррекции линейных искажений при синтезе и восстановлении голограмм.

Пусть корректируемый сигнал $\mathbf{b}=\{\beta_r\}$ может быть представлен как сумма результата действия линейного оператора А на некоторый неискаженный сигнал $\mathbf{a}=\{\alpha_r\}$ и аддитивного независимого шума $\mathbf{n}=\{\varkappa_r\}$ с ненулевым средним значением:

$$\mathbf{b} = \Lambda\mathbf{a} + \mathbf{n}.$$

Будем считать, что $\{\beta_r\}$ и $\{\alpha_r\}$ –спектры сигналов в базисе ДПФ и что линейная искажающая система задается в том же базисе диагональной матрицей $\Lambda=\{\lambda_r\}$ отсчетов своей частотной характеристики. Это соответствует модели пространственно-инвариантной линейной системы. Найдем скалярный фильтр $\mathbf{H}=\{\eta_r\}$, корректирующий искажения в такой системе и оптимальный по критерию МСКО (7.5). По-прежнему для простоты используем одномерную индексацию спектров и сигналов.

Так как в данном случае $\beta_r=\lambda_r\alpha_r+\varkappa_r$, из (7.6) получим:

$$\eta_{r,\,\mathrm{opt}} = (\lambda_r^* \langle|\alpha_r|^2\rangle + \langle\alpha_r\rangle\bar{\varkappa}_r)/\langle|\beta_r|^2\rangle$$

или, поскольку $\bar{\varkappa}_r=0$

$$\eta_{r,\,\mathrm{opt}} = \lambda_r^* \langle|\alpha_r|^2\rangle/\langle|\beta_r|^2\rangle \qquad (7.12)$$

Формулу (7.12) можно преобразовать к виду, предполагающему оценку неискаженного спектра сигнала и интенсивности шума по искаженному изображению:

$$\eta_{r,\,\mathrm{opt}} = \begin{cases} (\langle|\beta_r|^2\rangle - |\varkappa_r|^2)/\lambda_r\langle|\beta_r|^2\rangle; & \lambda_r \neq 0; \\ 0; \quad \lambda_r = 0. \end{cases}$$

В этих формулах $<|\overline{\beta_r}|^2>$ и $\overline{|x_r|^2}$ имеют тот же смысл и так же определяются, как и в формуле (7.9). Приняв во внимание высказанные соображения по поводу оценки $<|\overline{\beta_r}|^2>$ в этой формуле, получим аналогично (7.10):

$$\eta_{r,\,opt} = \begin{cases} (\overline{|\beta_r|^2} + \varepsilon^2 - \overline{|x_r|^2})/\lambda_r|\beta_r|^2; & \lambda_r \neq 0, \\ 0; & \lambda_r = 0. \end{cases}$$

Соответственно, для критерия ВСС характеристика корректирующего скалярного фильтра определяется выражением

$$\eta_{r,\,opt} = \begin{cases} ((\overline{|\beta_r|^2} + \varepsilon^2 - \overline{|x_r|^2})/|\lambda_r|^2|\beta_r|^2)^{1/2}; & \lambda_r \neq 0; \\ (<|\alpha_r|^2>/|\beta_r|^2)^{1/2}; & \lambda_r = 0, \end{cases}$$

где $<|\alpha_r|^2>$ – значения среднего по вариациям объектов интерпретации спектра неискаженного изображения для тех номеров r,для которых $\lambda_r = 0$. Оно должно быть известно априори или может быть найдено интерполяцией величины $(<|\overline{\beta_r}|^2> - \overline{|x_r|^2})/|\lambda_r|^2$ по соседним точкам, как это делалось при диагностике помех с узким спектральным составом в § 6.4.

Эксперименты со спектрами Фурье изображений показывают, что качество изображений при искажении их спектров Фурье определяется главным образом искажением фазовой части спектра, а не его модуля. Это иллюстрируется рис. 7.3, на котором пока-

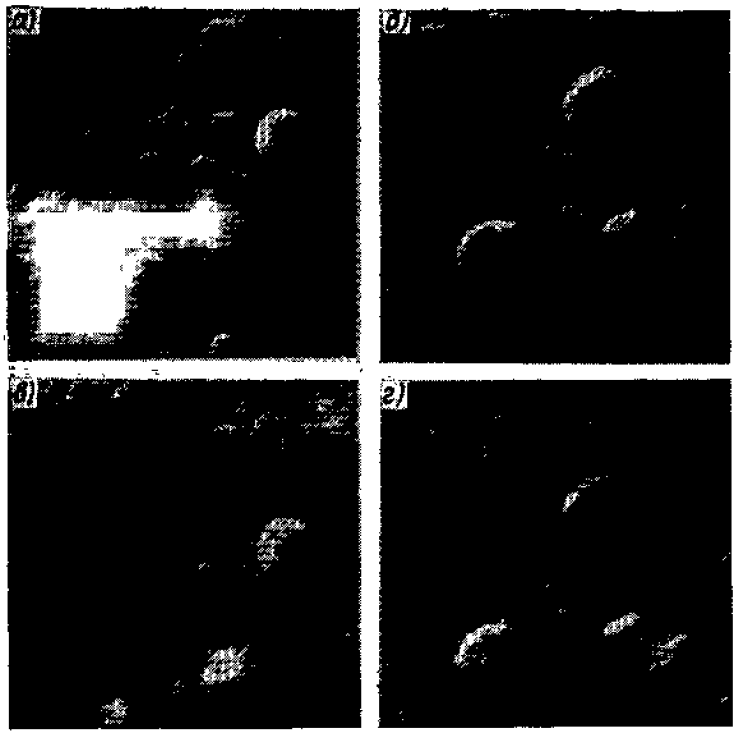


Рис. 7.3. Иллюстрация возможности грубой передачи модуля спектра Фурье изображений:
а, б – исходные изображения; в, г – результат восстановления этих изображений после обмена модулями их спектра Фурье при сохранении собственного фазового спектра

заны результаты восстановления двух разных изображений после обмена между ними модуля их спектра Фурье при сохранении собственного фазового спектра. Этот факт согласуется с тем, что в формулы для характеристик корректирующих фильтров входит квадрат модуля спектра неискаженного изображения $<|\alpha_r|^2>$,усредненный по случайным параметрам объектов интерпретации. Он также означает, что при коррекции искаженных изображений в качестве априорно задаваемого спектра неискаженного изображения $<|\alpha_r|^2>$, усредненного по параметрам объектов интерпретации, можно использовать некоторый «типовой» для данного класса изображений амплитудный спектр в базисе ДПФ. Обозначив его $|\tilde{\alpha}_r|^2$, получим для критерия ВСС следующую формулу:

$$\eta_{r,\,opt} = (|\lambda_r|/\lambda_r)\,(\overline{|\alpha_r|^2}/|\beta_r|^2)^{1/2} \qquad (7.13)$$

Следовательно, для коррекции изображения достаточно знать только фазовую характеристику искажающей системы. Из этой формулы вытекает, что если изображающая система не искажает фазы спектральных компонент изображения, то $\eta_r = (|\overline{\alpha_r}|^2/|\beta_r|^2)^{1/2}$ т.е. характеристика корректирующего фильтра не зависит от искажающей системы. Это значит, что возможна коррекция изображений при неизвестных характеристиках искажений, причем результат коррекции не зависит от характеристик искажающей системы.

Изображающие системы, не искажающие фазу спектра сигнала, составляют достаточно важный и распространенный класс систем. К ним относятся, например, системы наблюдения сквозь турбулентную атмосферу [64], системы с гауссовской апертурой, т.е. практически все системы, в которых формирование изображений производится электронным лучом. Моделирование [22] подтверждает эффективность такого способа коррекции (рис. 7.4).

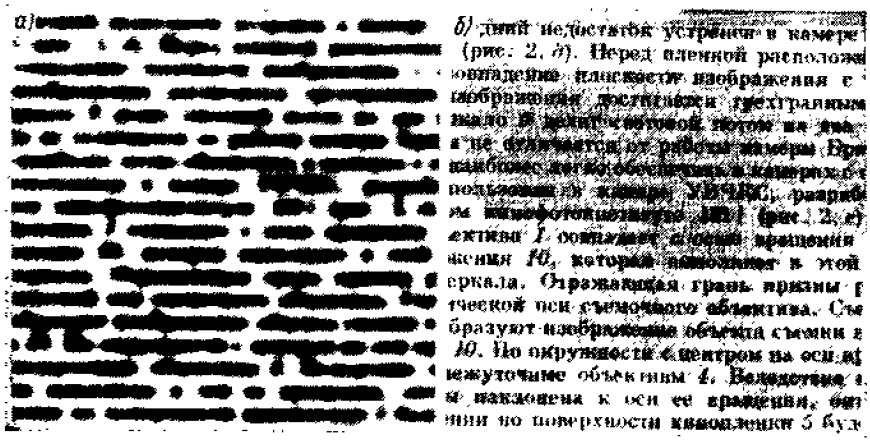


Рис. 7.4. Коррекция неизвестной гауссовской дефокусировки:
а – исходное дефокусированное изображение; б – результат коррекции фильтром (7.13)

Необходимо учитывать, что коррекция линейных искажений в изображающих системах производится обычно перед синтезом изображения. Фотографическое или другое устройство, используемое для синтеза скорректированного изображения, имеет отличную от идеальной частотную характеристику, которую необходимо учесть при коррекции. Если отсчеты частотной характеристики синтезирующего устройства обозначить $\{\mu_r\}$, то нетрудно получить оптимальный, например по критерию МОСШ, корректирующий фильтр, который будет определяться формулой

$$\eta_{r,\,\mathrm{opt}} = \begin{cases} (\langle|\beta_r|^2\rangle - \overline{|x_r|^2})/\lambda_r\mu_r < |\overline{\beta_r}|^2 >; & \lambda_r,\ \mu_r \neq 0; \\ 0, \quad \lambda_r = 0 \quad \text{или} \quad \mu_r = 0. \end{cases}$$

Возможны два способа цифровой реализации такого корректирующего фильтра: с помощью обработки дискретных спектров Фурье, используя алгоритмы БПФ, как описано в § 4.2, или с помощью цифровой фильтрации в пространственной области. В последнем случае найденная частотная характеристика используется для отыскания отсчетов импульсной реакции корректирующего цифрового фильтра. Для ослабления краевых эффектов здесь также целесообразно четное продолжение сигнала. Выбор между этими двумя способами реализации определяется требуемым объемом вычислений и емкостью запоминающих устройств. Практически •оказывается, что если корректирующий цифровой фильтр не может быть удовлетворительно аппроксимирован разделимым и рекурсивным (см. § 4.3), обработка в спектральной области с использованием алгоритмов БПФ обычно требует меньших вычислительных затрат.

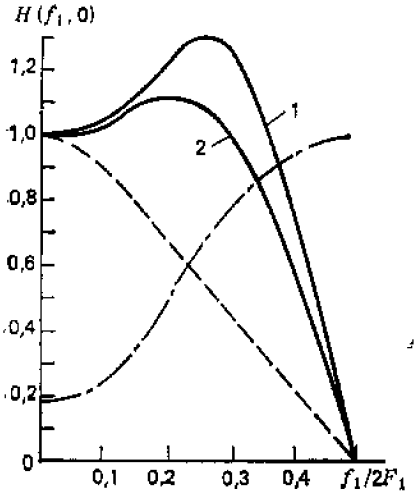


Рис. 7.5. Коррекция сквозной частотной характеристики фототелевизионной системы

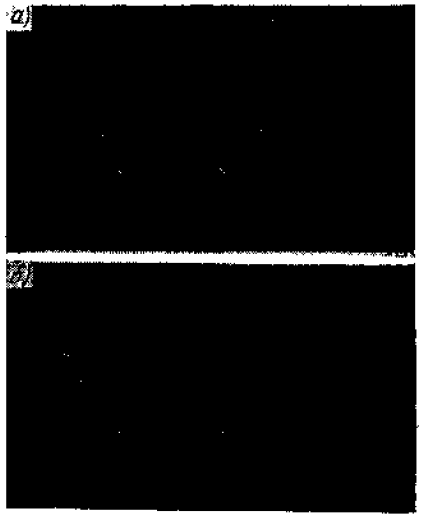


Рис. 7.6. Изображение до коррекции частотной характеристики фототелевизионной системы (*а*) и после нее (*б*)

На рис. 7.5, 7.6 приведены результаты коррекции линейных искажений с учетом характеристик фоторегистратора, используемого для синтеза скорректированного изображения [46J. Коррекция была выполнена с помощью простого разделимого рекурсивного цифрового фильтра, преобразующего отсчеты корректируемого видеосигнала $\{b_{k,l}\}$ по формуле:

$$\hat{a}_{k,l} = b_{k,l} + g\left[b_{k,l} - \frac{1}{(2N_1+1)(2N_2+1)} \sum_{m=-N_1}^{N_1} \sum_{n=-N_2}^{N_2} b_{k+m,\,l+n}\right].$$

Коэффициент усиления разностного сигнала g и размеры окрестности $(2N_1 + 1)$, $(2N_2+1)$, по которой производится усреднение, выбирались из условия аппроксимации требуемой частотной характеристики корректирующего фильтра, равной

$$\tilde{H}(f_1, f_2) = \{1 + g\,(1 - \mathrm{sincd}\,(2N_1+1;\ \pi f_1/2F_1) \times$$
$$\times\, \mathrm{sincd}\,(2N_2+1;\ \pi f_2/2F_2))\}\, H_0(f_1, f_2),$$

где $(2F_1,\ 2F_2)$ – размеры прямоугольника, ограничивающего пространственный спектр изображения, в соответствии с которым производилась дискретизация сигнала; $H_0(f_1,\ f_2)$– частотная характеристика фоторегистратора системы обработки изображений.

На рис. 7.5 штриховой линией показано сечение частотной характеристики системы, подлежащей коррекции, а штрих-пунктирной– корректирующая частотная характеристика для g=4, N_1=N2=1. Кривая 1 – частотная характеристика после коррекции без учета частотной характеристики фоторегистрирующего устройства; кривая 2 – суммарная характеристика. Видно, что в результате цифровой коррекции полоса пропускания пространственных частот по уровню 0,7 (половинной мощности) увеличилась более чем в два раза. Как это сказалось на визуально оцениваемой резкости изображения можно судить, например, по рис. 7.6, где показаны изображения до коррекции (а) и после нее (б).

Следует отметить, что использование для коррекции разделимого рекурсивного фильтра оказалось возможным благодаря тому, что искаженная характеристика системы имеет достаточно простой вид. Этот фильтр дал не вполне совершенную коррекцию. Так, на средних частотах он привел к некоторой перекоррекции. Но обработка изображений с помощью такого фильтра заняла в несколько раз меньше времени, чем потребовалось бы при обработке в спектральной области с помощью алгоритма БПФ.

Реализация коррекции линейных искажений в голографических системах имеет свои особенности. В задачах синтеза и анализа голограмм линейные искажения определяются главным образом тем, что апертуры устройств записи и считывания (измерения) голограмм и волновых полей имеют конечные размеры. Как следует из анализа процесса восстановления синтезированных голограмм, выполненного в § 10.4, конечные размеры апертуры

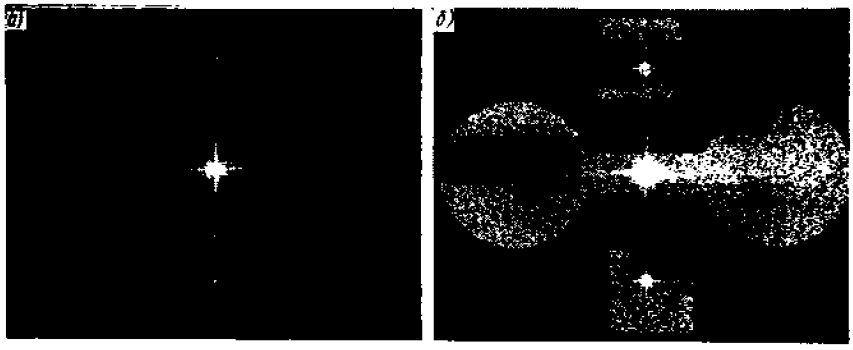


Рис. 7.7. Результат восстановления голограммы, синтезированной без коррекции затенения (*а*) и с коррекцией (*б*)

устройства записи голограмм и ограниченная разрешающая способность используемой для записи среды приводят к тому, что восстановленное изображение затеняется по полю маскирующей функцией. Эта функция пропорциональна квадрату модуля $|h(x, y)|^2$ преобразования Фурье импульсного отклика записывающего устройства (с учетом характеристик фотоматериала, используемого для записи голограмм). Скорректировать затенение можно соответствующим предыскажением исходного амплитудного распределения поля на объекте [49].

Для прямоугольной апертуры записывающего устройства с размерами $\Delta\xi, \Delta\eta$:

$$h(x, y) = \operatorname{sinc}(\pi\Delta\xi x/\lambda D)\operatorname{sinc}(\pi\Delta\eta y/\lambda D)$$

где λ –длина волны излучения, используемого для восстановления голограммы; а **D** – расстояние от точечного источника, освещающего голограмму, до плоскости наблюдения (см. § 10.1).

Поэтому, если отсчеты исходного поля обозначать *k, l* (k = 0, 1, ..., N–1; l = 0, 1, ..., N–1), то корректирующая функция, на которую необходимо умножить амплитуду распределения поля на объекте после симметрирования, должна быть следующей [см. (10.79)]:

$$c(k, l) = [\operatorname{sinc}(\pi(k/N - 1/2))\operatorname{sinc}(\pi(l/N - 1/2))]^{-2}$$

Влияние эффекта затенения и его коррекции иллюстрируется рис. 7.7, а, б.

Аналогично может производиться и коррекция конечных размеров датчиков сигнала при цифровом восстановлении голограмм и волновых полей.

7.4. ЛИНЕЙНЫЕ ФИЛЬТРЫ ДЛЯ ПРЕПАРИРОВАНИЯ ИЗОБРАЖЕНИИ

Многие методы линейной фильтрации, используемые при обработке изображений с целью препарирования, хорошо известны. Для подчеркивания мелких деталей изображений наиболее часто используется подавление нижних и усиление верхних пространственных частот спектра Фурье сигнала. Для подавления мешающих мелких деталей рекомендуется низкочастотная фильтрация – подавление верхних пространственных частот изображения [43, 55, 64, 69].

Для того чтобы обеспечить разумную основу для выбора параметров линейных фильтров, их целесообразно трактовать как оптимальную в определенном смысле линейную фильтрацию полезного сигнала на фоне помех. При этом под полезным сигналом понимаются детали изображения, которые необходимо в том или ином смысле выделить, под помехами – фоновая часть изображения, мешающая интерпретации деталей.

Следуя методике, описанной в § 7.1, найдем характеристику фильтра, минимизирующего средний квадрат модуля ошибки между сигналом выделяемого объекта (полезным сигналом) и результатом фильтрации наблюдаемого изображения. Усреднение будем производить по всем возможным вариациям сигнала полезных объектов {их положению на изображении, форме, ориентации и т.п.) и по реализациям шума датчика сигнала. Как и выше, ограничимся рассмотрением только наиболее просто реализуемых скалярных фильтров, описываемых диагональными матрицами, а наблюдаемый сигнал будем для упрощения выводов рассматривать как аддитивную смесь выделяемого объекта и фонового изображения.

Пусть $\{\alpha_r\}$, $\{\beta_r\}$, $\{\xi_r\}$, $\{\eta_r\}$ –коэффициенты представления соответственно выделяемого объекта, наблюдаемого изображения, фонового изображения и скалярного фильтра по некоторому ортонормальному базису $\{\varphi_r(k)\}$. Тогда среднее значение квадрата модуля ошибки фильтрации определяется выражением:

$$[\overline{<|\varepsilon|^2>}] = \left[\left\langle \sum_{r=0}^{N-1} |\alpha_r - \eta_r \beta_r|^2 \right\rangle\right]$$

где черта сверху, как и в § 7.1, означает усреднение по реализациям шума датчика сигнала, но в отличие от § 7.1 квадратные скобки означают усреднение по всем возможным положениям объекта на плоскости изображения, уголковые скобки – усреднение по другим случайным параметрам (форма, ориентация, масштаб и т.п.). По аналогии с (7.6) в этом случае значения η_r, минимизирующие ошибку, определяются уравнением:

$$\eta_r = [\overline{<\alpha_r \beta_r^*>}]/[\overline{<|\beta_r|^2>}] \qquad (7.14)$$

Подставив в (7.14) $\beta_r^* = \alpha_r^* + \xi_r^*$, получим

$$\eta_r = ([\overline{<|\alpha_r|^2>}] + [\overline{<\alpha_r>}]\,\xi_r^*)/[\overline{<|\beta_r|^2>}]. \qquad (7.15)$$

Так как $\alpha_r = \sum_{k=0}^{N-1} a_k \varphi_k(r)$, где $\{a_k\}$–отсчеты сигнала объекта, то

$$[\overline{<\alpha_r>}] = \sum_{k=0}^{N-1} [\overline{<a_k>}]\,\varphi_k(r) \qquad (7.16)$$

В простейшем и наиболее естественном случае, когда координаты объекта равномерно распределены по площади изображения, $[a_k]$ не зависит от ft: $[a_k] = [a]$, и (7.16) переходит в $[\overline{<\alpha_r>}] = \overline{[a]} \sum_{k=0}^{N-1} \varphi_k(r)$. Поскольку базис $\{\varphi_k(r)\}$ ортонормален, $\sum_{k=0}^{N-1} \varphi_k(r)$ отлична от нуля только для одной из базисных функций, ответственной за передачу постоянной составляющей сигнала. Поэтому второе слагаемое в числителе (7.15) сказывается только на значении г|о, влияющем на передачу несущественной постоянной составляющей по полю изображения. В дальнейшем его учитывать не будем. Кроме того, в задачах препарирования изображений, как и в задачах коррекции искажений, можно также считать, что выделяемые объекты занимают небольшую часть площади изображения, к влияние их вариаций на квадрат модуля спектра наблюдаемого сигнала можно учесть, добавив к нему небольшую константу ε^2. Таким образом, приходим окончательно к следующей формуле оптимального скалярного по МСКО-критерию фильтра

$$\eta_r = [\overline{<|\alpha_r|^2>}]/(\overline{|\beta_r|^2} + \varepsilon^2) \qquad (7.17)$$

Будем называть этот фильтр фильтром МСКО.

Если вместо условия минимума среднеквадратической ошибки фильтрации использовать критерий восстановления энергетического спектра сигнала, то придем к фильтру ВСС:

$$\eta_r = (|\overline{<|\alpha_r|^2>}]/|\beta_r|^2)^{1/2} \qquad (7.18)$$

Наконец, если задаться целью получить в результате фильтрации максимум отношения величины сигнала на искомом объекте в точке его локализации к среднеквадратическому значению сигнала на фоновом изображении, то придем к фильтру

$$\eta_r = \overline{<\alpha_r^*>}/(\overline{|\beta_r|^2} + \varepsilon^2), \qquad (7.19)$$

который можно назвать фильтром по максимуму отношения сигнал-шум (МОСШ) (см. гл. 8).

Таким образом, мы получили семейство адаптивных фильтров (7.17), (7.18) и (7.19), которые можно использовать при препарировании для выделения объектов на мешающем фоне.

Рассмотренные три типа фильтров охватывают как частные случаи упомянутые выше известные рекомендации о полезности подавления нижних пространственных частот изображений при выделении мелких деталей и подавления верхних пространственных частот при сглаживании изображений. Действительно, спектр изображений – это, как правило, функция, быстро убывающая с ростом пространственной частоты (индекса r). Таким образом, (7.17) – (7.19) – это фильтры, положение максимума пропускания которых меняется в зависимости от размера выделяемых объектов, влияющего на числители формул. Если это

малоразмерные объекты – максимум пропускания фильтра находится в области высоких пространственных частот. Если выделяются детали крупных размеров – максимум смещается в область низких пространственных частот.

Эксперименты с обработкой изображений для геологической и медицинской интерпретации [5] показывают, что эти фильтры дают значительный эффект. На рис. 7.8 показан пример фильтрации, направленной на усиление различимости микрокальцинатов на маммограммах (рентгенограммах грудной железы). На рис. 7.9 показаны примеры использования аналогичной обработки для повышения заметности кровеносных сосудов на ангиограммах (а – исходная рентгенограмма головного мозга, б – результат обработки ее фильтром, рассчитанным на выделение произвольно ориентированных сосудов). Интересный эффект псевдорельефа наблюдается на рис. 7.9, в, полученном обработкой рентгенограммы рис. 7.8, в анизотропным фильтром, рассчитанным на выделение вертикально идущих сосудов. Подобная обработка ангиограмм может явиться альтернативой введению больному при обследовании контрастирующих веществ – болезненной, а зачастую и опасной операции.

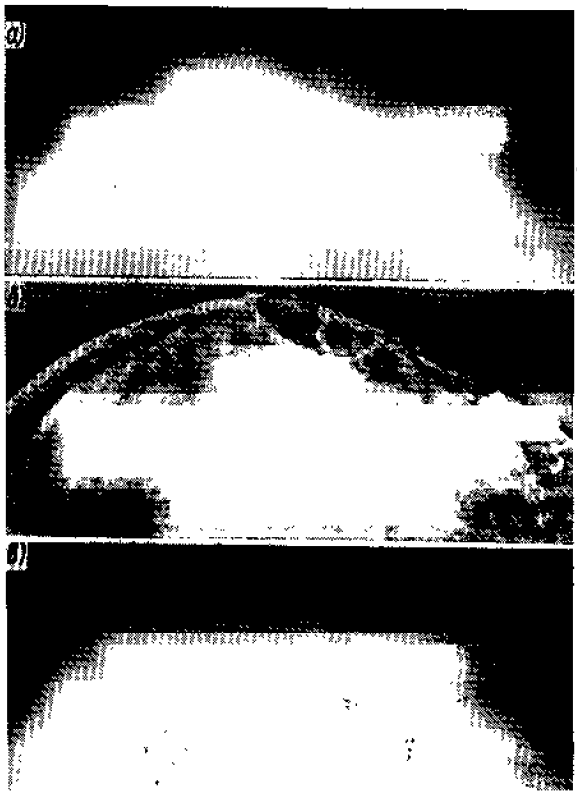


Рис. 7.8. Оптимальная фильтрация для повышения различимости микрокальцинатов на маммограммах:

а — исходная маммограмма; *б* — результат оптимальной МОСШ-фильтрации; *в* — отметка обнаруженных точек на исходном снимке

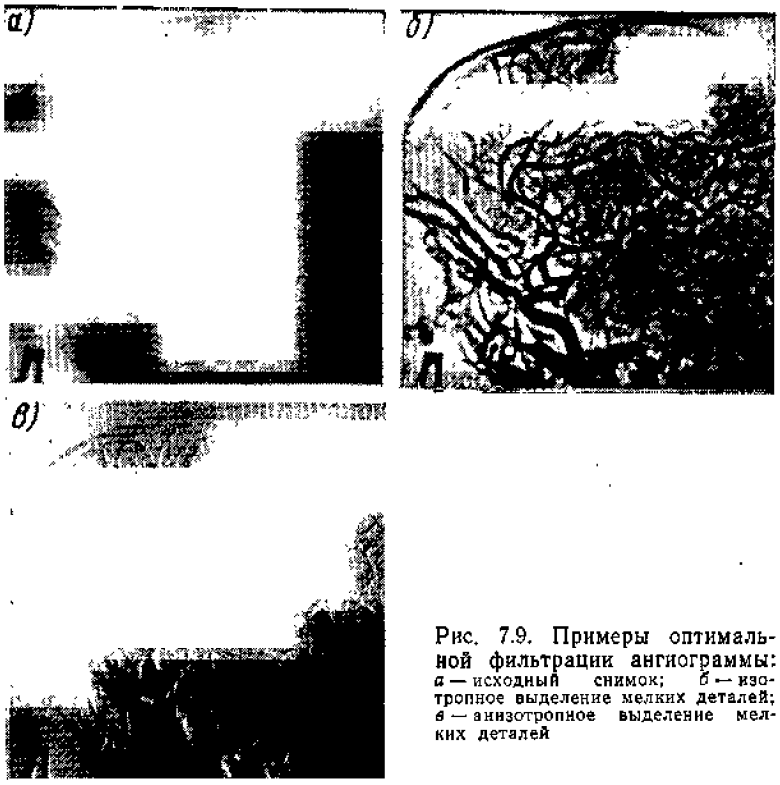


Рис. 7.9. Примеры оптимальной фильтрации ангиограммы: *а* — исходный снимок; *б* — изотропное выделение мелких деталей; *в* — анизотропное выделение мелких деталей

Рис. 7.10 иллюстрирует применение линейной фильтрации для подавления изображения ребер на флюорограммах и усиления контраста изотропных деталей средних размеров.

Важным вопросом использования линейной фильтрации для препарирования является ее вычислительная реализация, так как

Рис. 7.10. Подавление структуры ребер и усиление контраста деталей средних размеров линейной фильтрацией:
***а* — исходная рентгенограмма; *б* — результат фильтрации**

для обработки в диалоговом режиме требуется производить ее достаточно быстро. Одним из наиболее быстродействующих способов реализации оптимальной фильтрации является использование однократной или многократной – параллельной или последовательной (каскадной) фильтрации сигнала двумерным разделимым рекурсивным фильтром типа (4.21). Форма импульсной реакции этого фильтра – прямоугольная, и он, таким образом, оптимален для выделения прямоугольных по форме и ориентированных по растру объектов. Многократной параллельной фильтрацией можно сформировать произвольно ориентированную импульсную реакцию, соответствующую ориентации деталей изображения. Последовательная (каскадная), или итеративная обработка позволяет сформировать более плавную, в частности, более изотропную импульсную реакцию.

В некоторых случаях фильтрацию удобнее производить в спектральной области. Так целесообразно поступать, если необходимо подавить или усилить отдельные спектральные компоненты сигнала или узкие участки спектра сигнала (как на рис. 7.10).

Здесь важно отметить, что быстродействие существующих и потенциальных цифровых процессоров недостаточно для осуществления произвольных линейных преобразований в диалоговом режиме в реальном времени пользователя. Например, для локальной адаптации по

спектру при обработке изображения размером 1024x1024 элемента требуются порядка $K*2^{20}$ операций, где К.– коэффициент сложности, который даже для лучших рекуррентных алгоритмов не может быть меньше нескольких десятков. Так как в диалоговом режиме на обработку кадра должны затрачиваться .доли секунды, то требуемое быстродействие цифрового процессора составляет сотни миллионов операций в секунду.

Как известно, в операциях линейной пространственной фильтрации оптическая техника значительно превосходит цифровую по быстродействию. Существует простая оптическая реализация •описанных в этой главе адаптивных фильтров для коррекции и препарирования изображений в когерентно-оптических фурье-системах пространственной фильтрации [48]. Если в фурье-плоскость таких систем поместить нелинейную оптическую среду, прозрачность которой зависит от энергии падающего излучения, оптическая система превратится в адаптивную, реализующую фильтры с частотными характеристиками типа (7.17) – (7.19).

Действительно, энергия излучения в фурье-плоскости оптической фурье-системы пропорциональна квадрату модуля спектра изображения во входной плоскости системы. Следовательно, если прозрачность оптической среды обратно пропорциональна интенсивности или амплитуде падающего света, среда формирует фильтр, определяемый знаменателями формул (7.17) – (7.19).

Глава 8

АДАПТИВНЫЕ ЛИНЕЙНЫЕ ФИЛЬТРЫ ДЛЯ ЛОКАЛИЗАЦИИ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ

8.1. ПОСТАНОВКА ЗАДАЧИ

Одно из главных назначений изображений — нести информацию о взаимном пространственном расположении объектов. Обнаружение объектов, измерение их координат (локализация) является важной практической задачей, встречающейся во многих приложениях. Кроме того, она является составной частью многих других задач автоматической интерпретации изображений, в особенности задач распознавания объектов.

Локализации и обнаружению объектов на изображении посвящено много работ. Однако, по существу, все методы сводятся к вычислению функции корреляции заданного объекта с наблюдаемым изображением и последующему сравнению ее с порогом. Для обоснования этого подхода обычно привлекается аддитивная модель, согласно которой наблюдаемое изображение трактуется как сумма искомого объекта и коррелированного независимого шума с известной корреляционной функцией [13, 64]. Между тем экспериментально установлено, что на достаточно сложных изображениях корреляционный обнаружитель имеет большую вероятность ошибочного отождествления искомого объекта с посторонними объектами фона. Для того чтобы увеличить надежность обнаружения, из эвристических соображений предлагаются разного рода усовершенствования: 'квантование сигналов, пространственное дифференцирование, предыскажение формы коррелируемых объектов и т.п.

По существу, корреляционный обнаружитель-измеритель является разновидностью схемы линейного обнаружителя-измерителя, в котором решение о наличии искомого объекта и его координатах принимается в зависимости от сигнала в каждой точке поля на выходе некоторого линейного фильтра, действующего на наблюдаемое изображение. Назначение линейного фильтра в такого рода устройствах — так преобразовать пространство сигналов, чтобы затем решение можно было принимать не по всему сигналу в целом, а независимо по отдельным его координатам в преобразованном пространстве. Благодаря разбиению на независимые линейный и нелинейный безынерционные блоки значительно упрощается анализ и реализация подобного устройства в цифровых и аналоговых процессорах. Этим объясняется популярность корреляционного метода обнаружения и локализации объектов на изображениях, несмотря на его недостаток — невысокую надежность локализации. В данной главе будет показано, что, опираясь на представление о линейном обнаружителе-измерителе как совокупности линейного фильтра и нелинейного решающего блока, можно найти оптимальные характеристики линейного фильтра, обеспечивающего лучшую достоверность локализации, чем у стандартного коррелятора.

Качество измерения координат объекта определяется двумя видами ошибок. Ошибки первого рода возникают вследствие неверного отождествления искомых объектов с отдельными деталями на наблюдаемом изображении. Они дают большие отклонения результата измерения координат от истинного значения, превышающие размеры искомого объекта. Будем называть их аномальными. Аномальные ошибки характеризуются вероятностью неправильного отождествления объекта в расчете на элемент изображения. Ошибки второго рода, или нормальные ошибки, имеют порядок размеров объекта и связаны в основном только с искажениями сигнала искомого объекта шумом датчика. Можно считать, что нормальные ошибки в среднем характеризуются своим среднеквадратическим значением, и оптимальным с точки зрения минимума среднеквадратического значения нормальных ошибок является классический измеритель с согласованным фильтром. Однако он дает много аномальных ошибок.

Найдем характеристики линейного фильтра-измерителя, оптимального по отношению к аномальным ошибкам, основываясь на идеях оптимальной адаптивной линейной фильтрации, изложенных в гл. 7. Здесь в отличие от гл. 7 рассмотрим задачу синтеза непрерывного линейного фильтра с тем, чтобы полученные результаты можно было приложить также к аналоговым оптическим системам.

Определим точный смысл оптимальности. Для того чтобы учесть возможную пространственную неоднородность критерия оптимальности, будем считать, что изображение разбито на N фрагментов с площадью Sn, n = 0, 1, ..., N—1. Пусть $h_n(b, x_1^0, x_2^0)$ — распределение значений видеосигнала b (x_1, x_2) на выходе фильтра, измеренное для *n*-го фрагмента по точкам, не занятым объектом, при условии, что объект находится в точке с координатами (x_1^0, x_2^0), а b_0 — сигнал на выходе фильтра в точке локализации объекта (без ограничения общности можно считать, что $b_0>0$). Тогда, поскольку рассматриваемый линейный измеритель принимает решение о координатах искомого объекта по координатам абсолютного максимума сигнала на выходе линейного фильтра, интеграл

$$Q_n(x_1^0, x_2^0) = \int_{b_0}^{\infty} h_n(b, x_1^0, x_2^0)\, db \qquad (8.1)$$

представляет собой долю точек *n*-го фрагмента, которые могут быть ошибочно приняты решающим устройством за координаты объекта.

Величину b_0 следует рассматривать, вообще говоря, как случайную, поскольку на нее влияет шум датчика видеосигнала, условия съемки и освещения, ориентация объекта при съемке, соседние объекты и другие случайные факторы. Для того чтобы их учесть, введем функцию $q(b_0)$ — априорную плотность вероятностей значений b_0. Координаты объекта также нужно считать случайными. Кроме того, в задачах локализации вес ошибок измерения для разных участков изображения может быть неодинаков. Для учета этих факторов введем весовые функции $w_n(x_1^0, x_2^0)$ и W_n, характеризующие априорную значимость ошибок определения координат в пределах *n*-го фрагмента и для каждого n-го фрагмента соответственно и нормированные так, что

$$\iint_{S_n} w_n(x_1^0, x_2^0)\, dx_1^0 dx_2^0 = 1; \quad \sum_{n=0}^{N-1} W_n = 1$$

Тогда качество измерения координат рассматриваемым измерителем может описываться средневзвешенным по $q(b_0)$, $w_n(x_1^0, x_2^0)$ и W_n значением интеграла (8.1):

$$\overline{Q} = \int_{-\infty}^{\infty} q(b_0)\, db_0 \sum_{n=0}^{N-1} W_n \iint_{S_n} w^{(n)}(x_1^0, x_2^0)\, dx_1^0 dx_2^0 \times$$
$$\times \int_{b_0}^{\infty} h^{(n)}(b, x_1^0, x_2^0)\, db. \qquad (8.2)$$

Если нас интересует качество работы измерителя в среднем по некоторому набору изображений, то величину Q нужно усреднить по этому набору. Оптимальным будем считать измеритель, обеспечивающий минимум $\overline{Q}$

8.2. ЛОКАЛИЗАЦИЯ ТОЧНО ИЗВЕСТНОГО ОБЪЕКТА ПРИ ПРОСТРАНСТВЕННО-ОДНОРОДНОМ КРИТЕРИИ ОПТИМАЛЬНОСТИ

Предположим, что объект задан точно. Это означает, что отклик на этот объект любого фильтра точно известен, т.е. $q(b_0) = \delta(b_0 - \overline{b_0})$, где $\overline{b_0}$ —известное значение отклика. Тогда (8.2), определяющее критерий качества локализации, перейдет в

$$\overline{Q} = \sum_{n=0}^{N-1} W_n \int_{\overline{b_0}}^{\infty} \overline{h}_n(b)\, db,$$

$$\overline{h}_n(b) = \iint_{S_n} w^{(n)}(x_1^0,\ x_2^0)\, h_n(b,\ x_1^0,\ x_2^0)\, dx_1 dx_2.$$

где

Предположим, что критерий оптимальности является пространственно однородным, т.е. веса W_n не зависят от n и равны $1/N$.

$$\overline{h}(b) = \left(\sum_{n=0}^{N-1} \overline{h}_n(b) \right) / N$$

Пусть — распределение значений сигнала на выходе фильтра, измеренное для всего изображения и усредненное по неизвестным координатам объекта, так что

$$\overline{Q} = \int_{\overline{b}_0}^{\infty} \overline{h}(b)\, db.$$

Найдем сначала частотную характеристику $H(f_1,f_2)$ фильтра, минимизирующего значение $\overline{Q}$. Выбор $H(f_1, f_2)$ влияет как на величину b_0, так и на распределение $\overline{h}(b)$. Поскольку b_0 — значение отклика фильтра в месте локализации объекта, его можно найти, зная спектр объекта $\alpha_0(f_1, f_2)$ как

$$\overline{b}_0 = \int_{-\infty}^{\infty} \alpha_0(f_1,\ f_2) H(f_1,\ f_2)\, df_1 df_2.$$

Связь $h(b)$ и $H(f_1, f_2)$, вообще говоря, имеет сложный характер. Явную зависимость от $H(f_1, f_2)$ можно выписать только для второго момента распределения $\overline{h}(b)$, воспользовавшись соотношением Парсеваля для преобразования Фурье:

$$m_2^2 = \int_{-\infty}^{\infty} b^2 \overline{h}(b)\, db = \iint_S w(x_1^0,\ x_2^0)\, dx_1^0 dx_2^0 \int_{-\infty}^{\infty} b^2 h(b,\ x_1^0,\ x_2^0) db =$$

$$= \frac{1}{S_1} \iint_S w(x_1^0,\ x_2^0)\, dx_1^0 dx_2^0 \int_{S_1} b^2(x_1,\ x_2)\, dx_1 dx_2 =$$

$$= \frac{1}{S_1} \iint_S w(x_1^0,\ x_2^0)\, dx_1^0 dx_2^0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\alpha_\Phi(f_1,\ f_2)|^2 \times$$

$$\times |H(f_1,f_2)|^2 df_1 df_2 = \frac{1}{S_1} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{|\alpha_\Phi(f_1, f_2)|^2} |H(f_1,\ f_2)|^2 df_1 df_2,$$

где S_1—площадь анализируемого изображения за вычетом площади, занимаемой сигналом от искомого объекта на выходе фильтра: $\alpha_\Phi(f_1,\ f_2)$ — спектр Фурье изображения, в котором значения сигнала на участке, занятом искомым объектом, заменены нулевыми значениями {спектр фоновой части изображения), а

$$\overline{|\alpha_\Phi(f_1,\ f_2)|^2} = \iint_{S_1} w(x_1^0,\ x_2^0)\, |\alpha_\Phi(f_1,\ f_2)|^2 dx_1^0 dx_2^0. \qquad (8.3)$$

Поэтому будем опираться на известное в теории вероятностей неравенство Чебышева:

$$\overline{Q} = \int_{\overline{b}_0}^{\infty} \overline{h}(b)\, db \leqslant m_2^2 / \overline{b}_0^2.$$

и потребуем, чтобы отношение $g = m_2^2/\overline{b}_0^2$ было минимальным. Строго говоря, неравенство Чебышева — слишком грубая оценка $\overline{Q}$. Поэтому требование минимума g является необходимым и достаточным условием минимума $\overline{Q}$, только если $h(\overline{b})$ есть нормальная (гауссовская) плотность распределения. Однако, как показывают эксперименты на реальных изображениях, распределение значений $\overline{h}(b)$ сигнала на выходе фильтра, найденного из условия минимума g, близка к гауссовской. Это результат нормализующего действия линейного фильтра. Условие минимума g эквивалентно условию максимума величины

$$G = \frac{b_0^2}{S_1 m_2^2} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \alpha_0(f_1,\ f_2) H(f_1,\ f_2)\, df_1 df_2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{|\alpha_\Phi(f_1,\ f_2)|^2} |H(f_1,\ f_2)|^2 df_1 df_2}.$$

Для отыскания минимума G по $H(f_1, f_2)$ воспользуемся неравенством Шварца:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\alpha_0(f_1, f_2)/|\overline{\alpha_\Phi(f_1, f_2)}|)(H(f_1, f_2)|\overline{\alpha_\Phi(f_1, f_2)}|)\,df_1df_2 \leqslant$$

$$\leqslant\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{|\alpha_0(f_1, f_2)|^2}{|\alpha_\Phi(f_1, f_2)|^2}df_1df_2\int_{\infty}^{\infty}\int_{\infty}^{\infty}|H(f_1, f_2)|^2\times\right.$$

$$\left.\times|\overline{\alpha_\Phi(f_1, f_2)}|^2df_1df_2\right)^{1/2},$$

из которого вытекает, что максимальное значение G

$$G_{\max}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\alpha_0(f_1, f_2)|^2/|\overline{\alpha_\Phi(f_1, f_2)}|^2\,df_1df_2 \tag{8.4}$$

достигается при

$$H_{\text{opt}}(f_1, f_2)=\alpha_0^*(f_1, f_2)/|\overline{\alpha_\Phi(f_1, f_2)}|^2. \tag{8.5}$$

Выразим $\overline{|\alpha_\Phi(f_1, f_2)|^2}$ через спектр наблюдаемого изображения $\alpha_и(f_1, f_2)$ и спектр искомого объекта $\alpha_0(f_1, f_2)$. Очевидно,

$$\alpha_\Phi(f_1, f_2)=\alpha_и(f_1, f_2)-\alpha_0(f_1, f_2)\exp[-i2\pi(f_1x_1^0+f_2x_2^0)]$$

Тогда, подставив это выражение в (8.3), получим:

$$|\overline{\alpha_\Phi(f_1, f_2)}|^2=\overline{|\alpha_и(f_1, f_2)-\alpha_0(f_1, f_2)\exp[2\pi(f_1x_1^0+f_2x_2^0)]|^2}=$$

$$=\overline{|\alpha_и(f_1, f_2)|^2}+|\alpha_0(f_1, f_2)|^2-\alpha_и(f_1, f_2)\alpha_0^*(f_1, f_2)\times$$

$$\times W(f_1, f_2)-\alpha_и^*(f_1, f_2)\alpha_0(f_1, f_2)W^*(f_1, f_2),$$

где $W(f_1, f_2)$ — спектр Фурье весовой функции $w(x_1^0, x_2^0)$

Наибольший практический интерес представляет случай, когда весовая функция $w(x_1^0, x_2^0)$ примерно постоянна по площади изображения. Тогда ее спектр $W(f_1, f_2)$ является функцией, сосредоточенной в нуле (в дискретном случае, если весовая функция постоянна по площади обрабатываемого изображения, ее спектр есть .дельта-функиня). Поэтому последние два слагаемых в этой формуле (перекрестные спектры) заметно сказываются только на значении $\overline{|\alpha_\Phi(f_1, f_2)|^2}$ при $f_1=f_2=0$, т.е. влияют только на несущественное среднее (по площади изображения) значение сигнала на выходе фильтра. Отсюда следует, что усредненный спектр фоновой части изображения можно приближенно заменить усредненным спектром всего обрабатываемого изображения с некоторой поправкой ε^2, учитывающей спектр искомого объекта и перекрестные спектры:

$$\overline{|\alpha_\Phi(f_1, f_2)|^2}\approx\overline{|\alpha_и(f_1, f_2)|^2}+\varepsilon^2.$$

Эту поправку в первом приближении можно оценить, как $\varepsilon^2\approx S_0\alpha_{\text{скв}}^2$, где S_0—площадь, занимаемая объектом; $\alpha_{\text{скв}}^2$—среднеквадратическое значение сигнала объекта:

$$\alpha^2_{\text{скв}} = \frac{1}{S_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\alpha_0(f_1, f_2)|^2 df_1 df_2.$$

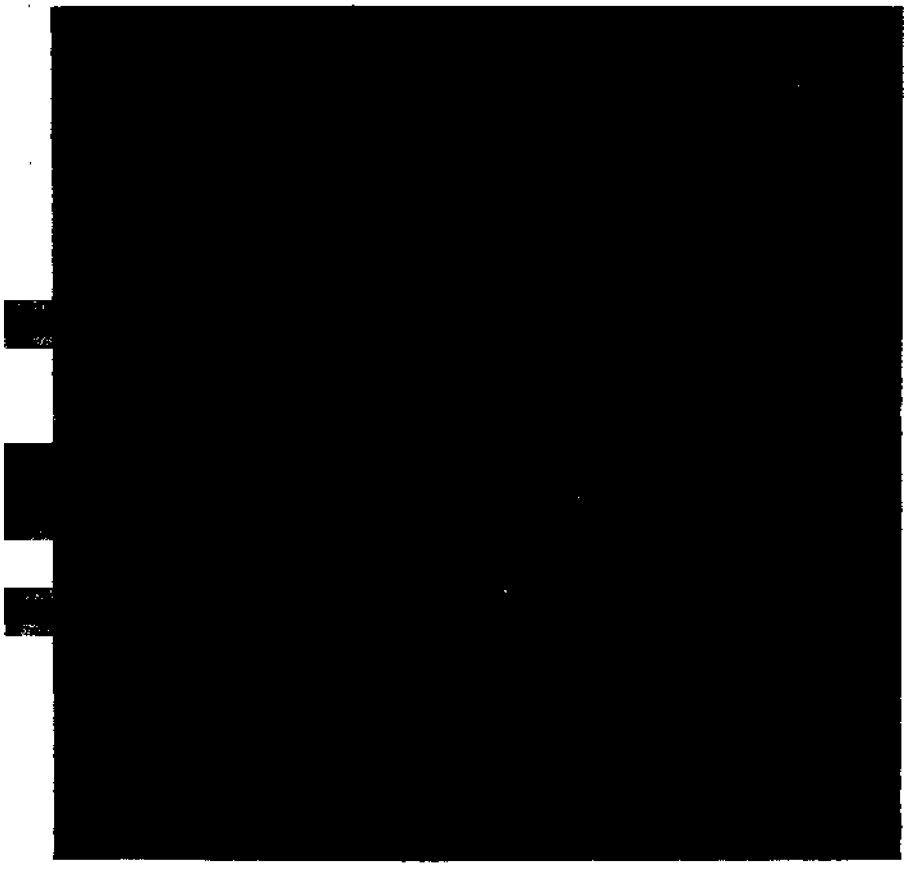


Рис. 8.1. Тестовый аэрофотоснимок с нанесенными квадратными метками

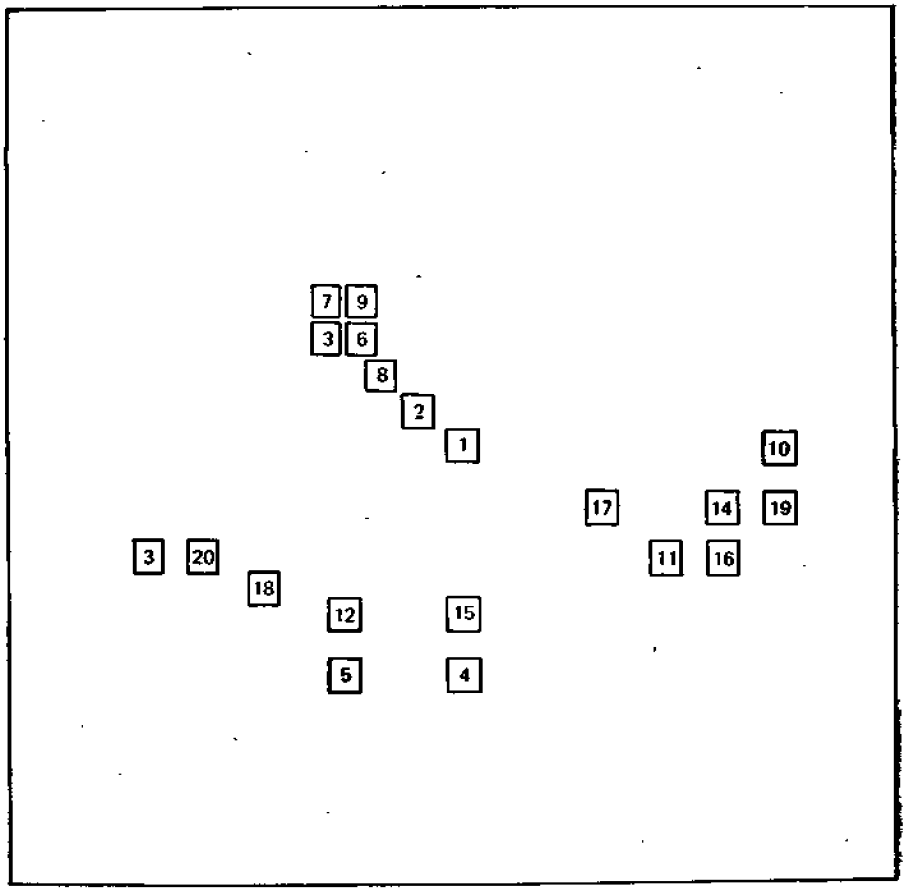


Рис. 8.2. Схема расположения меток на рис. 8.1.

Задача локализации объекта с минимальной вероятностью аномальных ошибок имеет смысл, когда площадь объекта мала по сравнению с площадью всего изображения. Поэтому поправка ε^2 мала: ее отношение к среднему квадрату модуля спектра изображения по порядку величины равно отношению площади объекта к площади изображения. С учетом этого получим следующую формулу для синтеза оптимального фильтра:

$$H_{\text{opt}}(f_1, f_2) = \alpha_0^*(f_1, f_2)/(|\overline{\alpha_{\text{и}}(f_1, f_2)}|^2 + \varepsilon^2) \qquad (8.6)$$

Очевидно, если требуется построить оптимальный фильтр для набора изображений, то в эту формулу следует подставлять вместо $\overline{\alpha_{\text{и}}|f_1, f_2|^2}$ результат усреднения спектров изображений по заданному набору.

Найденный оптимальный фильтр может быть сравнительно просто реализован оптическими средствами в адаптивной оптической системе с нелинейным элементом в фурье-плоскости [48]. При реализации его в цифровых процессорах естественней всего использовать обработку сигнала в частотной области, поскольку частотная характеристика (8.6) оптимального фильтра основывается на измерении спектра наблюдаемого изображения.

Результаты экспериментов по моделированию на цифровой ЭВМ описанного оптимального линейного измерителя, подтверждающие его преимущества по сравнению с традиционным корреля-

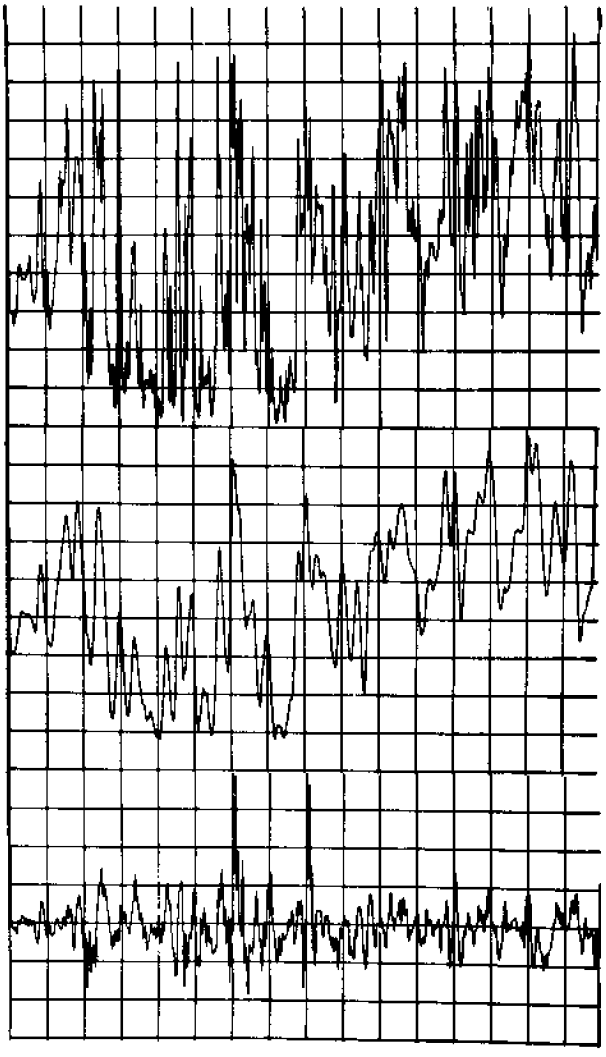


Рис. 8.3. График, сечения видеосигнала рис. 8.1, сигнала на выходе стандартного коррелятора и сигнала на выходе оптимального фильтра

тором, показаны на рис. 8.1—8.4. На рис. 8.1 приведено изображение, с которым проводились эксперименты по измерению координат наложенных на него 20 тестовых квадратных темных меток с линейными размерами в одну сотую размеров кадра изображения. Схема расположения меток показана пронумерованными квадратами на рис. 8.2. Как видно из этой схемы, тестовые объекты нанесены на различных по структуре участках аэрофотоснимков,, что позволяет

оценить работу коррелятора и оптимального линейного измерителя в разных условиях. Контраст меток составлял примерно 25% от размаха видеосигнала на аэрофотоснимке. Отношение амплитуды метки к среднеквадратическому значению видеосигнала на фоновом изображении — около 1,5. На рис. 8.3 показаны (сверху вниз) графики сечений исходного видеосигнала, выходного сигнала стандартного коррелятора и оптимального фильтра, проходящие через центры меток 12 и 15. На графике выходного сигнала коррелятора хорошо видны автокорреляционные пики тестовых меток и ложные корреляционные пики, в том числе превышающие автокорреляционный. Эти ложные пики дают ложные решения (рис. 8.4). Сравнивая график сигнала коррелятора с нижним графиком на рис. 8.3, можно оценить, насколько оптимальный фильтр облегчает решающему устройству локализацию.

Рис. 8.4. Схема решений на выходе стандартного коррелятора

8.3. УЧЕТ НЕОПРЕДЕЛЕННОСТИ В ЗАДАНИИ ОБЪЕКТА И ПРОСТРАНСТВЕННОЙ НЕОДНОРОДНОСТИ КРИТЕРИЯ. ЛОКАЛИЗАЦИЯ НА «СМАЗАННЫХ» ИЗОБРАЖЕНИЯХ. ХАРАКТЕРИСТИКИ ОБНАРУЖЕНИЯ

Локализация неточно заданного изображения. В этом случае $q(b_0)$ нельзя считать дельта-функцией. Критерий будем считать по-прежнему пространственно однородным, и потребуем, чтобы оптимальный измеритель обеспечивал минимум интеграла

$$\overline{Q}_1 = \int_{-\infty}^{\infty} q(b_0)\, db_0 \int_{b_0}^{\infty} \overline{h}(b)\, db. \qquad (8.7)$$

Возможны два варианта измерителя.

Измеритель с перебором. Разобьем интервал возможных значений b_0 на подынтервалы, в пределах которых $q(b_0)$ можно считать постоянной. Тогда

$$\overline{Q}_1 \approx \sum_i q_i \int_{b_0^{(i)}}^{\infty} h(b)\, db,$$

где $b_0^{(i)}$ —представитель i-го интервала; q_i —площадь под $q(b_0)$ на 1-м интервале. Поскольку $q_i \geqslant 0$, Q_1 минимально, если минимальны

$$\overline{Q}_1^{(i)} = \int_{b_0^{(i)}}^{\infty} h(b)\,db.$$

Таким образом, задача свелась к предыдущей задаче локализации точно известного объекта. Разница только в том, что теперь нужно строить измеритель с фильтром

$$H_{\text{opt}}^{(i)}(f_1, f_2) = \alpha_0^{*(i)}(f_1, f_2)/(|\overline{\alpha_{\text{н}}(f_1, f_2)}|^2 + \varepsilon^2) \qquad (8.8)$$

отдельно для каждого «представителя» объекта из всех возможных его вариаций, т.е. считать, что имеется не один заданный объект, а несколько, отличающихся друг от друга значениями неизвестных параметров. Это, конечно, приводит к потерям времени обработки на перебор.

Измеритель, настроенный на усредненный объект. Если разброс параметров невелик, можно ценой некоторого увеличения частоты аномальных ошибок решать задачу так, как если бы объект был известен точно, скорректировав только оптимальный фильтр с учетом разброса параметров объекта. Чтобы найти скорректированную характеристику фильтра, сделаем в (8.7) замену переменных $b_1 = b - b_0$ и изменим порядок интегрирования:

$$\overline{Q}_1 = \int_0^{\infty} db_1 \int_{-\infty}^{\infty} q(b_0)\,\overline{h}(b_1 + b_0)\,db_0.$$

Внутренний интеграл здесь представляет собой свертку двух распределений, или распределение разности двух независимых величин b и b_0 Обозначим это распределение $\overline{h}_q(b_1)$. Его среднее значение равно разности средних значений b_0 и b_{cp} распределений $q(b_0)$ и $\overline{h}(b)$, а дисперсия — сумме дисперсий этих распределений, т.е. $[m_2^2 - (b_{\text{ср}})^2] + \sigma_q^2$, где σ_q^2—дисперсия распределения $q(b_0)$. Поэтому

$$\overline{Q}_1 = \int_0^{\infty} h_q(b_1)\,db_1 = \int_{b_0}^{\infty} h_q(b_1 - \overline{b}_0)\,db_1.$$

Тем самым задача свелась к рассмотренной в § 8.2, и по аналогии с (8.6) можно записать следующее выражение для частотной характеристики оптимального фильтра:

$$\overline{H}_{\text{opt}}(f_1, f_2) = <\alpha_0^*(f_1, f_2)> / (|\alpha_{\text{н}}(f_1, f_2)|^2 + |\alpha_{\text{эф}}(f_1, f_2)|^2 + \varepsilon^2), \qquad (8.9)$$

где $<\alpha_0^*(f_1, f_2)>$ — функция, комплексно-сопряженная со средним по множеству значений неизвестных параметров объекта его спектром $<\alpha_0(f_1, f_2)>$ (результат усреднения по $q(b_0)$), а $|\alpha_{\text{эф}}(f_1, f_2)|^2 = <|\alpha_0(f_1, f_2) - <\alpha_0(f_1, f_2)>|^2>$ — такое же среднее квадрата разности $\alpha_0(f_1, f_2) - <\alpha_0(f_1, f_2)>$. Отсюда видно, что оптимальный фильтр несколько видоизменяется по сравнению с детерминированным случаем, когда объект точно известен: он строится на основе «усредненного» объекта и энергетического спектра фонового изображения, скорректированного на среднеквадратическое значение энергетического спектра.

Локализация в случае пространственно неоднородного критерия. Обратимся к общей формуле (8.2). В зависимости от реализационных ограничений в этом общем случае также можно выбрать одну из двух возможностей достижения минимума $\overline{Q}$.

Перестраиваемый измеритель с пофрагментной оптимальной фильтрацией. При заданных W_n минимум Q достигается при минимуме всех

$$\overline{Q}_2^{(n)} = \int_{-\infty}^{A} q(b_0)\,db_0 \iint_{S_n} w^{(n)}(x_1, x_2)\,dx_1 dx_2 \int_{b_0}^{\infty} h_n(b, x_1^0, x_2^0)\,db. \qquad (8.10)$$

Это означает, что линейный фильтр, преобразующий изображение, должен быть перестраиваемым и производить обработку изображения по фрагментам, в пределах которых производится усреднение в (8.10). Для каждого фрагмента оптимальная характеристика фильтра находится по (8.6) на основе измерения локального наблюдаемого энергетического спектра фрагментов. В соответствии с критерием (8.8) переход от фрагмента к фрагменту

происходит без перекрытия. Но если имеется возможность осуществить перестраиваемый измеритель, пофрагментный алгоритм естественно обобщается на скользящий алгоритм обработки, основанный на оценке текущего локального энергетического спектра изображения. При пофрагментной обработке и скользящей обработке перестраиваемым фильтром характеристика фильтра не зависит от весов W_n или непрерывной весовой функции, соответствующей случаю скользящей обработки.

Неперестраиваемый измеритель. Если перестраиваемый измеритель с пофрагментной или скользящей обработкой реализовать невозможно, измеритель должен настраиваться на усредненный по W_n энергетический спектр фрагментов изображения. Действительно, из (8.2) вытекает, что

$$Q=\int_{-\infty}^{\infty} q(b_0)\,db_0 \int_{b_0}^{\infty}\left(\sum_{n=0}^{N-1} W_n \iint_{S_n} w^{(n)}(x_1^0, x_2^0)\,dx_1^0 dx_2^0 h_n(b, x_1, x_2)\,db\right),$$

откуда по аналогии с (8.6) и (8.9) можно заключить, что

$$H_{\text{opt}}(f_1, f_2)=\langle \alpha_0^*(f_1, f_2)\rangle / (|\overline{\alpha_n(f_1, f_2)}|^2 + |\alpha_{\text{эф}}(f_1, f_2)|^2+\varepsilon^2), \qquad (8.11)$$

$$\overline{|\alpha_n(f_1, f_2)|^2}=\sum_{n=0}^{N-1} W_n \overline{|\alpha_n^{(n)}(f_1, f_2)|^2},$$

где

Таким образом, в этом случае передаточная характеристика оптимального фильтра зависит от весов $\{W_n\}$

Локализация на дефокусированных изображениях. Пусть изображение искажено линейной пространственно-инвариантной системой с частотной характеристикой $Hs(f_1,f_2)$. Оптимальный измеритель (для простоты будем исходить из формулы (8.6) для точно известного объекта и пространственно однородного критерия), очевидно, должен настраиваться на объект, прошедший то же преобразование, что и наблюдаемое изображение, т.е. передаточная характеристика фильтра должна определяться соотношением:

В зависимости $H_{\text{opt}}(f_1, f_2)=H_S^*(f_1, f_2)\alpha_0^*(f_1, f_2)/(|\overline{\alpha_n(f_1, f_2)}|^2+\varepsilon^2)$. от того, как удобно реализовать этот фильтр и в каком виде задан эталонный объект, возможны различные модификации этой формулы. Например, представление $H_{\text{opt}}(f_1, f_2)$ в виде:

$$H_{\text{opt}}(f_1, f_2)=H_1H_2=(H_S^*/|H_S|(|\overline{\alpha_n}|^2+\varepsilon^2)^{1/2})\times$$
$$\times(\alpha_0^*|H_S|/(|\overline{\alpha_n}|^2+\varepsilon^2)^{1/2})$$

соответствует измерителю, в котором наблюдаемое дефокусированное изображение спектра подвергается «разбеливанию» фильтром H_1—операции, делающей его энергетический спектр почти равномерным — и затем корректированию со скорректированным эталоном (фильтр H_2). Отношение $(|\overline{\alpha_n}|^2+\varepsilon^2)^{1/2}/|H_S|$ можно рассматривать как спектр изображения на выходе фильтра, обратного расфокусирующему, т.е. как спектр изображения, скорректированного обратным фильтром. Здесь прослеживается связь между задачами локализации на дефокусированных изображениях и коррекции искаженных линейной системой изображений (см. §7.3).

Характеристики обнаружения. Иногда требуется с определенной достоверностью обнаружить объект, не задаваясь априори, что он присутствует на изображении. Достоверность обнаружения характеризуется, как известно, условными вероятностями пропуска объекта и ложной тревоги (ложного обнаружения). Особенность рассматриваемой нами задачи локализации и обнаружения на изображении в том, что возможность пропуска объекта и ложной тревоги определяется разными случайными факторами: первая— наличием шума датчика сигнала, вторая — наличием посторонних объектов и (в меньшей степени) шумом датчика сигнала.

Вероятность ложной тревоги можно определить, если посторонние объекты заданы статистически, например в виде закона распределения выбросов сигнала на выходе оптимального фильтра. Описанный выше выбор линейного фильтра гарантирует только, что она минимальна для наблюдаемого набора посторонних объектов.

Шум датчика видеосигнала вполне удовлетворительно описывается аддитивной гауссовской моделью. Поэтому вероятность пропуска объекта можно найти по формуле

$$P_{\text{пр}} = \Phi((b_{\text{п}} - b_0)/\sigma_{\text{ш}})$$

где b_0 — максимальное значение сигнала искомого объекта на выходе оптимального фильтра; b_n — выбранное значение порога обнаружения; $\sigma_{\text{ш}}$ — стандартное отклонение шума датчика; $\Phi(x)$ — интеграл ошибок:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt.$$

8.4. ОПТИМАЛЬНАЯ ЛОКАЛИЗАЦИЯ И КОНТУРЫ ИЗОБРАЖЕНИЙ. ВЫБОР ОБЪЕКТОВ С ТОЧКИ ЗРЕНИЯ НАДЕЖНОСТИ ЛОКАЛИЗАЦИИ. ИЗБЫТОЧНОСТЬ СТЕРЕОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЙ С ТОЧКИ ЗРЕНИЯ ЗАДАЧИ ЛОКАЛИЗАЦИИ ОБЪЕКТОВ

«Разбеливание» и контуры. Для того чтобы понять смысл операций над наблюдаемым изображением, которые выполняются найденным оптимальным линейным фильтром, удобно представить характеристику (8.6) этого фильтра в виде:

$$H_{\text{opt}}(f_1, f_2) = (1/(|\overline{\alpha_{\text{и}}}|^2 + \varepsilon^2)^{1/2})(\alpha_0^*(f_1, f_2)/(|\overline{\alpha_{\text{и}}}|^2 + \varepsilon^2)^{1/2}).$$

В этом представлении действие фильтра сводится к уже упоминавшемуся в § 8.3 разбеливанию изображения и последующему коррелированию разбеленного изображения с точно так же преобразованным искомым объектом.

Интересной особенностью оптимального фильтра является то, что действие разбеливающей компоненты приводит обычно к оконтуриванию наблюдаемого изображения за счет усиления его верхних пространственных частот. Действительно, энергетический спектр изображений является, как правило, достаточно быстро убывающей функцией пространственных частот и, следовательно, функция $(|\overline{\alpha}|^2 + \varepsilon^2)^{1/2} / |Hs|$ является возрастающей функцией частот.

Этот вывод иллюстрируется рис. 8.5, на котором представлено разбеленное изображение, показанное на рис. 8.1, а также результатами разбеливания тестового изображения, представленными на рис. 8.6. Тем самым получает рациональное объяснение известная эмпирическая рекомендация, что для более надежной локализации

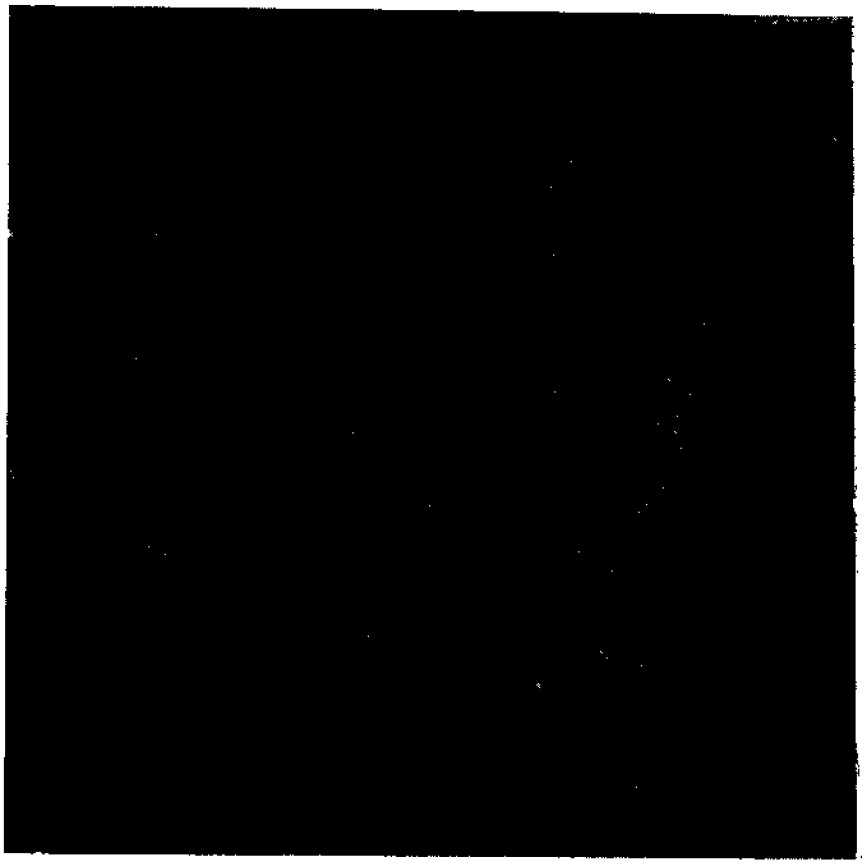


Рис. 8.5. Результат «разбеливания» изображения на рис. 8.1

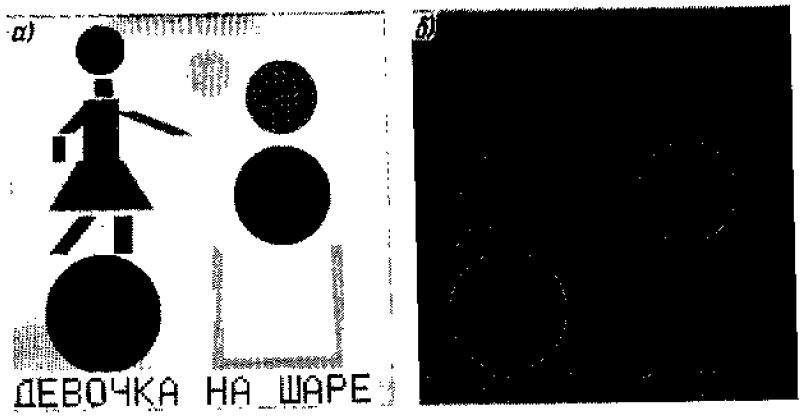


Рис. 8.6. «Разбеливание» тестового изображения геометрических фигур и букв:
а — исходное изображение; *б* — результат разбеливания

целесообразно перед коррелированием подвергать изображение оконтурнрованию с помощью тех или иных методов пространственного дифференцирования или грубо квантовать изображение для получения резких границ.

Кроме того, этот результат проливает новый свет на то, что следует называть контурами изображения и почему контуры так важны для зрительной системы. Понятие контуров часто встречается и по-разному определяется в работах по обработке и распознаванию изображений. С точки зрения локализации объектов на изображении оптимальным линейным измерителем, контуры — это то, что получается в результате разбеливания изображения. Чем интенсивнее эта контурная часть сигнала, описывающего объект (в частности, чем резче изображение объекта), тем надежнее локализация. Возможно, что с этих позиций можно также объяснить тот известный в психофизике зрения эффект, что визуальная за меткость помех и искажений вблизи резких перепадов яркости (границ объектов) ниже, чем там, где яркость меняется плавно, т.е. интенсивность «контурного» сигнала мала.

Отметим, что обычно, когда говорят о выделении контуров, чаще всего имеют в виду изотропные дифференцирующие процедуры выделения [64, 69]. Оптимальное же для локализации разбеливание необязательно изотропно и необязательно обладает дифференцирующим эффектом, поскольку оно определяется спектром фоновой части изображения (а в пространственно неоднородном измерителе — фрагментом изображения), на котором должен быть произведен поиск заданного объекта. Более того, по этой же причине оно адаптивно, т, е. характеристика разбеливающего фильтра подстраивается под наблюдаемое изображение, и оно по-разному сказывается на разных изображениях. Так, у прямоугольников и параллелограммов на фоне кругов подчеркиваются угловые точки, в изображении текста оконтуриваются вертикальные и горизонтальные фрагменты букв (от них тоже остаются практически только угловые точки), но почти не меняются наклонные фрагменты, как редко встречающиеся (см. рис. 8.6,6).

Выбор объектов с точки зрения надежности локализации. Имеется много задач, где объект локализации не задан и его необходимо выбрать. Спрашивается, как наилучшим образом осуществить этот выбор. В этом состоит проблема так называемых «характерных точек» изображения, которая стоит в стереограмметрии и в некоторых задачах «искусственного интеллекта». В работах по стереограмметрии обычно рекомендуется в качестве эталонных объектов выбирать фрагменты изображений с резко выраженными локальными характерными особенностями: перекрестья дорог, излучины рек, отдельно стоящие строения и т.п. Иногда в качестве таких объектов рекомендуют выбирать участки изображений, на которых достигается экстремум некоторых специально введенных функций информативности. Подобные качественные рекомендации встречаются в работах по распознаванию образов.

Представленный выше анализ дает решение этой задачи. Действительно, из формулы (8.4) для максимального значения отношения сигнал-шум, которое может быть достигнуто на входе решающего блока линейного измерителя, вытекает, что наилучшими эталонами будут те фрагменты изображения, энергия разбеленного спектра $\alpha_0/(|\alpha_и|^2+\varepsilon^2)^{1/2}$ которых максимальна. Тогда такие эталоны дадут наибольший отклик на выходе оптимального фильтра и, следовательно, обеспечат минимум ошибок ложного отождествления.

Отсюда вытекает следующая рекомендация к выбору эталонных объектов, например в задаче стереограмметрии. Одно из изображений стереопары необходимо разбить на достаточно малые по площади фрагменты и найти отношение их спектра $\alpha(f_1, f_2)$ к модулю скорректированного спектра второго изображения $\overline{|\alpha_и(f_1, f_2)|}^2+\varepsilon^2$. Затем для каждого фрагмента вычислить интеграл (8.4) (при цифровой обработке — соответствующую ему сумму; в силу соотношения Парсеваля такой же результат можно получить, интегрируя квадрат модуля разбеленного сигнала), и из полученных результатов выбрать нужное количество наибольших по величине. Поскольку, как уже отмечалось, в большинстве случаев спектр изображений представляет собой быстро спадающую функцию, наилучшими эталонами будут такие, которые имеют медленно спадающий спектр, т.е. участки изображений, которые мы визуально оцениваем как содержащие интенсивные контуры.

Эти рекомендации были экспериментально проверены в [6] и иллюстрируются рис. 8.7 и 8.8. Степень надежности обнаружения объекта (фрагмента исходного изображения) передана на рисунках степенью почернения на фотоотпечатках. На рисунках хорошо видно, что наилучшие фрагменты выделяются там, где на исходном изображении имеются какие-либо резко выраженные локальные особенности: перепады яркости, изменение рисунка текстур и т.п.

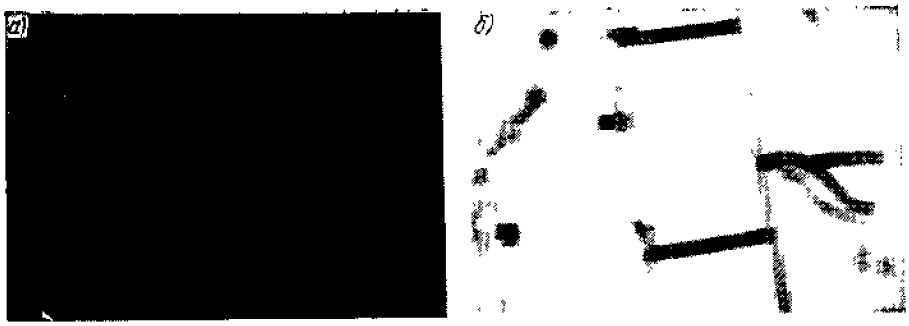


Рис. 8.7. Автоматическое выделение опорных объектов на аэрофотоснимке:
а — исходное изображение; *б* — результат выделения фрагментов 32×32 элемента

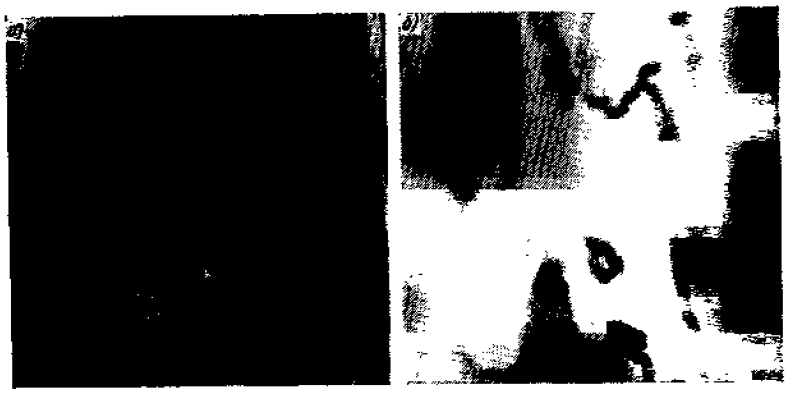


Рис. 8.8. Автоматическое выделение опорных объектов на космическом снимке:
а — исходное изображение; *б* — результат выделения фрагментов 32×32 элемента

Описанный алгоритм определения опорных объектов требует достаточно трудоемких вычислений, особенно при скользящей обработке. Поэтому могут представлять интерес более простые в вычислительном отношении алгоритмы, аппроксимирующие оптимальный. Эксперименты показывают [6], что в качестве таких упрощенных алгоритмов можно пользоваться алгоритмами вычисления локальной дисперсии или средних локальных значении градиента видеосигнала, для которых существуют быстрые рекурсивные алгоритмы.

Как уже отмечалось, все описанные в этой главе методы обработки наиболее эффективно могут быть реализованы в гибридной оптико-цифровой системе, базирующейся на адаптивном оптическом корреляторе с нелинейной средой в фурье-плоскости [48]. При чисто цифровой реализации этих методов для повышения быстродействия приходится идти на те или иные упрощения. Примерами могут служить грубое квантование разделенного сигнала [7], позволяющее резко сократить количество операций при вычислении корреляции разбеленного изображения и искомого объекта, и алгоритмы выделения реперных марок на аэро- и космических фотоснимках (см. § 8.5).

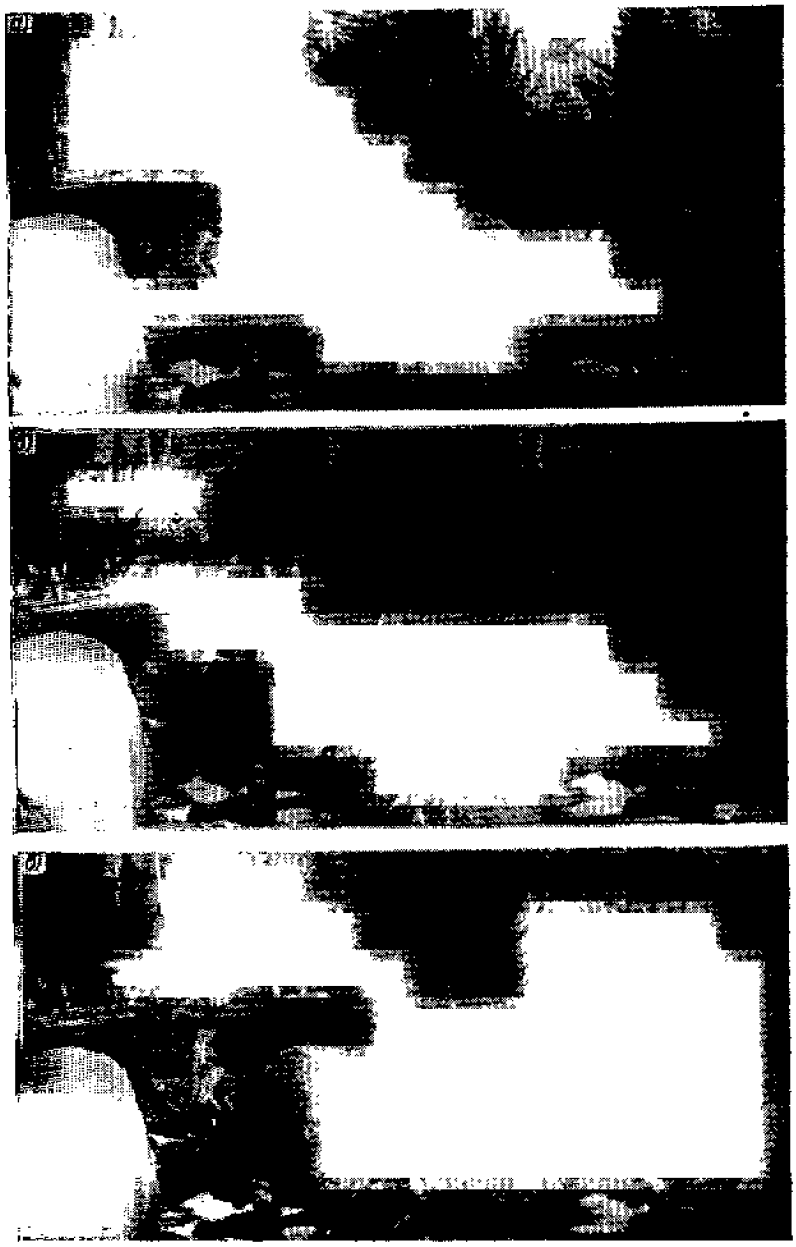
Избыточность стереоскопических изображений с точки зрения задачи локализации объектов. Как известно, стереоэффект является одним из основных механизмов объемного зрения. Это широко используется в различных проектах объемного телевидения и кинематографа, в прикладном телевидении, в аэрофотосъемке и картографии и многих других областях человеческой деятельности, связанных с получением и использованием визуальной информации. Поэтому большой практический интерес представляет оценка объема сигнала, соответствующего стереоскопическим изображениям, т.е. пропускной способности канала, необходимого для хранения и передачи стереоскопических изображений. Результаты анализа задачи оптимальной локализации объектов на изображении, описанные в этой главе, дают возможность получить такую оценку.

С информационной точки зрения два снимка, составляющие стереопару, эквивалентны одному из снимков и карте рельефа (глубин) изображаемой сцены. Действительно, по двум снимкам можно построить карту рельефа, и наоборот, по карте рельефа и по одному из снимков можно построить второй снимок стереопары. Поэтому приращение объема сигнала, даваемое

вторым снимком стереопары, равно объему сигнала, соответствующего карте рельефа. Как известно, количество разрешаемых глазом градаций глубины примерно равно количеству разрешаемых градаций яркости. Следовательно, относительное приращение объема сигнала будет определяться главным образом числом степеней свободы карты рельефа, т.е. количеством ее независимых отсчетов. Оценить это число можно с помощью следующих простых рассуждений.

Каждый отсчет карты рельефа можно найти путем отождествления соответствующих участков на снимках, составляющих стереопару, измерения их параллакса и пересчета этого параллакса в глубину рельефа (плана) с учетом геометрии съемки (наблюдения). Так работают все технические системы, в которых используются стереоизображения, и естественно предположить, что аналогично работает и механизм стереоскопического зрения. Количество независимых отсчетов карты рельефа, очевидно, равно отношению площади изображения к минимальной площади его участков, которые еще можно надежно отождествить на другом изображении стереопары. Очевидно также, что для надежного отождествления размеры отождествляемых участков должны превышать размеры элемента разрешения снимка и иметь площадь в несколько элементов разрешения. Из этого вытекает, что количество независимых отсчетов карты рельефа и, следовательно, приращение объема сигнала будет всегда в несколько раз меньше количества элементов разрешения снимка стереопары. Так, при размерах отождествляемых участков в 2x2 элемента приращение объема сигнала в 4 раза меньше объема сигнала, соответствующего одному снимку, при размере 3x3 элемента — в 9 раз меньше и г. д.

Результаты исследования оптимального линейного обнаружителя объектов на изображениях [6] показывают, что для надежного отождествления на сложных изображениях размеры участков должны превышать (8x8) — (10x10) элементов изображения. Это позволяет высказать гипотезу, что на сложных изображениях приращение объема сигнала для передачи стереоэффекта составляет проценты и даже доли процента объема сигнала одного снимка стереопары.



Для косвенной проверки этой гипотезы была проведена серия экспериментов по обработке стереоскопических снимков [52], заключавшихся в прореживании отсчетов одного из изображений стереопары и замене отброшенных отсчетов отсчетами, интерполированными по методу билинейной интерполяции. Цель экспериментов состояла в определении влияния прореживания на восприятие глубины и резкости наблюдаемого объемного изображения. Косвенными эти эксперименты являются потому, что они проверяют саму гипотезу, но не соображения, на которых она основана.

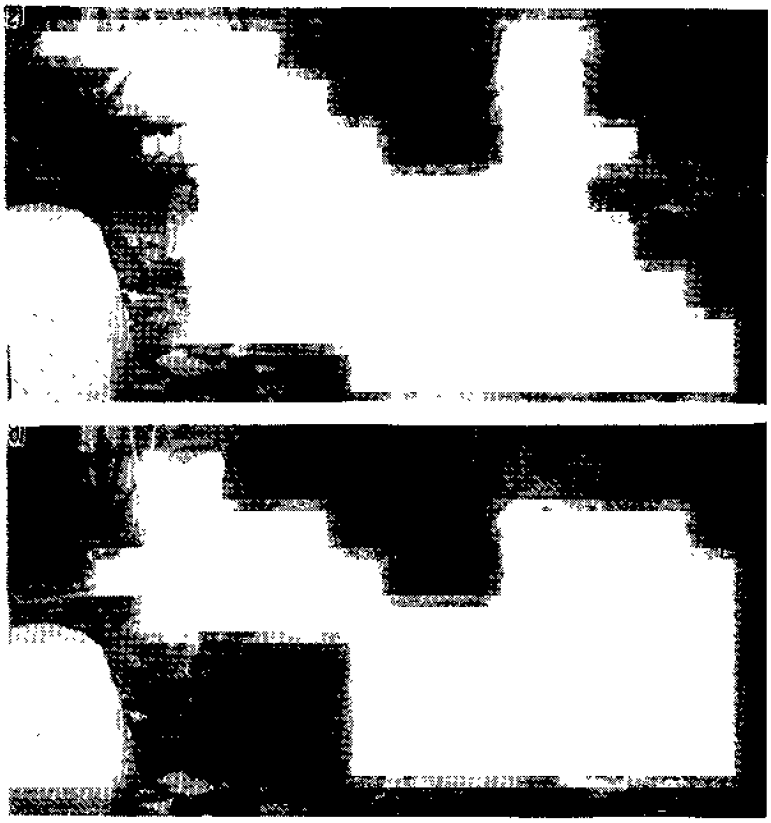


Рис. 8.9. Влияние прореживания одного из снимков стереопары на наблюдение стереоскопического эффекта:
а — исходная стереопара без прореживания; *б—д* — прореживание правого снимка с шагом — 2 : 1; 3 : 1; 4 : 1; 5 : 1

Рис. 8.10. Учебный аэрофотоснимок, использовавшийся в экспериментах по прореживанию

Эксперименты проводились с кадрами стереоскопического мультфильма (рис. 8.9) и учебным аэрофотоснимком {рис. 8.10). Кадры стереоскопического мультфильма представляли интерес как содержащие резкие скачкообразные изменения планов, на которых могла бы в более заметной степени сказаться потеря разрешения одного из снимков в результате прореживания и интерполяции отсчетов. Стереоскопический аэрофотоснимок использовался для количественной оценки влияния прореживания и интерполяции на точность измерения параллакса и тем самым на точность построения карты рельефа.

Наблюдая с помощью рисунков стереоскопические изображения, можно убедиться, что прореживание и интерполяция одного из снимков не сказывается заметным образом на качестве стереоскопического изображения даже при прореживании в 5x5 раз, когда объем сигнала

снимка уменьшается в 25 раз. Об этом же свидетельствуют и результаты измерения точности определения параллакса соответствующих точек, выполненные на стереокомпараторе для аэрофотоснимка рис. 8.10 по 31 случайно выбранным участкам. Эти результаты приведены на графиках рис. 8.11.

График на рис. 8.11, а показывает, что при прореживании до 1:3 среднеквадратическое значение ошибки измерения параллакса находится практически в пределах точности стереокомпаратора, характеризуемой ошибкой для нерастрированных, т.е. не подвергнутых дискретизации и последующему восстановлению изображений. Более того, растрирование и даже прореживание 1:2 несколько уменьшают ошибку. Это можно объяснить тем, что при дискретизации и восстановлении изображений прямоугольной апертурой на границах соседних отсчетов изображений возникают псевдоконтуры, которые несколько увеличивают точность локализации соответственных точек.

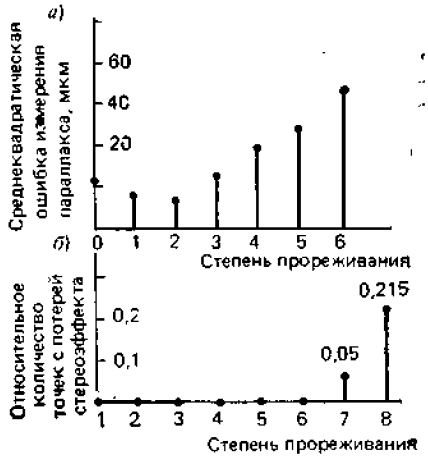


Рис. 8.11. Стандартное отклонение ошибки локализации (*а*) и вероятность аномальных ошибок локализации (*б*) как функции степени прореживания аэрофотоснимка рис. 8.10.

Как показывает график на рис. 8.11,6, заметная потеря стереоэффекта начинается только с прореживания 1:7. Все это говорит в пользу высказанной гипотезы. Ее на качественном уровне подтверждают также хорошо известные факты о возможности сильного искажения одного из снимков стереопары (снижения четкости, искажения перепадов полутонов, искажения и даже полной потери цвета) без существенного ухудшения стереоэффекта.

Однако рассуждения, использовавшиеся при оценке приращения объема сигнала, в свою очередь в известной мере объясняют эти явления.

Следует отметить, что соображения о минимальном размере участка для его отождествления по необходимости носят оценочный характер, поскольку можно представить себе специальные изображения и объекты (например, типа редких контрастных точечных или линейных объектов на абсолютно ровном фоне), для которых оценка приращения не будет столь оптимистична. Но для сложных, естественного происхождения изображений она, по-видимому, верна.

8.5. ОБНАРУЖЕНИЕ И ФИЛЬТРАЦИЯ ИМПУЛЬСНЫХ ПОМЕХ И СБОЕВ

Действие импульсных помех сказывается не на всем изображении, а в случайно расположенных точках или участках изображения, где значение сигнала заменяется случайной величиной. К этому же классу относятся так называемые сбои, которые приводят к тому, что па отдельных участках сигнал изображения пропадает и вместо истинного значения наблюдается другое, определяемое причиной сбоя: царапиной или повреждением эмульсии фотоматериала, сбоями радиоканала или иногда специальными служебными метками, вносимыми конструкцией изображающей системы. Поэтому любой алгоритм фильтрации таких помех состоит из двух этапов: обнаружения выбросов шума и исправления обнаруженных искаженных отсчетов сигнала. Ниже описаны алгоритмы фильтрации импульсных помех и выделения сбоев и крестообразных реперных марок на изображениях, использующие методы обнаружения помех и сбоев, основанные на рассмотренных выше принципах построения линейного обнаружителя-измерителя координат объектов на изображениях.

Алгоритм обнаружения и выделения сбоев и реперных марок на фотоснимках. В фотографических и фототелевизионных системах для аэрофото- и космосъемки иногда

предусматривается экспонирование изображений через специальные маски, содержащие непрозрачные реперные марки в виде точек или перекрестий, предназначенные для геометрической калибровки изображений. Одной из задач автоматической обработки изображений является обнаружение и выделение этих марок. Поскольку марки и сбойные участки изображения представляют собой малоразмерные и, как правило, контрастные объекты, их обнаружение возможно •с помощью упрощенного фильтра

$$\hat{a}_{k,l} = b_{k,l} - \frac{1}{(2N_1+1)(2N_2+1)} \sum_{n=-N_1}^{N_1} \sum_{m=-N_2}^{N_2} b_{k-n,\,l-m}. \qquad (8.12)$$

Такой фильтр подавляет нижние пространственные частоты изображения, и его параметры N_1 и N_2 можно подобрать так, чтобы аппроксимировать частотную характеристику оптимального фильтра (8.11), настроенного на локализацию объектов с равномерным спектром. Практически оказывается, что N_1 и N_2 следует выбирать порядка половины размера выделяемых марок и сбойных участков изображения. Такая аппроксимация в принципе снижает надежность обнаружения, но зато дает возможность путем простого сравнения сигнала на выходе фильтра с порогом выделять марки и сбойные участки практически без искажений их формы [46]. Результаты выделения марок и сбойных участков изображения таким алгоритмом иллюстрируются рис. 8.12.

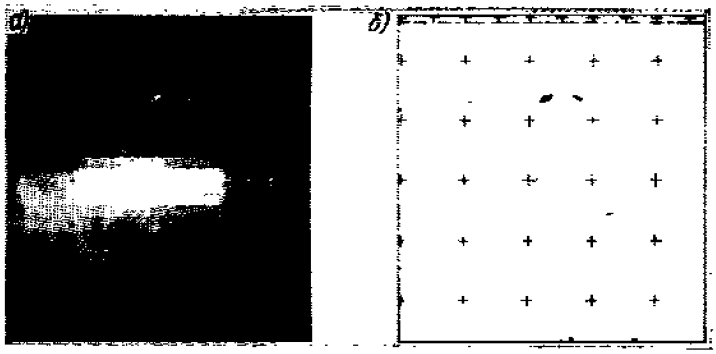


Рис. 8.12. Выделение реперных марок и сбоев:
а — исходное изображение; б — выделенные марки и сбойные участки изображения

Итеративный алгоритм фильтрации импульсных помех. Импульсные помехи представляют собой объекты размером в один элемент изображения. Поэтому для их обнаружения может быть использован фильтр (8.12) с $N_1=N_2=1$. Если сигнал на выходе такого фильтра превышает по модулю некоторый порог δ, то принимается решение о наличии выброса шума. Все обнаруженные таким образом искаженные отсчеты изображения отмечаются, после чего они исправляются путем замены значениями, усредненными по их не отмеченным соседним отсчетам.

Отсчеты видеосигнала, используемые при предсказании в окрестности 3x3 элемента и попадающие в апертуру фильтра (8.12), в свою очередь могут быть искажены. Поэтому алгоритм фильтрации должен быть итеративным с понижением порога δ в процессе итераций. Таким образом, фильтрация импульсных помех в соответствии с этим алгоритмом производится за несколько итераций, причем каждая итерация осуществляется за два прохода по изображению: один — на обнаружение выбросов шума и второй — на исправление искаженных отсчетов. Эксперименты показывают, что достаточно не более трех-четырех итераций.

На рис. 8.13 показан результат моделирования работы описанного алгоритма, а на рис. 8.14,а–е — графики видеосигнала для тех же изображений, что и на рис. 8.13, а–е, иллюстрирующие степень подавления помехи и связанные с фильтрацией искажения сигнала.

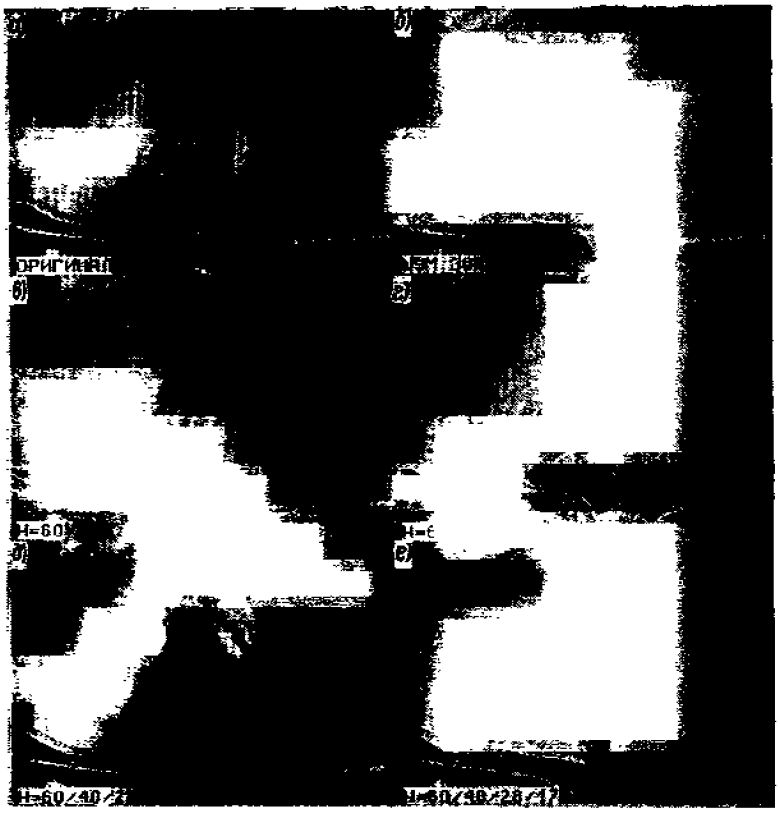


Рис. 8.13. Подавленке импульсного шума итеративным алгоритмом с предсказанием:

а — незашумленное изображение; *б* — зашумленное изображение с вероятностью искажения отсчетов 0,3; *в—е* — результаты фильтрации соответственно после первой — четвертой итераций

Рекурсивный алгоритм фильтрации импульсных помех. Итеративный алгоритм фильтрации импульсных помех обладает достаточно высокой эффективностью фильтрации, но чрезмерно сложен в вычислительном отношении, так как он требует нескольких проходов по изображению. Его можно упростить и осуществлять фильтрацию за один проход, если в фильтре (8.12) для вычисления лока.тьного среднего используются не все восемь отсчетов изображения, соседних с данным на прямоугольном растре, а только четыре уже обработанных отсчета, предшествующих данному при развертке слева направо и сверху вниз (три на предыдущей строке и один слева на той же строке). В результате мы приходим к обнаружению выбросов шума методом предсказания, описанным в § 6.4. Основанный на этом методе рекурсивный алгоритм фильтрации можно описать следующим соотношением:

$$\hat{a}_{k,l} = \begin{cases} b_{k,l}, & \text{если } |d| < \delta_1, \\ \bar{b}_{k,l} + \delta_2 \operatorname{sign} d & \text{в противном случае,} \end{cases}$$

где $d = b_{k,l} - \bar{b}_{k,l}$; $\bar{b}_{k,l} = (b_{k-1,\,l} + b_{k-1,\,l-1} + b_{k,\,l-1} + b_{k+1,\,l-1})/4$.

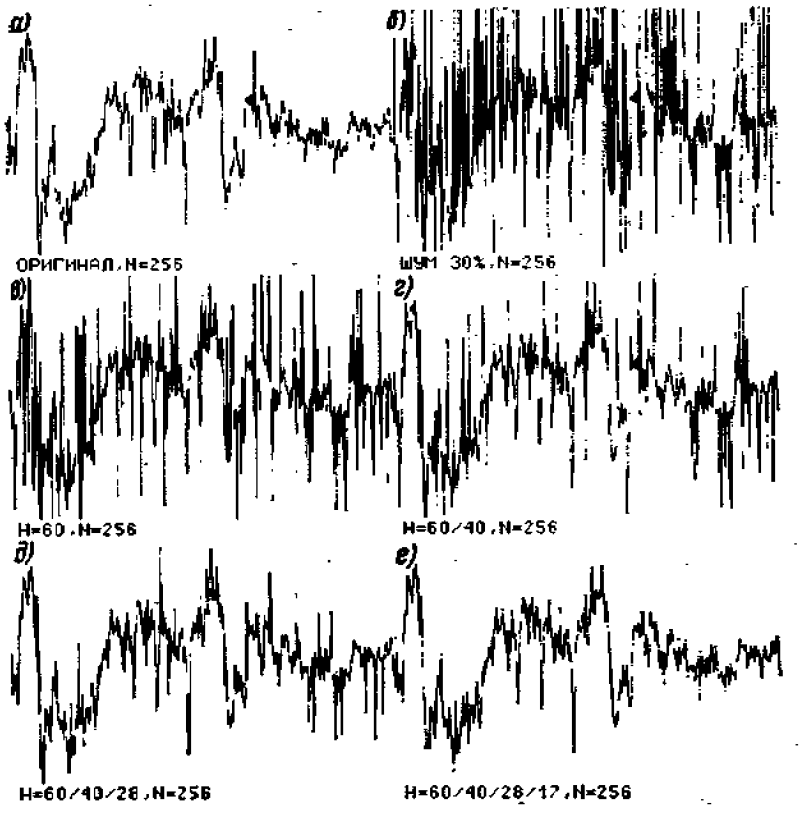


Рис. 8.14. Графики видеосигнала изображений рис. 8.13

Порог δ_1 можно находить автоматически по положению излома в гистограмме распределения сигнала d. Порог δ_2 вводится для того, чтобы обеспечить устойчивость алгоритма и выбирается равным 3—4% размаха значений видеосигнала ([46]). Работа алгоритма иллюстрируется рис. 8.15. Качество работы итеративного и рекурсивного алгоритмов можно сравнить по данным табл. 8.1, где приведены полученные моделированием значения вероятности пропуска и ложного обнаружения выбросов импульсного шума, и данным табл. 8.2, в которой приведены значения среднеквадратической и средней ошибки фильтрации для четырех значений вероятности искажения элементов изображения.

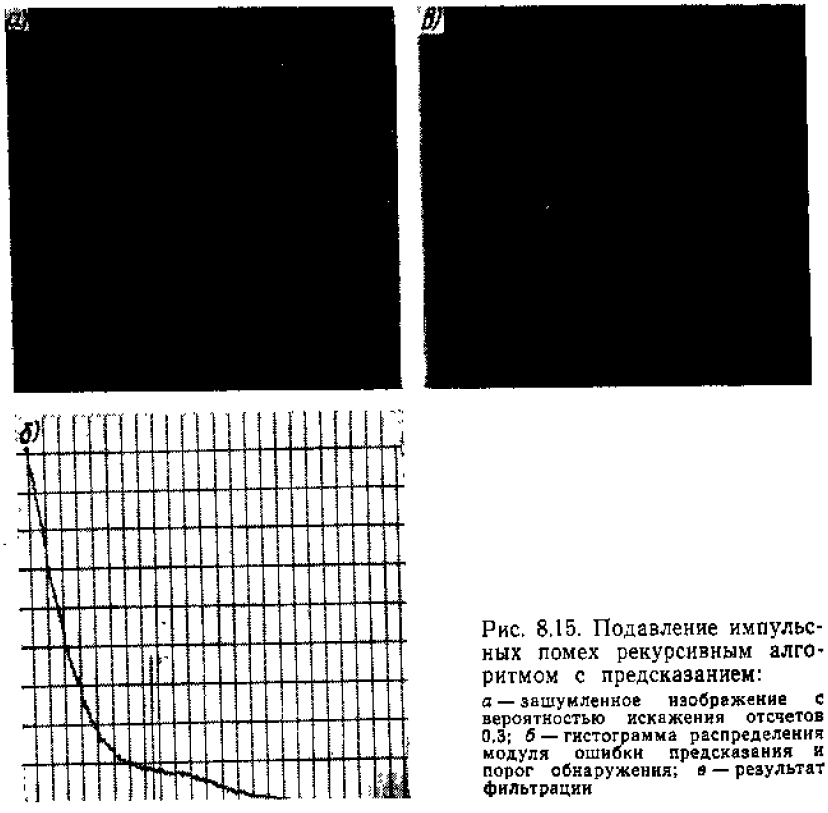


Рис. 8.15. Подавление импульсных помех рекурсивным алгоритмом с предсказанием:

а — зашумленное изображение с вероятностью искажения отсчетов 0,3; *б* — гистограмма распределения модуля ошибки предсказания и порог обнаружения; *в* — результат фильтрации

Таблица 8.1. *Сравнение алгоритмов фильтрации импульсных помех по критериям вероятностей пропуска и ложного обнаружения*

Вероятность ошибки на один элемент изображения, %	$P_{проп}$, %				$P_{ло}$, %			
	10	20	30	40	10	20	30	40
Итеративный алгоритм с предсказанием:								
- первая итерация	74	76	77	77	0	0	0	0
- вторая итерация	48	52	54	54	0	0	0	0
- третья итерация	30	33	34	34	1	1	1	1
- четвертая итерация	18	19	20	20	6	5	5	5
Рекурсивный алгоритм с предсказанием	11	11	11	8	8	15	21	28

Таблица 8.2. *Сравнение алгоритмов фильтрации импульсных помех по критериям среднеквадратичной и средней ошибок фильтрации*

Вероятность ошибки на один элемент изображения, %	Относительная среднеквадратическая ошибка.				Относительная средняя ошибка, %			
	10	20	30	40	10	20	30	40
Итеративный алгоритм с предсказанием:								
- третья итерация	2,5	3,6	4,6	5,6	0,06	0,12	0,13	0,15
- четвертая итерация	3,1	3,7	4,1	5	0,04	0,05	0,08	0,15
Рекурсивный алгоритм с предсказанием	4	4,6	5,4	6,1	0,1	0,16	0,26	0,38

Глава 9

РАНГОВЫЕ АЛГОРИТМЫ ОБРАБОТКИ ИЗОБРАЖЕНИИ

9.1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Методы адаптивной линейной фильтрации, рассмотренные в гл. 7, 8, являются важным классом методов обработки изображений. В своей вычислительной реализации они опираются на быстрые алгоритмы свертки и спектрального анализа сигналов (см. гл. 4, 5}. В данной главе рассматривается класс нелинейных алгоритмов обработки изображений, которые будем называть ранговыми и которые основаны на быстрых алгоритмах вычисления локальных гистограмм распределения и их характеристик.

Линейную фильтрацию последовательности отсчетов дискретного сигнала $\{a_k\}$ можно описать как преобразование вида: $b_k = L_k(a_k)$, где $L_k(a_k)$ – линейная функция величины a_k: $L_k(a_k) = = h_0 a_k + \bar{a}_k; \bar{a}_k$ – взвешенная сумма остальных отсчетов сигнала, составляющих заданную пространственную окрестность данного элемента (см. гл. 3, 4). Ранговые алгоритмы осуществляют нелинейное преобразование сигнала $b_k = \Phi_k(a_k)$, причем $\Phi_k(a_k)$ –функция, вид которой определяется некоторым заданным подмножеством так называемых рангов или ранговых (порядковых) статистик выборки отсчетов, образованной данным отсчетом сигнала и отсчетами, составляющими его заданную окрестность.

В дальнейшем используются следующие обозначения:

(k, l)–координаты текущего элемента изображения;

$v_{k,l}$ –квантованное исходное значение видеосигнала в элементе *(k,l)*, $v_{k,l} \in [0, Q-1]$

Q – количество уровней квантования исходного сигнала;

$\hat{v}_{k,l}$ – квантованное значение видеосигнала в элементе *(k,l)*, получаемое в результате преобразования;

$\mathcal{Q}$ – количество уровней квантования преобразованного сигнала;

S – окрестность элемента *(k, l)* – заданное определенным образом множество элементов изображения, окружающих «центральный элемент» *(k, l)*, включая его самого;

N_S – объем S-окрестности, т.е. количество составляющих ее элементов;

$v_S(r)$ – вариационный ряд из элементов S-окрестности, $r = 0, 1, ..., N-1$;

$\{h_S(q)\}$ – гистограмма значений видеосигнала в S-окрестности;

q – номер квантованного значения видеосигнала, q = 0, 1,…, Q-1;

$r_S(v)$ – ранг элемента с величиной видеосигнала и в вариационном ряду, построенном по S-окрестности;

M-окрестность – определенным образом заданное множество элементов вариационного ряда, включающее данный элемент *(k,l)*.

Варианты M-окрестности: KSN-окрестность– окрестность из K ближайших на растре соседей; KNV-окрестность – окрестность по K. ближайшим по величине сигнала соседям:

$$M_{KNV} = \left\{ v_S(i) : \sum_{i=p}^{p+K} | v_{k,l} - v_S(i)| = \min_p \right\};$$

ε_V -окрестность:

$$M_{\varepsilon V} = \{ v_S(i) : v_{k,l} - \varepsilon_V \leqslant v_S(i) \leqslant v_{k,l} + \varepsilon_V \};$$

ε_r -окрестность:

$$M_{\varepsilon r} = \{ v_S(i) : r_S(v_{k,l}) - \varepsilon_r < i \leqslant r_S(v_{k,l}) + \varepsilon_r \};$$

$r_{KNV}(v)$ – ранг величины v в вариационном ряду, построенном по KNV-окрестности;

$r_{\varepsilon_r}(v)$ – ранг величины v в вариационном ряду, построенном по ε_r-окрестности;

$v_M(R)$ – значение элемента в вариационном ряду по M-окрестности с рангом R;

$v_M(L)$ – значение элемента в вариационном ряду по M-окрестности с рангом L;

MEAN(M) – среднее арифметическое значение элементов M-окрестности:

$$\mathrm{MEAN}\,(\mathrm{M}) = \frac{1}{N_M} \sum_{(k,l)\in M} v_{k,l};$$

MED (M) – медиана элементов М-окрестности:

$$\mathrm{MED}\,(\mathrm{M}) = v_{\mathrm{M}}\,(r = (N_{\mathrm{M}} + 1)/2);$$

CUT(M)–«срезка» по М-окрестности:

$$\mathrm{CUT}\,(M) = \begin{cases} v_{k,l},\ v_{\mathrm{M}}\,(L) \leqslant v_{k,l} \leqslant v_{\mathrm{M}}\,(R); \\ v_{\mathrm{M}}\,(L),\ v_{k,l} < v_{\mathrm{M}}\,(L); \\ v_{\mathrm{M}}\,(R),\ v_{k,l} > v_{\mathrm{M}}\,(R); \end{cases}$$

RAND(M) – псевдослучайная величина, гистограмма распределения значений которой совпадает с гистограммой распределения значений элементов изображения по М-окрестности;

MIN(M), MAX(M)–минимальное и максимальное значения по окрестности.

Очевидно, что любую r-ю порядковую статистику $v_S(r)$ элемента изображения, имеющего координаты (k, l), заданную окрестность которого образуют остальные (N_{S-1}) элементов выборки, можно найти из локальной гистограммы $\{h_S(q)\}$ распределения значений элементов окрестности, решив уравнение:

$$N_S \sum_{q=0}^{v_S(r)} h_S(q) = r.$$

Вычисления локальных гистограмм для окрестностей каждого элемента изображения при последовательном сканировании изображения апертурой, охватывающей требуемую окрестность, осуществляются с помощью быстрого рекурсивного алгоритма, описанного в §6.3, так что вычислительная сложность алгоритмов ранговой фильтрации почти не зависит от размеров окрестности. Кроме того, при вычислении конкретных ранговых статистик и производных от них возможны дальнейшие упрощения, связанные, в частности, с информационной избыточностью изображений. Поэтому ранговые алгоритмы по простоте вычислений в принципе не уступают алгоритмам линейной фильтрации.

Ранговые алгоритмы не уступают алгоритмам линейной фильтрации и даже превосходят их и в другом отношении – в простоте локальной адаптации. Они являются адаптивными, а точнее

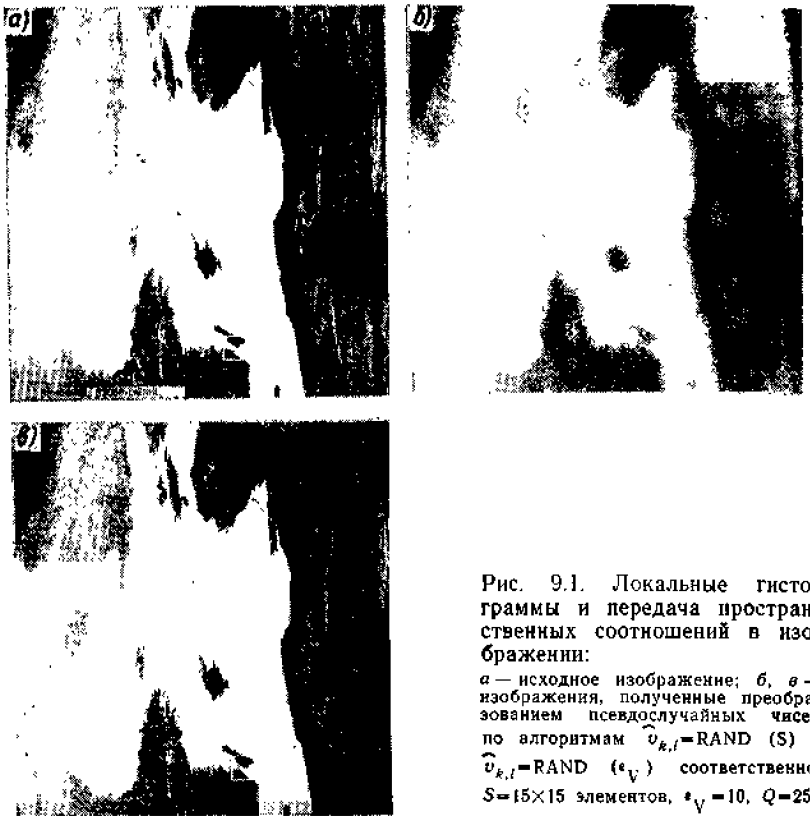


Рис. 9.1. Локальные гистограммы и передача пространственных соотношений в изображении:
$а$ — исходное изображение; $б$, $в$ — изображения, полученные преобразованием псевдослучайных чисел по алгоритмам $\hat{v}_{k,l}$=RAND (S) и $\hat{v}_{k,l}$=RAND (ε_V) соответственно; S=15×15 элементов, ε_V=10, Q=256

локально-адаптивными по своей сути, так как их параметры по определению являются функциями локальной характеристики изображений– их гистограмм. В то же время ранговые

алгоритмы лишены характерного недостатка методов линейной фильтрации – их пространственной инерционности: взаимное влияние отдельных деталей изображения простирается на результирующем изображении на расстояние порядка размеров апертуры фильтра. Это проявляется, в частности, в размазывании границ деталей при сглаживании изображений, в искажении формы деталей при их выделении из фона и т.п.

На первый взгляд может показаться, что поскольку ранговые алгоритмы переупорядочивают данные в вариационный ряд, они не используют пространственных связей между элементами изображений, и это является их принципиальным недостатком. Действительно, ранговые алгоритмы инвариантны к размерности сигнала. Однако, как это ни удивительно, это свойство является не недостатком, а преимуществом ранговых алгоритмов, еще одной стороной их адаптивного характера. Пространственные связи между элементами изображения, определяемые принадлежностью их к одной детали, проявляют себя в вариационном ряду через параметры условной гистограммы распределения значений сигнала в окрестности данного элемента, например через гистограммы по ε_v- и KNV-окрестностям. Высокую информативность этих гистограмм иллюстрирует рис. 9.1. Это отражение пространственных связей не зависит от ориентации деталей, и тем самым снимается необходимость априорного знания или измерения ориентации и формы деталей, которая существует при синтезе оптимальных линейных фильтров. Более того, это создает возможность определения неизвестных пространственных характеристик деталей.

Термин «ранговые алгоритмы» в обработке изображений появился сравнительно недавно. Однако многие алгоритмы обработки изображений, которые фактически относятся к этому классу, известны уже давно. Так, в [3] было введено понятие адаптивных амплитудных преобразований, включавшее тогда адаптивное квантование мод, скользящую эквализацию и степенную интенсификацию. Примерно с середины 70-х годов для сглаживания изображений стал использоваться предложенный Тьюки [71] алгоритм медианной фильтрации [32, 39, 44, 64]. Известны также алгоритмы экстремальной фильтрации, использующие значения минимума и максимума по окрестности. Теперь становится ясным, что псе они являются представителями единого большого класса ранговых алгоритмов [51].

Ранговые алгоритмы могут использоваться во всех процедурах обработки изображений. Рассмотрим их применения для стандартизации, сглаживания изображений, усиления детальности изображений, выделения объектов из фоновой части изображений, выделения границ объектов, определения статистических характеристик изображений и их искажений.

9.2. АЛГОРИТМЫ СГЛАЖИВАНИЯ ИЗОБРАЖЕНИЙ

Понятие сглаживания изображений имеет двоякий смысл. При коррекции искажений сигнала, внесенных изображающей системой, сглаживание – это подавление помех, связанных с несовершенством изображающей системы: аддитивных, флуктуационных, импульсных и др. При препарировании изображений сглаживание– это устранение деталей (обычно малоразмерных), мешающих восприятию нужных объектов на изображениях (так называемая генерализация изображения).

При коррекции искажений, вызванных изображающей системой, сглаживанию подвергается изображение на выходе изображающей системы. При препарировании сглаживание может применяться к изображению на любой стадии препарирования как одни из его этапов.

Понятие сглаживания всегда подразумевает некоторое представление об «идеально гладком» сигнале. Такой сигнал – цель сглаживания.

Для изображений таким «идеально гладким» сигналом можно считать сигнал, описываемый кусочно-постоянной моделью, т.е. «лоскутное» изображение с пятнами-деталями, имеющими постоянное значение сигнала в пределах каждого пятна. Действительно, представление изображения в виду кусочно-постоянного есть не что иное, как сегментация изображений, являющаяся конечной целью анализа изображений для построения их описания. На первый взгляд может показаться, что оно применимо только к «детальным» изображениям. Но это справедливо и для «текстурных» изображений, только в этом случае оно относится не к первичному видеосигналу, а к его признаку, характеризующему текстуру.

Понятие сглаживания подразумевает также представление о том, что должно быть подавлено при сглаживании. Будем называть подавляемую часть сигнала шумом. Рассмотрим ранговые алгоритмы сглаживания для двух наиболее характерных моделей шума – аддитивной и импульсной.

Сглаживание для аддитивной модели. Аддитивная модель шума предполагает, что наблюдаемый сигнал представляет собой сумму полезного сигнала и шума. Ранговые алгоритмы сглаживания аддитивного шума проще всего обосновывать с позиций кусочно-постоянной модели изображения. Действительно, любая S-окрестность «идеального» кусочно-постоянного изображения имеет гистограмму распределения значений видеосигнала вида:

$$h_S^0(q) = \sum_n H_n \delta(q - q_n), \qquad (9.1)$$

где q_n–значение сигнала на пятне с номером n, имеющем площадь H_n. Гистограмма распределения значений видеосигнала в любой S-окрестности наблюдаемого негладкого изображения отличается от $h_S^0(q)$ тем, что в ней чистые моды, определяемые дельта-функциями в (9.1), размыты вследствие тех или иных факторов (шума, искажений, наличия посторонних деталей и т.п.), так что она может быть записана в виде:

$$h_S(q) = \sum_n H_n \Delta_n(q - q_n),$$

где $\Delta_n(\cdot)$ – некоторая унимодальная функция, характеризующая размытие моды.

Таким образом, $h_S(q)$ –это, вообще говоря, мультимодальная функция, в которой первичные моды $\{\delta_n(q - q_n)\}$ проявляются в виде некоторых кластеров, или локальных максимумов.

Сглаживание теперь можно определить как оценку параметра кластера, к которому принадлежит данный элемент. Для того чтобы найти эту оценку, необходимо определить границы кластера. Можно предложить два способа определения границ кластера: адаптивное квантование мод [3, 46, 69] и «выращивание» кластера [28].

Адаптивное квантование мод заключается в том, что анализируется гистограмма распределения значений сигнала изображения (это может быть сигнал значений яркости изображения, плотности фотокегатива или значений того или иного скалярного признака, измеренного на изображении) и в ней отыскиваются границы между локальными максимумами. Эти границы рассматриваются как границы интервалов квантования, и все значения сигнала на изображении, попавшие в тот или иной интервал, заменяются значением, равным положению максимума (моды) гистограммы в этом интервале (рис. 9.2).

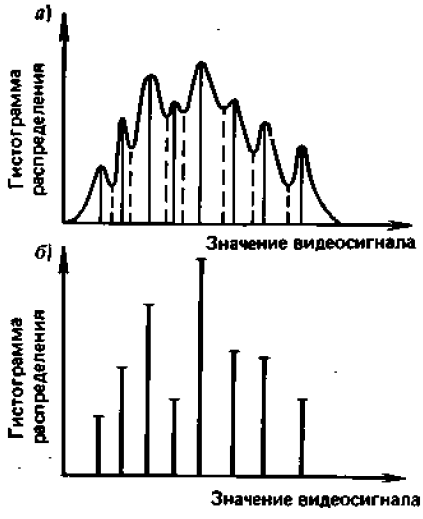


Рис. 9.2. К понятию адаптивного квантования мод:
а – исходная гистограмма распределения значений видеосигнала; б – гистограмма после адаптивного квантования; – – – – границы между модами, —— – значения сигнала в максимумах гистограммы (модах)

Качество адаптивного квантования мод зависит от того, насколько хорошо разделяются моды гистограммы. Степень «размытия» мод определяется степенью однородности объектов на изображении по выбранному для знали-

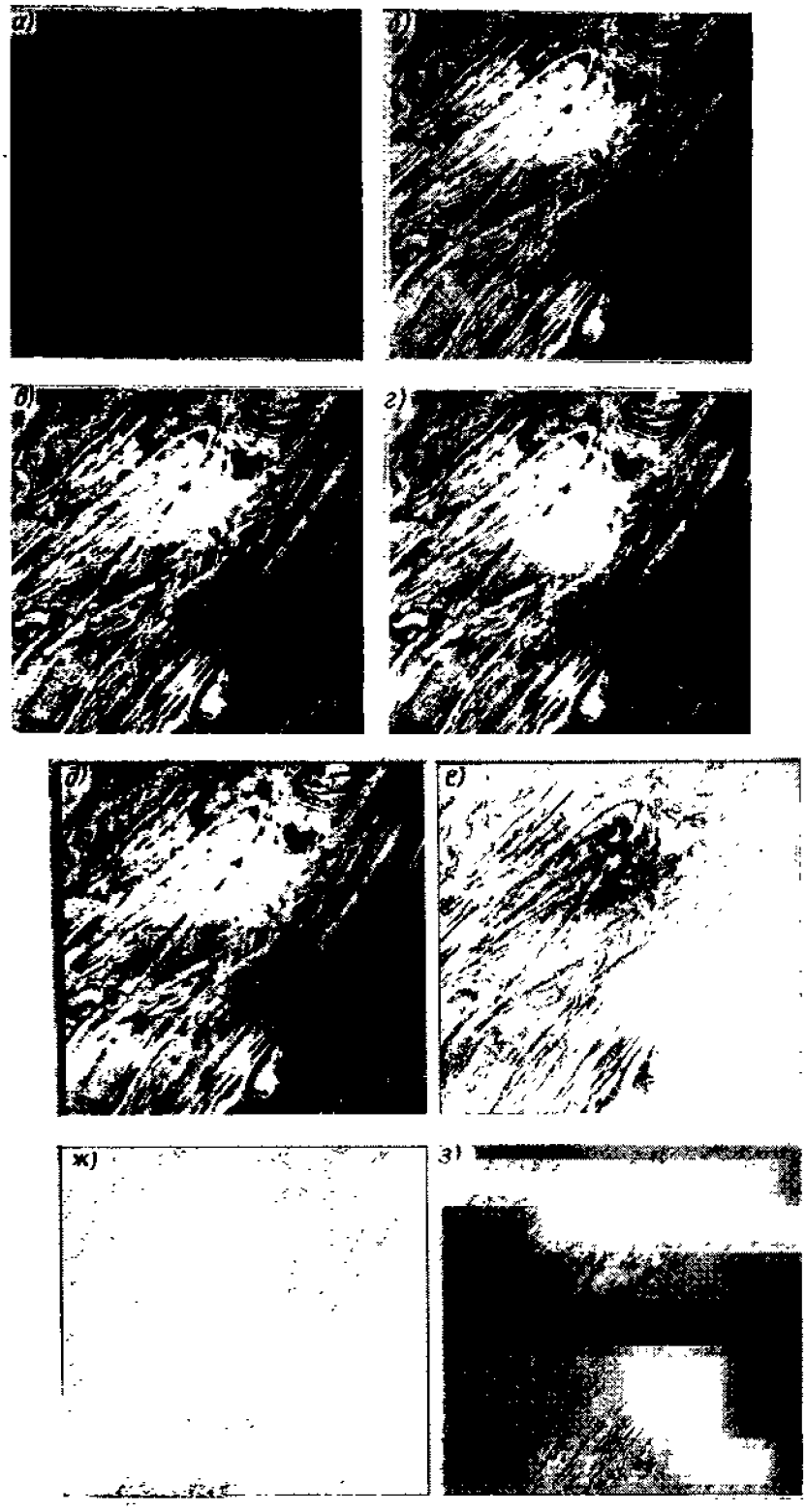


Рис. 9.3. Адаптивное квантование мод:
а – исходное изображение; б– д – квантование с порогом отбраковки мод соответственно» 4, 5, 7 и 10% (три уровня квантования); е – одна из мод изображения д; ж –границы деталей, образующих эту моду; з – положение границ на исходном изображении

за признаку, т.е. степенью соответствия изображения кусочно-постоянной модели, а также наличием искажений изображения: шумом датчика видеосигнала, дефокусировкой и т.п.

Для улучшения разделимости мод и повышения достоверности адаптивного квантования его целесообразно производить по отдельным фрагментам, размер которых выбирается так, чтобы они содержали небольшое число деталей изображения (т.е. чтобы в гистограмме было небольшое число мод). Кроме того, хорошие результаты дает использование условной гистограммы распределения, которая строится по значениям видеосигнала только в тех

элементах изображения, где эти значения незначительно отличаются от значений в соседних элементах [3, 46]. Степень допустимого отличия может задаваться априори или определяться автоматически в зависимости от локальной дисперсии видеосигнала на изображении.

При адаптивном квантовании мод может оказаться, что выделяются моды, площадь которых, т.е. количество элементов изображения, ей принадлежащих, относительно невелико. На изображении такие моды проявляются обычно в виде разбросанных точек, которые разбивают границы деталей изображения, образующих более мощные моды. Поэтому адаптивное квантование мод целесообразно сочетать с отбраковкой выделяемых мод по их мощности. Если площадь моды в гистограмме (ее мощность) меньше заданной пороговой величины, эта мода объединяется с соседней более мощной. Примеры адаптивного квантования мод показаны на рис. 9.3.

При адаптивном квантовании мод определяются границы всех кластеров гистограммы изображения или его фрагментов. При скользящей обработке, когда нужно принять решение о принадлежности к тому или иному кластеру только одного, центрально-то элемента анализируемого фрагмента, определять границы всех кластеров гистограммы – слишком трудоемкая задача. В этом случае целесообразнее использовать метод «выращивания» кластера. Он заключается в том, что кластер, к которому принадлежит центральный элемент анализируемой окрестности, определяется путем последовательных приближений. Сначала выбирается центр выращивания – предварительная оценка положения моды, относительно которой строится некоторая М-окрестность, являющаяся оценкой границ кластера. После этого по М-окрестности выбирается следующее приближение к центру кластера, относительно которой строится новая М-окрестность, к т.д. Возможные варианты выбора центра выращивания и М-окрестности для определения границ кластера, к которому принадлежит данный элемент, показаны в табл. 9.1.

Таблица 9.1. *Варианты центра выращивания и М-окрестности*

Центр выращивания	Л1-окрестность		
	KNV-окрестность	ε_V -окрестность	ε_r- окрестность
$v_{k,l}$	KNV($v_{k,l}$)	ε_V ($v_{k,l}$)	ε_r ($v_{k,l}$)
Среднее арифметическое по М	KNV(MEAN(M))	ε_V(MEAN(M))	ε_r (MEAN(M))
Медиана по М	KNV(MED(M))	ε_V (MED(M))	ε_r (MED(M))
Срезка по М	KNV(CUT(M))	ε_V(CUT(M))	ε_r(CUT(M))

Выбор в качестве центра выращивания той или иной предварительной оценки элемента *(k, l)*, а также выбор вида окрестности определяются характером обрабатываемого изображения и наличием априорной информации. Выбор KNV-окрестности позволяет учесть априорную информацию о геометрических размерах деталей изображения, которые нужно сохранить. Как правило, можно дать следующую рекомендацию: К должно быть порядка площади деталей, которые должны быть сохранены при сглаживании. Выбор ε_V-окрестности позволяет учитывать априорную информацию о минимальных перепадах, которые нужно сохранить, или о дисперсии шума, который нужно подавить. Третий вид окрестности – ε_r -окрестность – нашел свое применение в алгоритмах фильтрации импульсных помех и в описываемых ниже алгоритмах выделения границ. Размер S-окрестности должен быть примерно равен удвоенному размеру минимальной детали, которую нужно сохранить при сглаживании. Благодаря адаптивным свойствам ранговых алгоритмов форма S-окрестности слабо влияет на качество сглаживания.

Для оценки $\hat{v}_{k,l}$ текущего элемента изображения после определения границ кластера, к которому он принадлежит, можно рекомендовать среднее арифметическое или медиану в пределах кластера, а также «срезку» по границам кластера.

Выбор оценки текущего элемента изображения в пределах кластера, равно как и положения центра выращивания, определяется характером распределения значений шума, который необходимо подавить. Если распределение шума является гауссовским, то среднее арифметическое дает лучшую оценку; если распределение имеет более «тяжелые хвосты», то лучше использовать медиану или срезку. При этом необходимо иметь в виду возможное стирание мелких деталей на первом этапе и неточное определение кластера на втором этапе процедуры сглаживания.

Таким образом, выбирая разные способы оценки центра выращивания кластеров и окрестностей, можно получить семейство ранговых алгоритмов сглаживания. Некоторые из представителей этого семейства описаны в литературе. Так, в [54] описан алгоритм усреднения по К ближайшим соседям, который в наших обозначениях можно записать как

$$\hat{v}_{k,l} = \mathrm{MEAN}\,(\mathrm{KNV}\,(v_{k,l})).$$

В [58] описан фильтр, работающий по следующему алгоритму:

$$\hat{v}_{k,l} = MEAN(KNV(v_{k,l})) \qquad (9.2a)$$

причем $\varepsilon_V = 1{,}5\sigma$, где σ– стандартное отклонение гауссовского шума, которое считается известным. В [63] описан аналогичный фильтр с внутренним окном, правда, не для обработки изображений, а для сглаживания речевых сигналов:

$$\hat{v}_{k,l} = \mathrm{MEAN}\,(\varepsilon_V\,(\mathrm{MED}\,(\mathrm{KSN}\,(v_{k,l})))). \qquad (9.26)$$

Очевидно, к этому же семейству принадлежит простейший и наиболее известный ранговый метод сглаживания – медианная фильтрация [32, 39, 44, 64]:

$$\hat{v}_{k,l} = \mathrm{MED}\,(\mathrm{S}).$$

Отметим, что в литературе по математической статистике срезка по ε_r-окрестности называется винзоризацией, а среднее по ε_v -окрестности – усеченным средним. Для ранговых алгоритмов

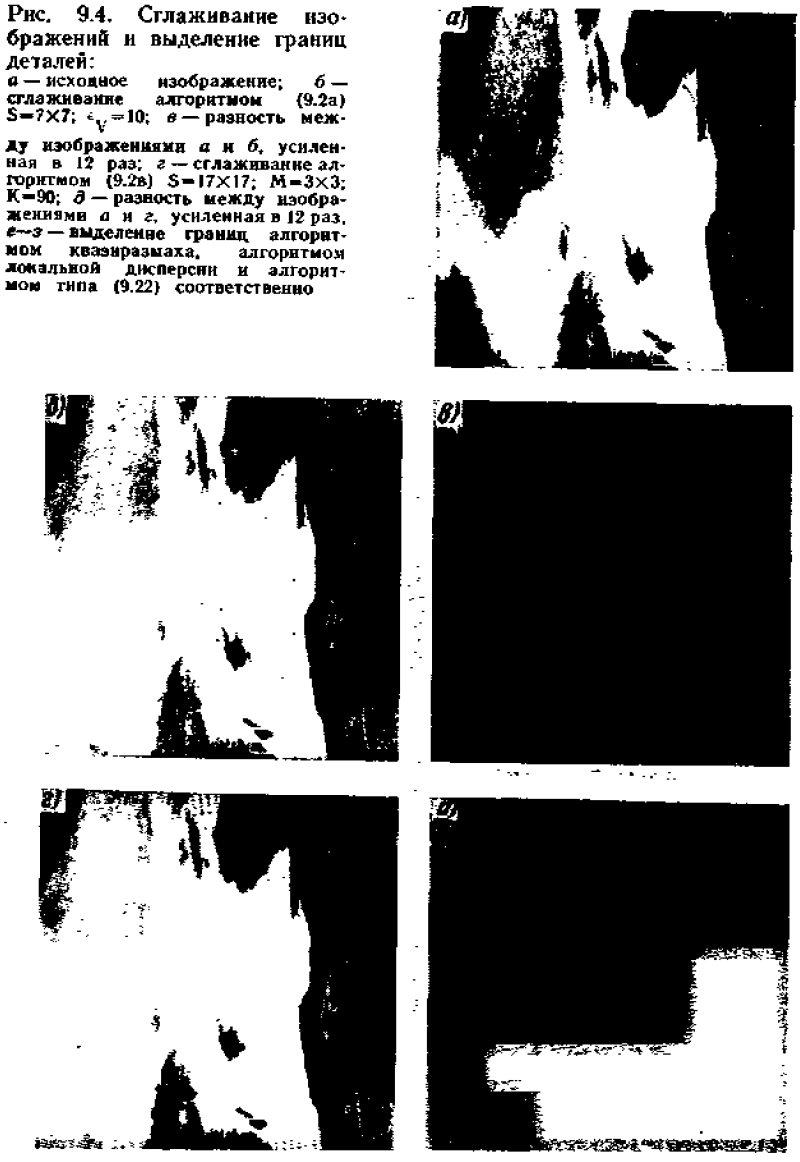


Рис. 9.4. Сглаживание изображений и выделение границ деталей:
а — исходное изображение; *б* — сглаживание алгоритмом (9.2а) S=7×7; ε_V=10; *в* — разность между изображениями *а* и *б*, усиленная в 12 раз; *г* — сглаживание алгоритмом (9.2в) S=17×17; M=3×3; K=90; *д* — разность между изображениями *а* и *г*, усиленная в 12 раз, *е—з* — выделение границ алгоритмом квазиразмаха, алгоритмом локальной дисперсии и алгоритмом типа (9.22) соответственно

обработки изображений в ряде случаев более удобны срезки по KNV- и ε_V-окрестностям.
На рис. 9.4 показаны примеры сглаживания алгоритмами (9.2а) и

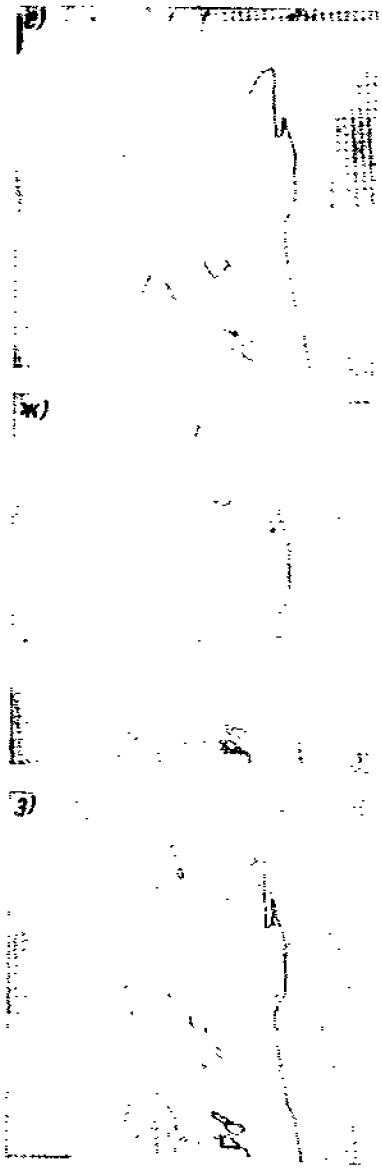
$$\hat{v}_{k,l} = \mathrm{MEAN}\,(\mathrm{KNV}\,(\mathrm{MED}\,(\mathrm{M}))). \quad (9.2в)$$

M-окрестность в последнем случае составляли элементы в KSK-окрестности 3x3 по отношению к центральному элементу.
На рис. 9.5 показан еще один пример сглаживания алгоритмом (9.2в). Сравнение рисунков 9.4, в и 9.4, д показывает, как в зависимости от параметров алгоритмов сглаживания меняется понятие шума: подавляется слабо коррелированная составляющая, которую можно отнести к шуму датчика видеосигнала (рис. 9.4,*в*), или, при более сильном сглаживании, подавляется составляющая видеосигнала, содержащая мелкие текстурные детали изображения (рис. 9.4,*д*).

Оценки центров выращивания в описанных алгоритмах можно получать итеративно, например, следующим образом:

$$\hat{v}_{k,l}^{(0)} = v_{k,l};$$
$$\hat{v}_{k,l}^{(i)} = \mathrm{MEAN}\,(\mathrm{KNV}\,(\hat{v}_{k,l}^{(i-1)})).$$

С ростом числа итераций такие алгоритмы будут давать результаты, близкие к квантованию алгоритмом адаптивного квантования мод, с границами мод в минимумах гистограммы.



Сглаживание для модели импульсных помех. Модель импульсных помех предполагает, что с некоторой вероятностью элемент сигнала заменяется случайной величиной. Сглаживание импульсного шума, очевидно, требует обнаружения искаженных элементов сигнала и последующего оценивания их значений по значениям неискаженных элементов. Вообще говоря, алгоритмы сглаживания импульсных помех должны быть двухпроходовыми, с разметкой искаженных элементов на первом проходе и оценкой их сглаженных значений на втором проходе. Но для упрощения можно сделать алгоритм однопроходовым, совмещая операции обнаружения и оценивания в одном проходе.

Разметка элементов изображения на искаженные шумом и не искаженные (обнаружение выбросов шума) может быть выполнена на основании проверки гипотезы о принадлежности централь-

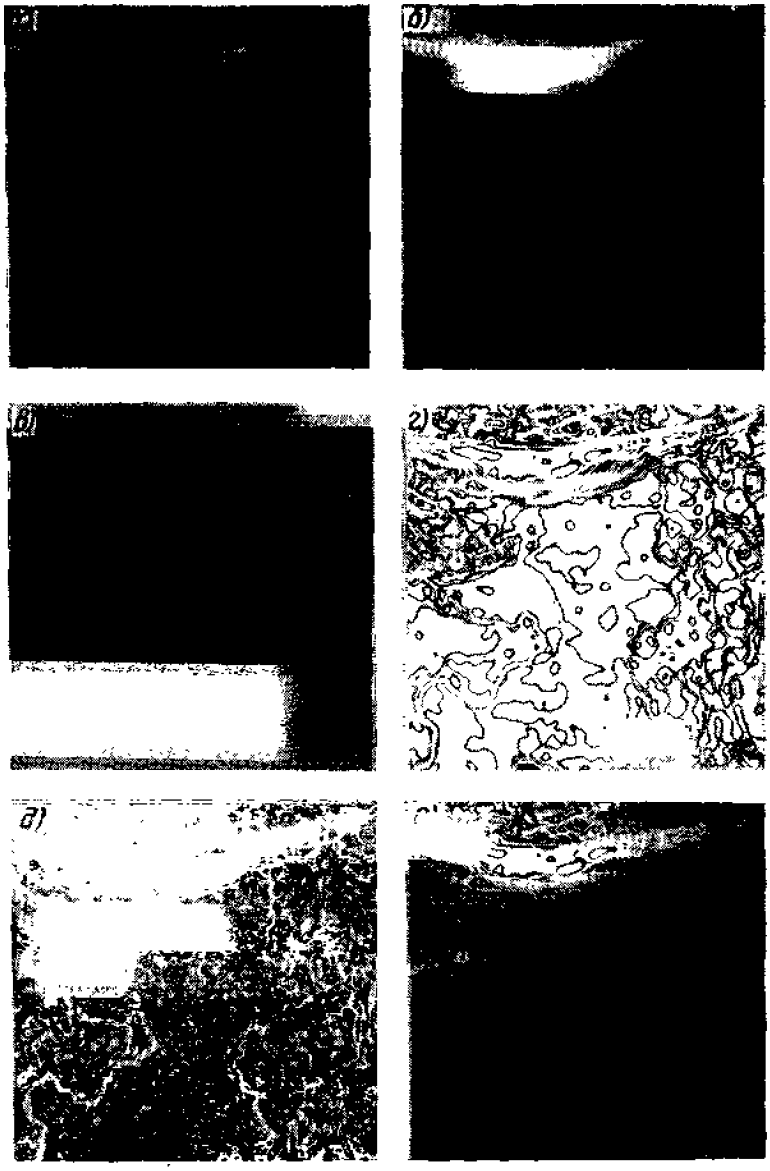


Рис. 9.5. Сглаживание изображений и выделение границ деталей:
а — исходное изображение; *б* — сглаживание алгоритмом (9.2в); S=15×15; M=3×3; K=60; *в* — разность изображений *а* и *б*, усиленная в 6 раз; *г* — выделение границ алгоритмом квазиразмаха на изображении *б*; *д* — то же на изображении *а*; *е* — исходное изображение с наложенными границами (сумма изображений *а* и *г*)

ного элемента S-окрестности той же выборке, что и заданное большинство остальных элементов окрестности, или выпадения ее из этой выборки. Это достаточно типичная задача математической статистики, и для ее решения обычно рекомендуются алгоритмы, основанные на ранговых статистиках.

Наиболее простым ранговым способом проверки гипотезы о принадлежности центрального элемента S-окрестности к выборке из большинства остальных элементов окрестности является голосование (см. также § 6.4), т.е. проверка попадания ранга $r_s(v_{k,l})$ в ε-окрестность медианы, задаваемую в зависимости от вероятности появления импульсных помех на элемент изображения: если

$$| r_S(v_{k,l}) - (N_S + 1)/2 | < \varepsilon_r,$$

то принимается решение об отсутствии помехи, в противном случае элемент *(k, l)* помечается как искаженный помехой. Такой способ обнаружения помехи предполагает, что импульсная помеха, как правило, принимает экстремальные значения. Отметим, что ранг как критерий проверки гипотезы о принадлежности элемента к данной выборке является частным случаем критерия Вилкоксона [24], проверяющего наличие сдвига между двумя выборками с одинаковым законом распределения.

Проверку гипотезы о наличии или отсутствии выброса помехи в центральном элементе S-окрестности можно производить также путем сравнения не по его рангу, а по его значению. Например, критерием может служить знак разности

$$\Delta = \varepsilon_V - | v_{k,l} - SMTH(v_{k,l}) |$$

где $\mathrm{SMTH}(v_{k,l})$ – сглаженное значение $v_{k,l}$, полученное одним из описанных выше алгоритмов сглаживания для аддитивной модели, а ε_v подбирается в зависимости от распределения значений помехи и разброса значений самого сигнала. Этот способ лучше согласован с особенностями изображений как сигналов, проявляющимися в том, что значения сигнала в геометрически соседних элементах изображений, как правило, близки друг к другу.

Порог ε_v может быть выбран сразу для всего изображения, но его можно и адаптивно подстраивать в зависимости от локального разброса значений сигнала. В качестве оценки локального разброса можно использовать, например, квазиразмах по ε_r -окрестности:

$$\mathrm{QDISP}(S) = v_S((N_S + 1)/2 + \varepsilon_r) - v_S(N_S + 1)/2 - \varepsilon_r),$$

являющийся, как известно [18], устойчивой к распределению оценкой разброса значений в выборке.

После этапа обнаружения элементы изображения, отмеченные как выбросы импульсного шума, должны быть заменены их оценкой. В качестве оценки можно использовать значения, полученные тем или иным сглаживанием по окрестности этих элементов, причем из этой окрестности исключаются элементы, отмеченные при обнаружении выбросов шума.

Таким образом, алгоритмы сглаживания импульсного шума могут быть двух типов:

$$\hat{v}_{k,l} = \begin{cases} v_{k,l}, & \varepsilon_r \geqslant | r_S(v_{k,l}) - (N_S + 1)/2 | ; \\ \mathrm{SMTH}(M), & \varepsilon_r < | r_S(v_{k,l}) - (N_S + 1)/2 | ; \end{cases}$$

$$\hat{v}_{k,l} = \begin{cases} v_{k,l}, & \varepsilon_V \geqslant | v_{k,l} - \mathrm{SMTH}(v_{k,l}) | ; \\ \mathrm{SMTH}(M), & \varepsilon_V < | v_{k,l} - \mathrm{SMTH}(v_{k,l}) | , \end{cases}$$

где SMTH(M) означает сглаживание по окрестности, из которой исключены точки, подлежащие исправлению. Пример фильтрации импульсных помех алгоритмом первого типа, в котором в качестве S-окрестности использовалась окрестность 3x3 элемента, а SMTH(M) =MED(S), показан на рис. 9.6.

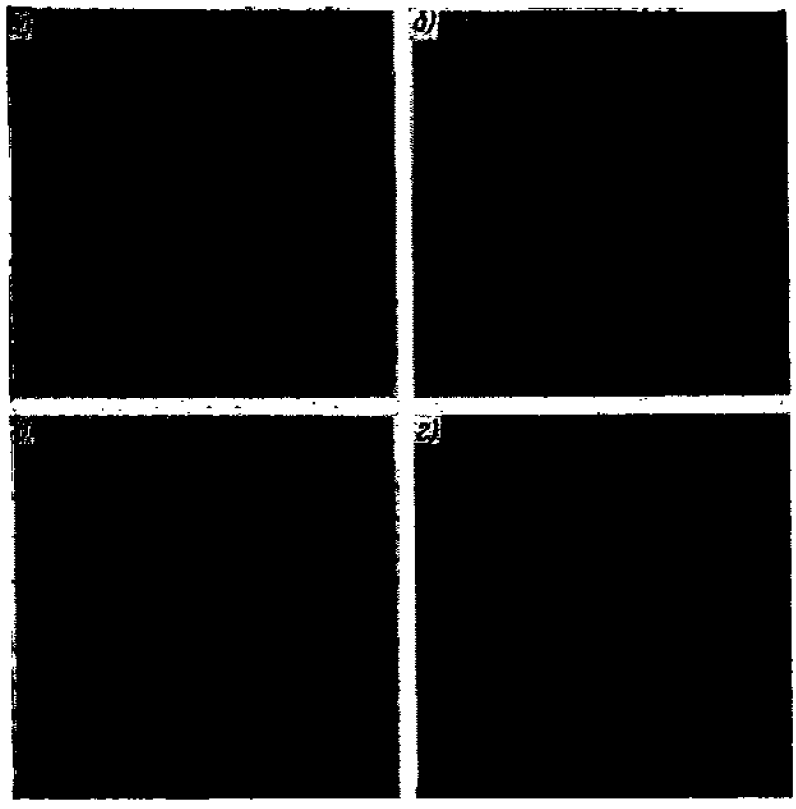


Рис. 9.6. Подавление импульсных помех алгоритмом с обнаружением методом голосования:

а. – зашумлевное изображение (вероятность искажения отсчетов 0,3); б – результат фильтрации по M=3x3, $\varepsilon_г$=3; в – то же при $e_г$=2; г –то же при $\varepsilon_г$=1

Характерными при сглаживании импульсных помех являются ошибки ложного обнаружения, которые приводят к нежелательному сглаживанию деталей изображения, и ошибки пропуска, из-за которых на изображении могут остаться несглаженные выбросы помехи. Доля этих ошибок зависит от порогов: с увеличением порогов доля ложных обнаружений падает, а доля пропусков возрастает. Следует учитывать, что число ложных обнаружений и пропусков возрастает также из-за возможного наличия в S-окрестности не одного, а нескольких больших выбросов помех. Поэтому для повышения качества сглаживания импульсных помех его целесообразно проводить итеративно, начиная с больших значений порогов, и по мере удаления больших выбросов помехи понижая пороги на каждом шаге итерации.

9.3. УВЕЛИЧЕНИЕ ДЕТАЛЬНОСТИ ИЗОБРАЖЕНИЙ

Увеличение детальности изображений – понятие, противоположное сглаживанию. Если при сглаживании стираются различия деталей изображения, то'при увеличении детальности они должны, наоборот, усиливаться. Поэтому увеличение детальности изображений называют также повышением локальных контрастов. Это, по существу, основная операция при препарировании изображений.

Повышение локальных контрастов достигается путем измерения отличий значения сигнала в каждом элементе изображения от его значений в элементах, окружающих данный, и усиления этих отличий.

Наиболее известный и очевидный метод определения и усиления отличий – так называемая нерезкая маска. При этом вычисляется разность между значениями элементов изображения и усредненными значениями по окрестности этих элементов, эта разность усиливается и добавляется к усредненному изображению:

$$\hat{v}_{k,l} = g\,(v_{k,l} - \bar{v}_{k,l}) + \bar{v}_{k,l},$$

где $\bar{v}_{k,l}$ – сумма элементов в S-окрестности, взятых с некоторыми весами; g–коэффициент усиления [ср. (8.12)]. Отметим, что из этой формулы вытекает возможность обобщения метода нерезкой маски на использование ранговых алгоритмов. Она заключается в том, чтобы вместо взвешенного среднего по S-окрестности (величины $\bar{v}_{k,l}$) использовать сглаженное значение SMTH(M) сигнала, полученное с помощью ранговых алгоритмов сглаживания, описанных в предыдущем разделе:

$$\bar{v}_{k,l} = g d_{k,l} + \mathrm{SMTH}\,(M), \quad (9.3)$$

где

$$d_{k,l} = v_{k,l} - \mathrm{SMTH}\,(M).$$

Преимущества нерезкого маскирования с ранговым сглаживанием вместо линейного– адаптивность и меньшая пространст-

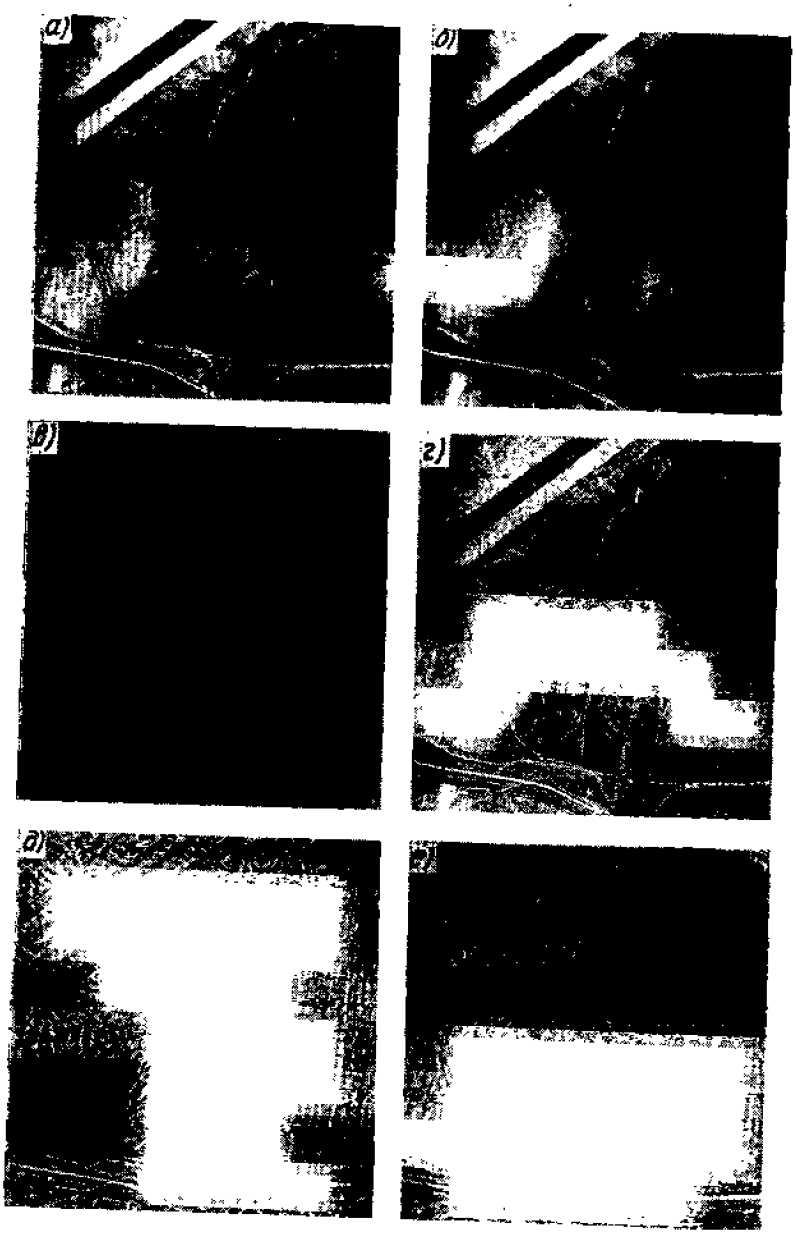


Рис. 9.7. Повышение детальности изображений: а- исходной изображение; б–результат сглаживания алгоритмом (9.2б); S=15x15; M=3x3; ε_V=10; в–разность изображений а и б, усиленная в 6 раз; г–сумма сглаженного изображения б и разностного изображения, усиленного в 3 раза; д–скользящая эквализация по окрестности S=15x15; е–взвешенная сумма исходного изображения а (вес 2/3) и изображения д (вес 1/3).

венная инерционность – вытекают из преимуществ рангового сглаживания. Пример работы такого алгоритма показан на рис. 9.7, *а–г*, и рис. 9.8, *а –в*.

Целью обработки изображения рис. 9.7 было усиление его текстуры на участках полей. Для сглаживания был использован алгоритм (9.26). Разностное изображение рис. 9.7, а показывает, какие элементы текстуры подверглись усилению.

Целью обработки изображения рис. 9.8 было увеличение контраста деталей изображения, меньших по размерам, чем ширина ребер на снимке. Сглаживание производилось алгоритмом:

$$SMTH(M) = MED(KNV(v_{k,l})).$$

Сглаженное изображение рис. 9.8, б и разностное рис. 9.8, в показывают, что при сглаживании практически не исказились границы ребер, а детали размером менее ширины ребер оказались подавленными. Это означает, что в результирующем изображении увеличился контраст только этих деталей.

Другой путь повышения локальных контрастов связан с использованием скользящей эквализации гистограмм и ее обобще-

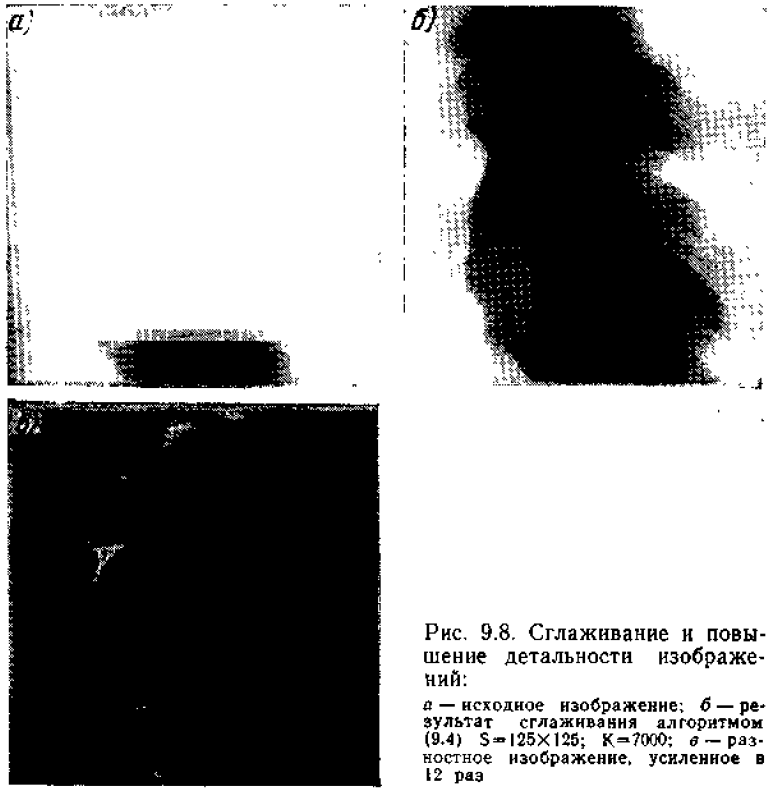
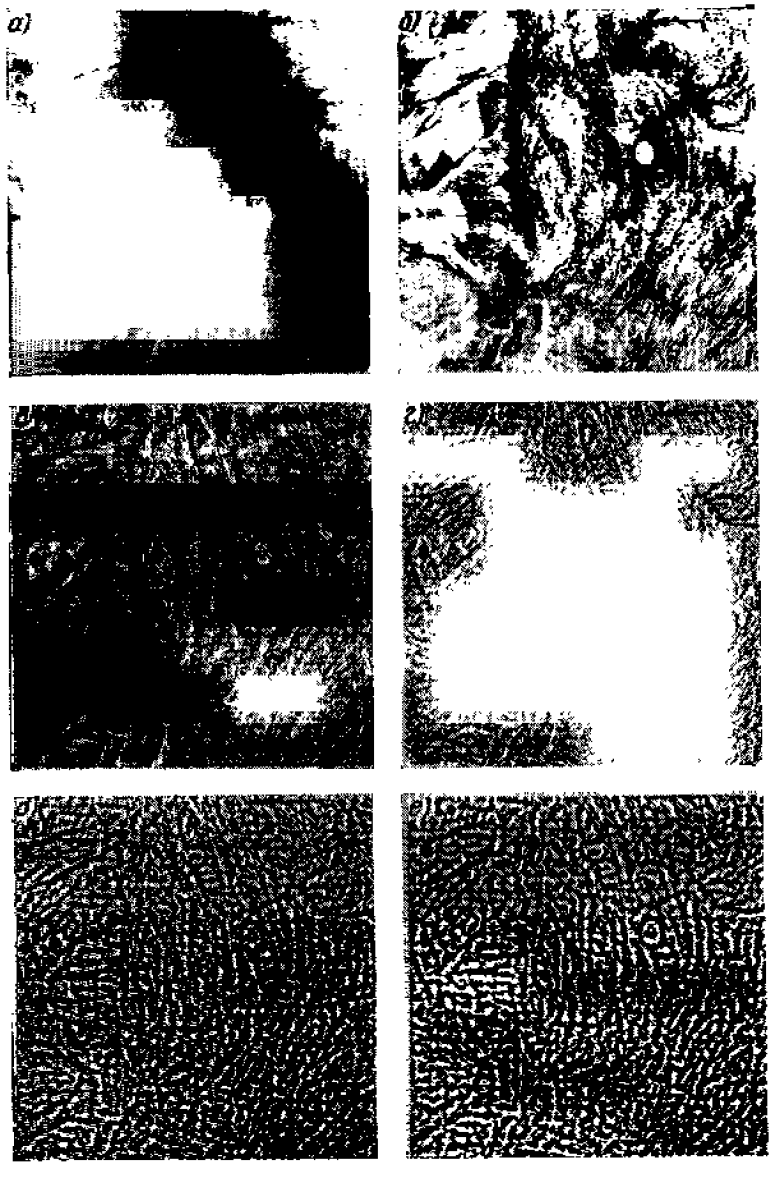


Рис. 9.8. Сглаживание и повышение детальности изображений:

а — исходное изображение; *б* — результат сглаживания алгоритмом (9.4) S=125×125; K=7000; *в* — разностное изображение, усиленное в 12 раз



ния – степенной интенсификации [3, 46]. При этом изображение подвергается поэлементному нелинейному преобразованию $F(v_{k,l})$ крутизна которого в точке $v_{k,l}$ пропорциональна некоторой степени p от значения локальной гистограммы в точке $q = v_{k,l}$:

$$\Delta F(v_{k,l})/\Delta v_{k,l} = (h_S(q))^p \mid_{q=v_{k,l}} \quad (9.4а)$$

Скользящая эквализация соответствует случаю p=1:

$$\Delta F(v_{k,l}) = h_S(q) \mid_{q=v_{k,l}} \quad (9.46)$$

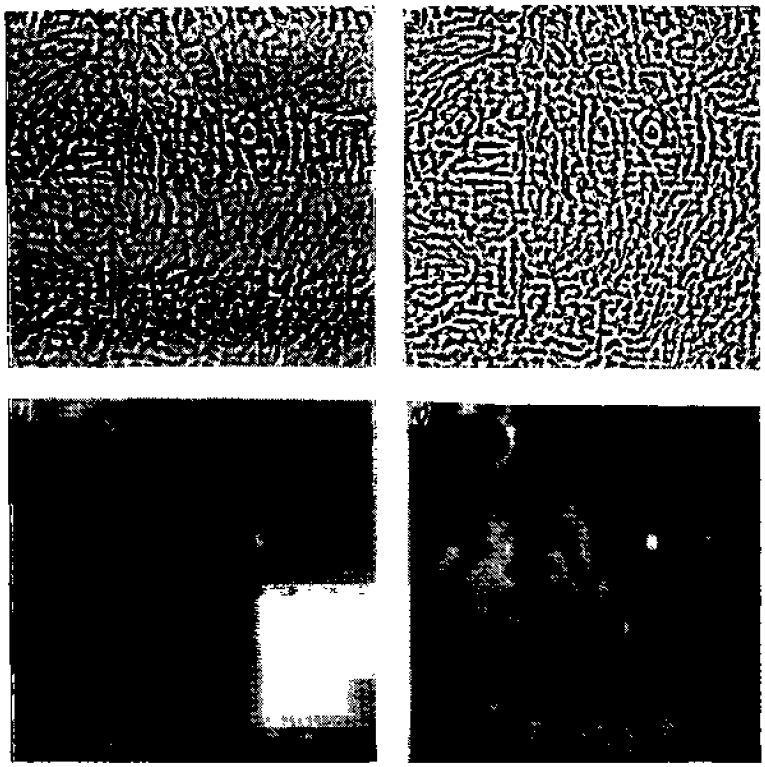


Рис. 9.9. Повышение детальности и выделение деталей изображений: а –исходное изображение; б–г – скользящая эквализация соответственно по S=125x125, S = 35x35 и S = 15x15; д– з – последовательные стадии обработки алгоритмами эквализации – сглаживания по S = 15x15, и – обнаружение деталей по критерию максимума отклонения локальных гистограмм; *к* –то же по критерию суммы модулей отклонения гистограмм

Решив это разностное уравнение, можно получить, что преобразование значений $v_{k,l}$ при эквализации определяется следующей формулой:

$$\hat{v}_{k,l} = F(v_{k,l}) = C_1 \sum_{q=0}^{v_{k,l}} h_S(q) + C_2 \quad . \quad (9.4в)$$

где C_1 – нормировочная константа; C_2 – константа смещения. В дальнейшем будем их опускать, имея в виду очевидность значений этих констант. Поэтому перепишем (9.48) в виде

$$\hat{v}_{k,l} = \sum_{q=0}^{v_{k,l}} h_S(q)$$

Из этой формулы вытекает, что скользящая эквализация есть не что иное, как замена значения видеосигнала $v_{k,l}$ его рангом в вариационном ряду, построенном по S-окрестности:

$$\hat{v}_{k,l} = r_S(v_{k,l}). \quad (9.5)$$

При степенной интенсификации ранг определяется не по гистограмме, а по значениям гистограммы, возведенным в степень p. В работе [5] показано, что при обработке медицинских и аэрокосмических изображений степенная интенсификация при значениях 0<p<1 зачастую дает лучшие результаты, чем чистая эквализация (p=1).

Степенная интенсификация – это, конечно, частный случай преобразования сигнала, крутизна которого пропорциональна некоторой функции от значения локальной гистограммы:

$$\Delta F(v_{k,l})/\Delta v_{k,l} = f(h_S(q)).$$

Такое преобразование можно назвать *f*-эквализацией. Практический интерес представляет вариант линейной *f*-эквализации, когда $f(h_S(q))$ – линейная функция:

$$f(h_S(q)) = g_1 h_S(q) + g_2. \quad (9.6)$$

В этом случае преобразованное изображение представляет собой взвешенную сумму эквализованного и исходного изображений:

$$\hat{v}_{k,l} = g_1 r_S(v_{k,l}) + g_2 v_{k,l}. \quad (9.7)$$

Примеры скользящей эквализации с разными размерами S-окрестностей показаны на рис. 9.9. Рис. 9.9, б и 9.9, в, г показывают, как размер деталей, контраст которых усиливается, зависит от размеров S-окрестности.

Пример «линейной» *f*-эквализации по (9.7) показан на рис. 9.6, где исходное (см. рис.9.6,а) и эквализованное изображения (см. рис. 9.7,д) сложены с весами 0,5 и 0,5.

«Ранговая» трактовка эквализации позволяет предложить некоторые ее естественные обобщения, связанные с изменением вида окрестности: эквализацию по $M_{\varepsilon V}$ и эквализацию по M_{KNV}-

При эквализации по $M_{\varepsilon V}$:

$$\hat{v}_{k,l} = r_{\varepsilon V}(v_{k,l}). \quad (9.8)$$

При эквализации по M_{KNV}.

$$\hat{v}_{k,l} = v_{KNV}(v_{k,l}). \quad (9.9)$$

Эквализация по ε_V или по KNV-окрестности позволяет избежать влияния деталей, принадлежащих к другому кластеру в S-окрестности. Различие эквализации по S и по $M_{\varepsilon V}$ иллюстрируется на модельном изображении, показанном на рис. 9.10. Видно, как при эквализации по $M_{\varepsilon V}$ границы квадрата оказались почти полностью подавленными, тогда как при эквализации по S они усиливаются наряду с усилением контраста мелкой текстуры.

Действие эквализации по S и эквализации по $M_{\varepsilon V}$ и M_{KNV} на реальном изображении можно сравнить на рис. 9.11. На этом рисунке видно, что эквализация по M_{KNV} дает результат, более близкий к скользящей эквализации по S. Объясняется это тем, что в данном случае при выбранных значениях К и ε_V размер М-окрестности, по которой происходит подсчет рангов, при эква лизации по M_{KNV} ближе ко всей S-окрестности, чем при эквализации по $M_{\varepsilon V}$ - Изменяя величины К и ε_V, можно варьировать преобразование от чистой эквализации (при K=Ns или ε_V=Q/2) до вырождения при К.=1 или ε_V=0.

Выбор между ε_V – и KNV-окрестностями, как и при сглаживании, определяется априорной информацией о разбросе значений сигнала и размерах деталей изображения, контраст которых подлежит усилению.

Другая возможность обобщения процедуры эквализации состоит в замене величины $v_{k,l}$ не ее рангом по той или иной окрестности, а некоторой функцией ранга:

$$\hat{v}_{k,l} = f(r_S(v_{k,l})). \quad (9.10)$$

Примером такой обработки может служить алгоритм *гиперболизации гистограмм* [64]:

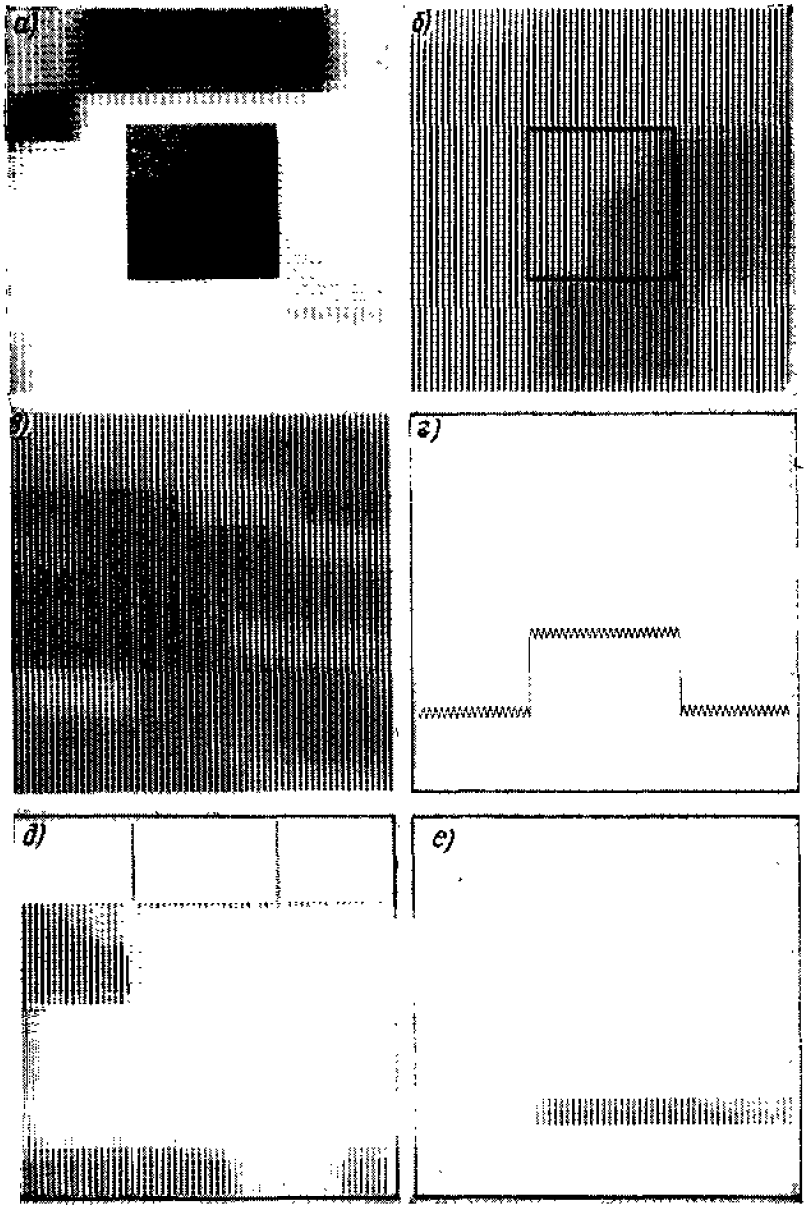
$$\hat{v}_{k,l} = \exp(r_S(v_{k,l})). \quad (9.11)$$

Действительно, при этом $\ln \hat{v}_{k,l}$ будет иметь равномерное распределение, что и требуется при гиперболизации.

Все сказанное об эквализации относится и к степенной интенсификации: возможна степенная интенсификация по $M_{\varepsilon V}$, по M_{KNV}, *f*-степенная интенсификация. Особый интерес представляет случай степенной интенсификации при p = 0. Тогда из уравнения (9.3) вытекает, что преобразованное значение сигнала находится следующим образом:

$$\hat{v}_{k,l} = v_{k,l} - \mathrm{MIN}(S) = v_{k,l} - v_S(r = 1). \quad (9.12)$$

Эта формула описывает известный алгоритм так называемой линейной коррекции видеосигнала [46].



9.10. Варианты скользящей эквализации:

ходное тестовое изображение; *б* — эквализация по S=15×15; *в* — эквализация по ркости; *г—е* — графики видеосигнала рис. *а—в* соответственно

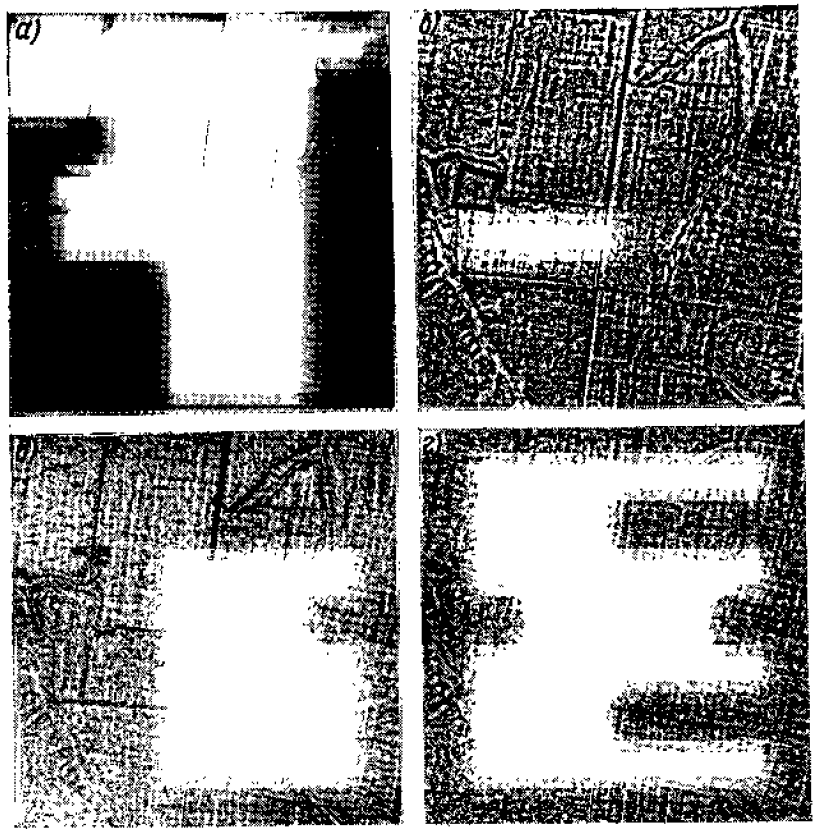


Рис. 9.11. Варианты скользящей эквализации:
а – исходное изображение; б –скользящая эквализация по 5=15x15; в – эквализацня по KNV-окрестиости (К=100, S=15x15); г – эквивализация по ε_V -окрестности (S=15x15; ε_V =10)

При степенной интенсификации (p = 0) по M_{KNV}:

$$\hat{v}_{k,l} = v_{k,l} - \mathrm{MIN}\,(\mathrm{KNV}\,(v_{k,l})). \quad (9.13)$$

Степенная интенсификация (p = 0) при M_ey

$$\hat{v}_{k,l} = v_{k,l} - \mathrm{MIN}\,(\varepsilon_V\,(v_{k,l}))$$

не имеет смысла: как правило, $\mathrm{MIN}\,(\varepsilon_V\,(v_{k,l})) = v_{k,l} - \varepsilon_V$, так что $\hat{v}_{k,l} = \varepsilon_V$ не зависит от $v_{k,l}$. Это заставляет вспомнить, что ε_V- и KNV-окрестности – это оценки положения кластера в вариационном ряду. При этом $v_{k,l}$ можно рассматривать как оценку «центра выращивания» кластера. Отсюда следует, что в формулах (9.4) – (9.13) в общем случае вместо $v_{k,l}$ должна входить сглаженная оценка SMTH(M), например

$$\hat{v}_{k,l} = r_S\,(\mathrm{SMTH}\,(M)); \quad (9.14)$$

$$\hat{v}_{k,l} = g_1 r_S\,(\mathrm{SMTH}\,(M)) + g_2\,\mathrm{SMTH}\,(M); \quad (9.15)$$

$$\hat{v}_{k,l} = v_{k,l} - \mathrm{MIN}\,(\mathrm{KNV}\,(\mathrm{SMTH}\,(M))); \quad (9.16)$$

$$\hat{v}_{k,l} = v_{k,l} - \mathrm{MIN}\,(\varepsilon_V\,(\mathrm{SMTH}\,(M))). \quad (9.17)$$

Таким образом, можно говорить о двух группах ранговых алгоритмов повышения локальных контрастов: разностных алгоритмах типа (9.3) с ранговым сглаживанием и чисто ранговых алгоритмах (9.8) –(9.17).

Формально алгоритмы (9.13), (9.16), (9.17) также можно отнести к разностным, хотя они являются как бы переходным мостиком между двумя группами. Но принципиального различия между этими группами, по-видимому, нет. Их родство просматривается, например, в алгоритме (9.17).

Разностные алгоритмы повышения детальности соответствуют представлению о деталях, аддитивно наложенных на фон. Лучше соответствует понятию деталей и фона в изображении представление о деталях, как бы врезанных в фон. Из него вытекает, что увеличению контраста деталей должно предшествовать их обнаружение, как в алгоритмах фильтрации импульсных помех. Этого можно достичь посредством переключательных алгоритмов вида:

$$\hat{v}_{k,l}=\begin{cases}\mathrm{SMTH}(M), & |d_{k,l}|\leqslant\varepsilon_V;\\ gd_{k,l}+\mathrm{SMTH}(M), & |d_{k,l}|>\varepsilon_V.\end{cases}$$

При усилении детальности путем повышения локальных контрастов не делается различия между деталями изображений, а отличие их от фоновой части изображения задается параметрами S-, ε_V, KNY-окрестностей. Поэтому увеличение детальности при обработке изображений обычно представляет собой только часть процедуры обработки, которую можно было бы назвать выделением деталей изображения, и в которой в том или ином виде закладываются параметры, описывающие одни детали как полезные, а другие как ненужные. Это означает, что результат повышения локальных контрастов подлежит последующему сглаживанию, устраняющему ненужные детали и оставляющему полезные.

В зависимости от конечной цели можно указать два вида алгоритмов выделения деталей изображения: выделение деталей с сохранением «фона» и отделение деталей от фона, т.е. представление их на постоянном фоне. Эти алгоритмы получаются сочетанием описанных алгоритмов повышения локальных контрастов и сглаживания в двухэтапной процедуре повышения локальных контрастов по некоторой Sc -окрестности и сглаживания по некоторой Ss -окрестности. Ss -окрестность, как правило, составляет часть Sc-окрестности. Приведем несколько примеров таких процедур.

Процедуры выделения деталей с сохранением фона:

$$\hat{v}_{k,l}=g_1\,\mathrm{SMTH}(\mathrm{S}_g(f(r_{\mathrm{S}_C}(\mathrm{SMTH}(\mathrm{S}_C)))))+g_2\,\mathrm{SMTH}(\mathrm{S}_C);\tag{9.18}$$

$$\hat{v}_{k,l}=\mathrm{SMTH}(\mathrm{S}_g(v_{k,l}-\mathrm{MIN}(\mathrm{KNV}(\mathrm{SMTH}(\mathrm{S}_C))))));\tag{9.19}$$

$$\hat{v}_{k,l}=\mathrm{SMTH}(\mathrm{S}_S(gd_{k,l}+\mathrm{SMTH}(S_C))).\tag{9.20}$$

Процедуры с отделением деталей от фона:

$$\hat{v}_{k,l}=\mathrm{SMTH}(\mathrm{S}_S(f(r_{\mathrm{S}_C}(\mathrm{SMTH}(\mathrm{S}_C))))),$$

$$\hat{v}_{k,l}=\mathrm{SMTH}(\mathrm{S}_g(d_{k,l}))=\mathrm{SMTH}(\mathrm{S}_g(v_{k,l}-\mathrm{SMTH}(\mathrm{S}_C))).\tag{9.21}$$

Пример сглаживания эквализованных изображений алгоритмом $\hat{v}_{k,l}=\mathrm{MED}(\mathrm{KNV}(\mathrm{MED}(M)))$ показан на рис. 9.9.

Отметим, что некоторые из описанных процедур выделения деталей в частных случаях совпадают с процедурами, используемыми в математической статистике при проверке гипотез. Так, (9.21) в частном случае $\hat{v}_{k,l}=\mathrm{MEAN}(\mathrm{S}_S(r_{\mathrm{S}_C}(v_{k,l})))$ совпадает с известным критерием Вилкоксона {18].

В случае $f(r)=r^t$, где t – показатель степени, получаем статистику критерия Тамуры [19]:

$$\hat{v}_{k,l}=\mathrm{MEAN}(\mathrm{S}_S(r^t_{\mathrm{S}_C}(v_{k,l}))).$$

В случае $f(r)=\mathrm{sign}\ (r_{\mathrm{S}_C}(v_{k,l})-(N_{\mathrm{S}_C}+1)/2)$ получаем так называемый медианный критерий [18]:

$$\hat{v}_{k,l}=\mathrm{MEAN}(\mathrm{S}_g(\mathrm{sign}\,(r_{\mathrm{S}_C}(v_{k,l})-(N_{\mathrm{S}_C}+1)/2))).$$

Рассмотрим теперь некоторые естественные обобщения описанных ранговых алгоритмов повышения локальных контрастов и выделения деталей изображений.

Нетрудно показать, что линейные алгоритмы повышения локальных контрастов (9.3) можно реализовать (см. § 4.3) с помощью U параллельных простейших фильтров, вычисляющих локальное среднее:

$$\hat{V}_{k,l}=g\left(v_{k,l}-\sum_{u=1}^{U}g_u\,\mathrm{MEAN}(\mathrm{S}_u)\right)+\sum_{u=1}^{U}g_u\,\mathrm{MEAN}(S_u),$$

где $\{g_u\}$ – веса, подбираемые для достижения наилучшего эффекта фильтрации.

Аналогия между (9.3) и (9.7), а также (9.15) позволяет предложить следующий более общий параллельный ранговый алгоритм повышения локальных контрастов с помощью эквализации:

$$\hat{v}_{k,l} = \sum_{u=1}^{U} g_u r_{S_u}(\mathrm{SMTH}(\mathrm{M})),$$

где r_{S_u} – ранг, вычисляемый по окрестности S_u. Такое суммирование с весами результатов эквализации по окрестностям разных размеров позволяет сбалансировать степень усиления контрастов для деталей разных размеров.

Другая возможность обобщения ранговых алгоритмов выделения деталей вытекает из общих формул (9.18) – (9.20), где появляются пары процедур «сглаживание – повышение локальных контрастов», которые естественно продолжить итеративно. За счет эффектов квантования сигнала последовательное повторение этих пар достаточно быстро сходится к бинарному препарату. Примеры полученных такой обработкой изображений показаны на рис. 9.9, е–з.

9.4. ОБНАРУЖЕНИЕ ДЕТАЛЕЙ И ИХ ГРАНИЦ

Связь алгоритмов повышения локальных контрастов и выделения деталей со статистиками ранговых критериев, а также очевидная их аналогия с алгоритмами обнаружения выбросов импульсных помех, рассмотренными выше, проливает новый свет на смысл этих алгоритмов. Эти алгоритмы можно трактовать как алгоритмы проверки гипотезы о несоответствии центрального элемента Sc -окрестности выборке, определяемой подмножеством по ε_V- или KNV-окрестности элементов Sc -окрестности, а вычисляемую ими оценку сигнала как критерий верности этой гипотезы, распределение значений которого по площади обрабатываемого изображения представляется пользователю как изображение-препарат.

Такая трактовка ведет к обобщению алгоритмов выделения деталей на задачу обнаружения деталей и их границ. В описанных алгоритмах использовались простейшие точечные критерии несоответствия элемента изображения заданной выборке: в разностных алгоритмах – разность между значением центрального элемента Sc -окрестности и оценкой среднего значения заданной выборки; в ранговых алгоритмах – количество элементов заданной выборки, не превышающих по своему значению значение центрального элемента, т.е. ранг центрального элемента в заданной выборке. Степень несоответствия трактовалась как контраст детали.

Для рангового обнаружения деталей изображения и их границ нужно, очевидно, измерять степень статистического несоответствия распределения значений элементов анализируемой окрестности заданному распределению значений сигнала в пределах деталей. При этом размер окрестности выбирают порядка размеров деталей, которые необходимо обнаруживать, или соответственно порядка размеров окрестности границ деталей. Для измерения степени несоответствия можно использовать известные в математической статистике критерии согласия.

Само по себе обнаружение состоит в сравнении измеренной степени соответствия с порогом. При препарировании изображений имеет смысл также предъявлять для визуализации саму величину соответствия, а не только бинарный результат сравнения с порогом. При этом обнаружение осуществляется оператором визуально. На рис. 9.9,*и*, *к* показан пример обнаружения деталей на изображении путем сравнения локальных гистограмм по критерию максимума отклонения и суммы модулей отклонения. В качестве опорного объекта было взято овальное серое пятно в правой верхней четверти снимка.

Ранговые алгоритмы обнаружения, основанные на сравнении гистограмм значений сигнала, нечувствительны к пространственному «перепутыванию» элементов изображения. Но пространственное «перепутывание» не входит, как правило, в число возможных искажений изображений в оптических и аналогичных изображающих системах, и поэтому опасность спутать при обнаружении деталь с последовательностью независимых отсчетов, имеющих то же распределение значений, что и распределение значений отсчетов сигнала на детали, маловероятна. В то же время ранговые алгоритмы устойчивы к таким распространенным искажениям сигнала, как монотонные изменения их значений при амплитудных искажениях, засорение распределений, изменения ориентации.

Приведем несколько примеров возможных ранговых алгоритмов обнаружения границ деталей. Признаком границы для кусочно-постоянной модели является двухмодовость локальной гистограммы. В качестве критерия двухмодовости можно использовать характеристики разброса локальной гистограммы, например, такую как квазиразмах: $\hat{v}_{k,l} = QDISP(S)$, когда ε_r -окрестность выбирается достаточно небольшой. В качестве практической рекомендации можно предложить выбирать $\varepsilon_r = \sqrt{N_S}$

Более общим является алгоритм вычисления квазиразмаха по М-окрестности. Примером может служить алгоритм вида:

$$\hat{v}_{k,l} = \mathrm{MAX}(\varepsilon_r(\mathrm{SMTH}(M))) + \mathrm{MIN}(\varepsilon_r(\mathrm{SMTH}(M))), \quad (9.22)$$

который, в принципе, способен выделять более тонкие границы. При использовании этих алгоритмов выделения границ целесообразно предварительно подвергать изображение сглаживанию для улучшения соответствия изображения кусочно-постоянной модели.

Пример выделения границ деталей алгоритмом квазиразмаха показан на рис. 9.4, е. Для сравнения на рис. 9.4, ж приведен результат обработки того же изображения алгоритмом, вычисляющим локальную дисперсию:

$$\hat{v}_{k,l} = (\mathrm{MEAN}(\mathrm{S}(v_{k,l} - \mathrm{MEAN}(\mathrm{S}))^2))^{1/2}.$$

На этом примере видно, что такое пространственная инерционность линейных алгоритмов и насколько она преодолевается ранговыми алгоритмами. Пример работы алгоритма типа (9.22) при-иеден на рис. 9.4,з. Еще один пример использования алгоритма квазиразмаха для очерчивания границ деталей, иллюстрирующий роль сглаживания перед выделением границ, показан на рис. 9.5,*г–е*.

9.5. ДРУГИЕ ПРИМЕНЕНИЯ РАНГОВЫХ АЛГОРИТМОВ

Кроме применений для сглаживания, усиления детальности, выделения деталей изображений и границ деталей, ранговые алгоритмы могут употребляться также для решения многих других более частных задач обработки изображений. Из них можно упомянуть диагностику искажений видеосигнала и определение их статистических характеристик, стандартизацию изображений, определение статистических характеристик самого видеосигнала и измерение текстурных признаков.

Автоматическая диагностика параметров помех и искажений видеосигнала. Она может основываться на принципе обнаружения и измерения аномалий в статистических характеристиках видеосигнала [47] (см. § 6.4). Для обнаружения аномалий можно использовать ранговые алгоритмы, такие как алгоритм голосования проверки принадлежности анализируемого элемента выборки к заданному числу крайних (наибольших или наименьших) значений упорядоченной выборки.

Стандартизация изображений. Стандартизация – это приведение характеристик изображений к некоторым заданным. С помощью ранговых алгоритмов может быть достаточно просто осуществлена стандартизация гистограмм, т.е. преобразование видеосигнала, делающее гистограмму распределения его значений заданной. Это можно осуществить с помощью следующего алгоритма: $\hat{v}_{k,l} = v_{\mathrm{ST}}(r_S(v_{k,l}))$, где $v_{\mathrm{ST}}(r)$ – значение r-го по рангу элемента вариационного ряда, построенного по заданной (стандартной) гистограмме. В зависимости от задачи могут использоваться глобальная и локальная стандартизация гистограмм. В первом случае $v_{\mathrm{ST}}(r)$ определяется по гистограмме всего стандартного изображения, а во втором случае – по гистограмме скользящего фрагмента стандартного изображения. В качестве стандартной может использоваться не вся гистограмма стандартного изображения или его локальные гистограммы, а соответствующие гистограммы по ε_V, KNV-, ε_V-окрестности. Стандартизация среднего арифметического и дисперсии изображения может проводиться алгоритмом

$$\hat{v}_{k,l} = \frac{g_1\sigma}{g_1\sigma_{k,l} + \sigma}\{v_{k,l} - \mathrm{MEAN}(\mathrm{S})\} + g_2 m + (1 - g_2) \times$$
$$\times \mathrm{MEAN}(\mathrm{S}),$$

где σ, m–эталонные среднеквадратическое отклонение и среднее арифметическое; g_1, g_2– задаваемые константы. В [70] предлагается использовать вместо среднего арифметического и средне-квадратического отклонения медиану и межквартильное расстояние. Как и стандартизация гистограмм, стандартизация параметров сдвига и разброса может быть сделана как глобальной, так и локальной. В первом случае эталонные параметры постоянны по всему изображению, во втором случае они вычисляются по скользящему фрагменту эталонного изображения.

Определение статистических характеристик видеосигнала и измерение текстурных признаков. Адаптивные свойства ранговых алгоритмов делают их удобным инструментом для измерения локальных статистических характеристик изображений: локального среднего, локальной дисперсии и других моментов распределения. Очевидно, что эти и другие подобные характеристики гистограмм являются также текстурными характеристиками изображений.

Ранговые алгоритмы могут служить для оценки не только гистограммных текстурных признаков, но и для оценки текстурных признаков, связанных с локальными пространственными статистическими характеристиками изображений. Одним из простейших признаков такого рода является число локальных экстремумов в S-окрестности обрабатываемого элемента. Ряд текстурных признаков связан с характеристиками пространственного распределения локальных экстремумов, т.е. среднего расстояния между ними, дисперсии расстояний между ними и т.д.

Более общими являются признаки, характеризующие пространственное распределение рангов в обрабатываемом фрагменте. В частности, текстурным признаком является число перемен знака первой производной по фрагменту эквализованного изображения в заданном направлении сканирования. Ряд текстурных признаков можно предложить как параметры пространственного распределения элементов, принадлежащих ε_V и KNV-окрестностям, описанным выше, в частности моменты распределения взаимных расстояний между ними.

Кодирование изображений. Возможность применения ранговых алгоритмов для кодирования изображений связана с использованием алгоритмов адаптивного квантования мод в режиме пофрагментной обработки. В этом случае анализируется гистограмма распределения значений элементов изображения в пределах фрагмента (или, как принято говорить в кодировании, блока), находятся границы кластеров, которые выбираются в качестве границ интервалов квантования, и производится квантование всех отсчетов фрагмента в соответствии с найденными границами. Как правило, при этом, если размеры фрагмента не слишком велики, количество уровней квантования Q_S отсчетов фрагмента намного меньше количества Q уровней квантования, выбираемого из условия качественного воспроизведения всего изображения. Нетрудно подсчитать, что количество бит, требуемых для передачи значений N_S отсчетов фрагмента, будет равно сумме $Q_S \log_2 Q$ бит на передачу таблицы квантования и $N_S \log_2 Q_S$ бит на передачу номера уровня квантования, т.е. на один отсчет изображения требуется в среднем $\log_2 Q_S + (Q_S \log_2 Q)/N_S$ бит вместо $\log_2 Q$ без адаптивного квантования по фрагментам. Отсюда вытекает, что площадь фрагментов целесообразно увеличивать до тех пор, пока количество уровней квантования Q_s не превысит нескольких единиц. Опыты, проведенные по пофрагментному квантованию мод, показывают, что это возможно при размерах фрагмента до 30x30 элементов. Следовательно, оценкой потенциальных возможностей кодирования изображений этим методом является величина порядка 1–2 бит на элемент.

Глава 10

СИНТЕЗ ГОЛОГРАММ

10.1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Рассмотрим математическую модель синтеза голограмм, пользуясь схемой визуального наблюдения объектов, приведенной на рис. 10.1. Положение наблюдателя относительно наблюдаемого объекта определяется на этом рисунке поверхностью наблюдения, на которой располагаются глаза наблюдателя, множество ракурсов рассматривания – углом охвата объекта. Чтобы наблюдатель мог видеть объект в заданном угле охвата, достаточно воспроизвести на поверхности наблюдения с помощью голограммы распределение интенсивности и фары световой волны, рассеянной объектом.

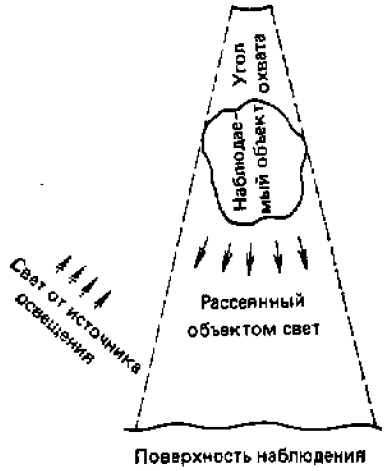


Рис. 10.1. Схема визуального наблюдения объектов

Для простоты будем рассматривать монохроматическое освещение объекта. Это позволит воспользоваться для описания процессов преобразования световых волн понятием комплексной амплитуды волны. Хотя процесс взаимодействия излучения с телом при отражении от его поверхности является сложным, для наших целей свойства объекта, определяющие его способность отражать и рассеивать падающее на него излучение, можно описать коэффициентами отражения излучения по интенсивности B(*x, y, z*) или амплитуде *b*(*x, y, z*), где *x, y, z* – координаты на поверхности объекта. Эти коэффициенты связывают интенсивность волны I_0(.x, y, z) и ее комплексную амплитуду A_0 (x, y, z) в точке (x, y, z) с интенсивностью I(x, y, z) падающей волны и ее амплитудой A (x, y, z):

(10.1)

$$I_0(x, y, z) = B(x, y, z)\, I(x, y, z);$$

$$A_0(x, y, z) = b(x, y, z)\, A(x, y, z). \quad (10.2)$$

Коэффициент отражения по амплитуде может рассматриваться как комплексная функция, которая может быть представлена в виде:

$$b(x, y, z) = |\, b(x, y, z)\,| \exp[i\rho(x, y, z)]. \quad (10.3)$$

Ее модуль $|b|$ и фаза ρ показывают, как изменяются модуль амплитуды A и фаза ω световой волны в точке (x, y, z) поверхности тела после отражения:

(10.4)

$$|\, A_0(x, y, z)\,| = |\, A(x, y, z)\,|\ |\, b(x, y, z)\,|;$$

$$\omega_0(x, y, z) = \omega(x, y, z) + \rho(x, y, z), \quad 10.5)$$

где

(10.6)

$$A_0(x, y, z) = |A_0(x, y, z)| \exp[i\omega_0(x, y, z)],$$
$$A(x, y, z) = |A(x, y, z)| \exp[i\omega(x, y, z)]. \quad (10.7)$$

В соответствии с (10.1) – (10.7) коэффициент отражения по интенсивности может быть определен через коэффициент отражения по амплитуде по формуле:

$$B = |b|^2 = bb^*. \quad (10.8)$$

Как известно, связь комплексной амплитуды $\Gamma(\xi, \eta, \zeta)$ поля световой волны на произвольной поверхности наблюдения, заданной в координатах (ξ, η, ζ) с комплексной амплитудой A_0 поля на поверхности объекта, можно описать интегральным соотношением [13, 27]:

$$\Gamma(\xi, \eta, \zeta) = \iiint_{S(x, y, z)} A_0(x, y, z) T(x, y, z; \xi, \eta, \zeta)\, dxdydz, \quad (10.9)$$

вид ядра $T(x, y, z; \xi, \eta, \zeta)$ которого зависит от пространственного расположения объекта и поверхности наблюдения, а интегрирование производится по поверхности объекта S(x, y, z). Это преобразование в принципе обратимо:

$$A_0(x, y, z) = \iiint_{0(\xi, \eta, \zeta)} \Gamma(\xi, \eta, \zeta) \overline{T}(\xi, \eta, \zeta; x, y, z)\, d\xi d\eta d\zeta, \quad (10.10)$$

где $\overline{T}$–ядро, взаимное Т, а интегрирование производится по поверхности наблюдения.

Функцию $\Gamma(\xi, \eta, \zeta)$ можно назвать математической голограммой. Задача синтеза голограммы заключается, таким образом, в вычислении функции $\Gamma(\xi, \eta, \zeta)$ по заданным из описания объекта и условий освещения функциям *b*(x, y, z) к A(x, y, z) и записи результата на физический носитель в форме, которая допускала бы взаимодействие с излучением для визуализации или восстановления A_0(x, y, z) в соответствии с (10.10).

Вычисление интегралов вида (10.10) является в общем случае сложной задачей. Но, учитывая естественные ограничения процесса визуального наблюдения, ее можно существенно упростить. Эти ограничения состоят в следующем:

1) размеры зрачка глаза наблюдателя намного меньше расстояния от поверхности наблюдения;
2) участки поверхности наблюдения размером в межзрачковое расстояние глаз можно считать плоскими;
3) глубина рельефа объектов, расположенных на удобном для рассматривания расстоянии от наблюдателя, обычно мала по сравнению с этим расстоянием.

Они позволяют прежде всего свести трехмерную задачу к двумерной. Для этого поверхность наблюдения можно разбить на участки, аппроксимируемые плоскостями, а распределение амплитуды и фазы волны на поверхности объекта заменить, пользуясь законами геометрической оптики, распределением амплитуды и фазы волны на плоскости, касающейся объекта (или просто достаточно близкой к нему, чтобы при пересчете амплитуды и фазы волны можно было пренебречь дифракцией) и параллельной данному плоскому участку поверхности наблюдения. Тем самым задача синтеза голограммы для всей поверхности наблюдения сводится к синтезу фрагментарных голограмм для плоских участков этой поверхности, и эта полная голограмма составляется, как мозаика, из фрагментарных голограмм. При этом упрощении для фрагментарной голограммы вместо (10.9) имеем:

$$\Gamma(\xi, \eta) = \iint_{(x, y)} \overline{A}_0(x, y)\, T_D(x, y; \xi, \eta)\, dxdy, \quad (10.11)$$

где $\overline{A}_0(x, y) = |\tilde{A}_0(x, y)| \exp[i\bar{\omega}_0(x, y)]$ – комплексная функция, полученная в результате пересчета амплитуды и фазы поля, отраженного объектом, на плоскость (x, y), касательную к нему и параллельную плоскости наблюдения (ξ, η); D – расстояние между этими плоскостями; $T_D(x, y; \xi, \eta)$–ядро преобразования.

Если геометрические размеры тела малы по сравнению с расстоянием D до плоскости наблюдения, то это вместе с условием малости площади участка поверхности наблюдения позволяет для вычисления интеграла (10.11) использовать его аппроксимацию интегралом Френеля [27]:

$$\Gamma(\xi,\ \eta) \approx \Gamma_{\Phi P}(\xi,\ \eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{A}_0(x,\ y) \exp\left\{ i \frac{\pi}{\lambda D} [(x-\xi)^2 + (y-\eta)^2] \right\} dx dy, \qquad (10.12)$$

где λ – длина волны излучения.

Голограммы, синтезированные в соответствии с этим соотношением, будем называть *синтезированными голограммами Френеля*.

Дальнейшее упрощение возможно, если

$$\pi (x^2 + y^2)/\lambda D \ll 1. \qquad (10.13)$$

Тогда интеграл (10.12) переходит в выражение

$$\Gamma_{\Phi P}(\xi,\ \eta) \approx \exp\left[i \frac{\pi}{\lambda D} (\xi^2 + \eta^2) \right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{A}_0(x,\ y) \times \times \exp\left[-i 2\pi \left(\frac{x\xi + y\eta}{\lambda D} \right) \right] dx dy, \qquad (10.14)$$

являющееся с точностью до экспоненциального фазового множителя преобразованием Фурье функции $\overline{A}_0(x,\ y)$:

$$\Gamma_{\Phi}(\xi,\ \eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{A}_0(x,\ y) \exp\left[-i 2\pi \left(\frac{x\xi + y\eta}{\lambda D} \right) \right] dx dy. \qquad (10.15)$$

Голограммы, синтезированные с использованием преобразования Фурье, будем называть *синтезированными голограммами Фурье*, или *фурье-голограммами*. Нетрудно видеть, что голограмма Френеля – это фурье-голограмма того же объекта, но наблюдаемого через линзу, и записанная также после линзы. Эту вторую линзу, описываемую множителем $\exp\{i\pi(\xi^2+\eta^2)/\lambda D\}$ в (10.14), можно не включать в процесс синтеза, так как в нее входит единственный общий для всей голограммы параметр D. Его можно воссоздать отдельно в процессе восстановления голограммы.

С точки зрения процесса восстановления волнового фронта объекта голограмма Френеля отличается от голограммы Фурье тем, что она в принципе обладает фокусирующими свойствами и способна воспроизвести конечное расстояние до объекта. Это значит, что если для восстановления голограмм Френеля использовать плоский волновой фронт от источника когерентного света, то она образует на расстоянии D сфокусированное изображение, определяемое функцией $\overline{A}_0(x,\ y)$. Голограммы Фурье восстанавливают объект, расположенный как бы на бесконечности, если для восстановления используется плоский волновой фронт, или в месте расположения источника освещения, если для восстановления используется сферический волновой фронт от точечного источника.

Математически восстановление объекта по голограммам Френеля и Фурье описывается обратными преобразованиями Френеля и Фурье соответственно. При визуальном наблюдении голограмм эти преобразования выполняются оптической системой глаза.

Перейдя от пространственной задачи к плоской, мы, строго говоря, потеряли возможность точного учета влияния глубины и рельефа объекта на волновой фронт в месте наблюдения. Даже в голограмму Френеля входит только расстояние от объекта до плоскости наблюдения, а не глубина рельефа объекта. Тем не менее остается возможность синтезировать поле, восстанавливающее в определенных условиях объект, а значит, остается наиболее важное свойство голографической визуализации – естественность наблюдения объекта. При этом передача рельефа, как показано в § 10.5, может осуществляться либо за счет изменения ракурсов при синтезе фрагментарных голограмм и восприниматься благодаря стереозрению, либо путем имитации диффузных свойств отражающей поверхности объекта.

10.2. ДИСКРЕТНОЕ ПРЕДСТАВЛЕНИЕ ГОЛОГРАММ ФУРЬЕ И ФРЕНЕЛЯ

Соотношения (10.12) и (10.15) являются исходными для синтеза голограмм Фурье и Френеля. Способ их реализации в цифровых процессорах в дискретной форме зависит от способа дискретного описания объекта и поля на объекте $\overline{A}_0(x,\ y)$, а также дискретного

представления самой голограммы. Наиболее простым и естественным способом является представление объекта и голограммы отсчетами, взятыми на прямоугольном растре с некоторым шагом $\Delta x, \Delta y;\ \Delta\xi, \Delta\eta$ по координатам (x, y) и (ξ, η) в соответствии с теоремой отсчетов. Обозначим эти матрицы $\bar{A}_0(k, l)$ и $\Gamma_\Phi(r, s)$ соответственно. Переход от $\bar{A}_0(k, l)$ к $\bar{A}_0(x, y)$ и от $\Gamma_\Phi(r, s)$ к $\Gamma_\Phi(\xi, \eta)$ осуществляется интерполяцией в аналоговых устройствах записи и восстановления голограмм.

Дискретное представление голограмм Фурье. Интерполяция математически может быть описана как свертка дискретного сигнала

$$A_д(x, y) = \sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1} \bar{A}_0(k, l)\,\delta\{[x-(k+u)\Delta x], [y-(l+v)\Delta y]\} \tag{10.16}$$

с некоторой интерполирующей функцией $h(\cdot, \cdot)$. Здесь $\delta(\cdot)$–дельта-функция Дирака, u и v – параметры, определяющие сдвиг растра отсчетов относительно системы координат (x, y), а количество отсчетов N_1 и N_2 определяются размерами объекта ($-X_{max}$, X_{max}, $-Y_{max}$, Y_{max}) и шагом дискретизации:

$$N_1 = \mathrm{int}(2X_{max}/\Delta x), \quad N_2 = \mathrm{int}(2Y_{max}/\Delta y). \tag{10.17}$$

Согласно теореме отсчетов точная интерполяция $\bar{A}_0(x, y)$ по $\bar{A}_0(k, l)$ возможна, если

$$\Delta x = \lambda D/2\xi_{max};\ \Delta y = \lambda D/2\eta_{max}, \tag{10.18}$$

где $(\pm\xi_{max}/\lambda D;\ \pm\eta_{max}/\lambda D)$ – прямоугольная область, за пределами которой пространственный спектр Фурье функции $\bar{A}_0(x, y)$ в координатах (ξ/λD; η/λD) можно считать равным нулю.

В соответствии с этим представлением достаточно синтезировать голограмму дискретного объекта $\bar{A}_д(x, y)$ по матрице отсчетов $\bar{A}_0(k, l)$; исходный же непрерывный объект $A_0(x, y)$ можно восстановить по $\bar{A}_д(x, y)$ аналоговым путем на этапе реконструкции синтезированной голограммы. Подставив (10.16) и (10.18) в (10.15), получим, что фурье-голограмма дискретного объекта $\bar{A}_д(x, y)$ может быть вычислена в виде конечной суммы

$$\Gamma_{\Phi, д}(\xi, \eta) = \left\{\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1} \bar{A}_0(k, l)\exp\left\{i2\pi\left[\frac{(k+u)\xi}{2\xi_{max}} + \frac{(l+v)\eta}{2\eta_{max}}\right]\right\}\right\} H(\xi, \eta), \tag{10.19}$$

где $H(\xi, \eta)$ –преобразование Фурье интерполирующей функции h (x, y].

Обозначим сумму в (10.19) $\hat{\Gamma}_{\Phi,д}(\xi, \eta)$. Поскольку объект $\bar{A}_д(x, y)$ имеет ограниченную протяженность ($-X_{max}$, X_{max}, $-Y_{max}$, Y_{max}), то функция $\hat{\Gamma}_{\Phi,д}(\xi, \eta)$ может быть определена путем интерполяции ее отсчетов $\hat{\Gamma}_\Phi(r, s)$

$$\hat{\Gamma}_\Phi(r, s) = \sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1} \bar{A}_0(k, l)\exp\left\{i2\pi\left[\frac{(k+u)(r+p)\Delta\xi}{2\xi_{max}} + \frac{(l+v)(s+q)\Delta\eta}{2\eta_{max}}\right]\right\}, \tag{10.20}$$

взятых с шагом

$$\Delta\xi = \lambda D/2X_{max};\ \Delta\eta = \lambda D/2Y_{max} \tag{10.21}$$

по прямоугольному растру, сдвинутому относительно системы координат (ξ, η) на $p\Delta\xi$ и $q\Delta\eta$ соответственно.

Таким образом, задача синтеза фурье-голограммы сводится к цифровому расчету матрицы отсчетов $\hat{\Gamma}_\Phi(r, s)$ и двум аналоговым процедурам: интерполяции для получения непрерывной голограммы $\hat{\Gamma}_{\Phi,д}(\xi, \eta)$ на этапе записи голограммы и интерполяции непрерывного объекта $\bar{A}_0(x, y)$ на этапе записи [функция $H(\xi, \eta)$] и восстановления голограммы.

Подставив (10.21) и (10.18) в (10.20), получим окончательно следующую формулу для

вычисления элементов матрицы $\{\hat{\Gamma}_\Phi(r, s)\}$ по матрице чисел $\{\bar{A}_0(k, l)\}$:

$$\hat{\Gamma}_\Phi(r, s) = \sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1} \bar{A}_0(k, l) \exp\left\{i 2\pi\left[\frac{(k+u)(r+p)}{N_1} + \right.\right.$$

$$\left.\left. + \frac{(l+v)(s+q)}{N_2}\right]\right\}. \qquad (10.22)$$

Эта формула является с точностью до нормирующего множителя формулой двумерного сдвинутого дискретного преобразования Фурье СДПФ $(u, v; p, q)$ (см. § 3.3).

Дискретное представление голограмм Френеля. Перепишем (10.12) в виде:

$$\Gamma_{\Phi P}(\xi, \eta) = \exp[i\pi(\xi^2+\eta^2)\lambda D_0] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \bar{A}_0(x, y) \exp\left[i\frac{\pi}{\lambda D} \times\right.$$

$$\left.\times (x^2+y^2)\right] \exp\left[-i\frac{2\pi}{\lambda D}(x\xi + y\eta)\right]. \qquad (10.23)$$

Как и в случае голограммы Фурье, функция $\Gamma_{\Phi P}(\xi, \eta)\exp\left[-i\frac{\pi}{\lambda D}\times\right.$ $\left.\times(\xi^2+\eta^2)\right]$ может быть восстановлена интерполяцией своих отсчетов:

$$\tilde{\Gamma}_{\Phi P}(\xi, \eta) = \Gamma_{\Phi P}(\xi, \eta) \exp\left[-i\frac{\pi}{\lambda D}(\xi^2+\eta^2)\right] =$$

$$= \sum_r \sum_s \Gamma_{\Phi P}(r\Delta\xi, s\Delta\eta) \exp\left[-i\frac{\pi}{\lambda D}((r+p)^2\Delta\xi^2 + \right.$$

$$\left. + (s+q)^2\Delta\eta^2)\right] \gamma[\xi - (r+p)\Delta\xi; \eta - (s+q)\Delta\eta] \qquad (10.24)$$

интерполирующей функцией $\gamma(\xi, \eta)$, где $\Delta\xi$ и $\Delta\eta$ определяются (10.21); p и q, как и для голограммы Фурье, – параметры, определяющие сдвиг растра отсчетов голограммы относительно системы координат (ξ, η)

Отсюда вытекает возможность восстановления $\Gamma_{\Phi P}(\xi, \eta)$:

$$\Gamma_{\Phi P}(\xi, \eta) = \tilde{\Gamma}_{\Phi P}(\xi, \eta) \exp\left[i\frac{\pi}{\lambda D}(\xi^2+\eta^2)\right] \qquad (10.25)$$

коррекцией голограммы $\tilde{\Gamma}_{\Phi P}(\xi, \eta)$ (10.24) путем освещения ее сферическим волновым фронтом

$$\Gamma_{C\Phi}(\xi, \eta) = \exp\left[i\frac{\pi}{\lambda D}(\xi^2+\eta^2)\right] \qquad (10.26)$$

Посмотрим теперь, как найти $\tilde{\Gamma}_{\Phi P}(\xi, \eta)$. Из (10.23) получаем

$$\Gamma_{\Phi P}(r\Delta\xi, s\Delta\eta) \exp\left\{-i\frac{\pi}{\lambda D}[(r+p)^2\Delta\xi^2 + (s+q)^2\Delta\eta^2]\right\} =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \bar{A}_0(x, y) \exp\left\{i\frac{\pi}{\lambda D}(x^2+y^2)\right\} \times$$

$$\times \exp\left\{-i\frac{2\pi}{\lambda D}[x(r+p)\Delta\xi + y(s+q)\Delta\eta]\right\} dxdy. \qquad (10.27)$$

Как и в случае голограмм Фурье, естественно считать, что $|A_0(x, y)|$ может быть восстановлена по ее отсчетам $|\bar{A}_0(k, l)|$ путем их интерполяции некоторой функцией $h(x, y)$ аналогично (10.16):

$$|\bar{A}_0(x, y)| = \sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1} |\bar{A}_0(k, l)| h[x - (k+n)\Delta x;$$

$$y - (l+v)\Delta y]. \qquad (10.28)$$

Тогда

$$\Gamma_{\Phi P}(r\Delta\xi,\ s\Delta\eta)\exp\left\{-i\frac{\pi}{\lambda D}\left[(r+p)^2\Delta\xi^2+(s+q)^2\Delta\eta^2\right]\right\}=$$

$$=\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}|\overline{A}_0(k,\ l)|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h[x-(k+u)\Delta x;\ y-(l+v)\Delta y]\times$$

$$\times\exp\left\{i\left[\frac{\pi}{\lambda D}(x^2+y^2)+\overline{\omega}_0(x,\ y)\right]\right\}\times$$

$$\times\exp\left\{-i\frac{2\pi}{\lambda D}[x(r+p)\Delta\xi+y(s+q)\Delta\eta]\right\}dxdy. \qquad (10.29)$$

Сделав замену переменных, перепишем (10:29) в виде

$$\Gamma_{\Phi P}(r\Delta\xi,\ s\Delta\eta)=\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}|\overline{A}_0(k,\ l)|\exp[i\omega_0(k\Delta x,\ l\Delta y)]\times$$

$$\times\exp\left\{i\frac{\pi}{\lambda D}[[(k+u)\Delta x-(r+p)\Delta\xi]^2+[(l+v)\Delta y-\right.$$

$$\left.-(s+q)\Delta\eta]^2]\right\}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,\ y)\exp\left\{-i\frac{2\pi}{\lambda D}[x(r+p)\Delta\xi+\right.$$

$$\left.+y(s+q)\Delta\eta]\right\}\exp\left\{i\frac{\pi}{\lambda D}[(x+(k+u)\Delta x)^2-(k+u)^2\Delta x^2+\right.$$

$$+(y+(l+v)\Delta y)^2-(l+v)^2\Delta y^2]\}\exp\{i[\overline{\omega}_0(x+(k+u)\Delta x;$$

$$y+(l+v)\Delta y]-\overline{\omega}((k+u)\Delta x;\ (l+v)\Delta y)]\}dxdy. \qquad (10.30)$$

Интерполирующая функция $h(x, y)$ обычно должна быть заметно отлична от нуля только внутри прямоугольника $(\pm\Delta x,\ \pm\Delta y)$. Поэтому при надлежащем наборе интервалов дискретизации Δx и Δy фазовыми ошибками в последних двух экспоненциальных множителях в (10.30) можно пренебречь, так что

$$\Gamma_{\Phi P}(r\Delta\xi,\ s\Delta\eta)=\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}\overline{A}_0(k,\ l)\exp\left\{i\frac{\pi}{\lambda D}[[(k+u)\Delta x-\right.$$

$$\left.-(r+p)\Delta\xi]^2+[(l+v)\Delta y-(s+q)\Delta\eta]^2]\right\}H(r,\ s), \qquad (10.31)$$

где маскирующая функция

$$H(r,\ s)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,\ y)\exp\left\{-i\frac{2\pi}{\lambda D}[x(r+p)\Delta\xi+\right.$$

$$\left.+y(s+q)\Delta\eta]\right\}dxdy, \qquad (10.32)$$

равна отсчетам преобразования Фурье интерполирующей функции h(x, y].

Сумма в (10.31) является с точностью до множителя дискретным преобразованием Френеля (см. § 3.4):

$$\Gamma_{\Phi P}(r,\ s)=\frac{1}{\sqrt{N_1N_2}}\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}\overline{A}_0(k,\ l)\times$$

$$\times\exp\left\{i\pi\left[\frac{(k\varkappa-r/\varkappa+w_x)^2}{N_1}+\frac{(l\nu-s/\nu+w_y)^2}{N_2}\right]\right\}, \qquad (10.33)$$

где

$$\varkappa^2=N_1\Delta x^2/\lambda D;\ \nu^2=N_2\Delta y^2/\lambda D;\ \Delta\xi=\lambda D/N_1\Delta x;\ \Delta\eta=\lambda D/N_2\Delta y;$$

$$w_x=u\sqrt{\Delta x/\Delta\xi}-p\sqrt{\Delta\xi/\Delta x};\ w_y=v\sqrt{\Delta y/\Delta\eta}-q\sqrt{\Delta\eta/\Delta y}. \qquad (10.34)$$

Таким образом, задача синтеза голограмм Френеля сводится к расчету с помощью дискретного преобразования Френеля матрицы $\{\Gamma_{\Phi P}(r,\ s)\}$ по матрице отсчетов $\{\overline{A}_0(k,\ l)\}$, описывающих комплексную амплитуду поля на объекте, и аналоговой интерполяции этих отсчетов в соответствии с (10.24) – (10.26). Интерполяция при восстановлении объекта с голограммы производится маскированием голограммы функцией $H(r, s)$ (10.32).

10.3. МЕТОДЫ И СРЕДСТВА ЗАПИСИ СИНТЕЗИРОВАННЫХ

Получающиеся в результате расчета по формулам (10.22) и (10.33) отсчеты голограмм Фурье и Френеля являются *математическими голограммами*, т.е. массивами чисел, вообще говоря, комплексных, которые определяют значения отсчетов амплитуды и фазы синтезируемого волнового фронта. Для того чтобы математическую голограмму превратить в физическую, способную сформировать требуемый волновой фронт, эти числа нужно преобразовать в параметры оптических сред, модулирующие соответственно амплитуду и фазу световой волны, используемой для реконструкции голограммы.

Существующие оптические среды можно разделить на три класса: *амплитудные*, *фазовые* и *комбинированные* или *амплитудно-фазовые*.

В амплитудных средах управляемым параметром является их коэффициент пропускания или отражения по интенсивности. Это наиболее распространенный класс сред. Характерный пример – стандартные галогенидо-серебряные эмульсии, используемые в фотографии и в оптической голографии.

В фазовых средах пропускание по интенсивности света не поддается управлению, но можно управлять оптической толщиной среды, например путем изменения ее коэффициента преломления или физической толщины или того и другого одновременно. Таковы фототермопластические среды, фоторезисты, отбеленные фотоматериалы, среды на основе бихромированной желатины, фотополимеры и т.п.

В комбинированных средах допускается независимое управление коэффициентом пропускания света по интенсивности и оптической толщиной. В настоящее время это двухслойные или многослойные фотоматериалы, каждый слой которых чувствителен к излучению с различной длиной волны, что позволяет независимо путем экспонирования каждого слоя на своей длине волны управлять прозрачностью одних слоев и оптической толщиной других.

Для управления оптическими параметрами этих сред в соответствии с результатами расчета волнового поля необходимы специальные *устройства записи голограмм*, управляемые цифровыми сигналами. В настоящее время таких устройств нет, и для этой цели используются различные дисплейные устройства, созданные для вывода из вычислительных машин символьных, графических и полутоновых изображений.

Особенность символьных и графических дисплеев заключается в том, что они способны осуществлять только бинарную, или двухуровневую модуляцию оптических параметров среды. Поэтому используемые с этими устройствами среды будем называть бинарными, считая, что их управляемые оптические параметры могут принимать только два значения. По информационной емкости бинарные среды уступают амплитудным и фазовым, так как возможности записи на них информации определяются только их пространственными степенями свободы (разрешающей способностью), тогда как в случае амплитудных и фазовых сред можно использовать также степени свободы по пропусканию (отражению, преломлению). Основное достоинство бинарных сред – простота экспонирования, фотохимической обработки, копирования.

Устройства вывода полутоновых изображений применяются для записи на амплитудных и фазовых средах. Цветные дисплейные устройства используются для записи голограмм на комбинированных средах. Управление оптическими параметрами среды в таких устройствах чаще всего осуществляется модулированным по интенсивности пучком света или электронов, который воздействует на (экспонирует) отдельные элементарные площадки (ячейки разрешения) чувствительной среды, записывая на них определенным образом закодированные значения отсчетов математической голограммы.

Важнейшими характеристиками устройств записи являются шаг дискретизации (расстояние $\Delta\xi$ и $\Delta\eta$ между соседними отдельно экспонируемыми, ячейками разрешения) и общее количество экспонируемых ячеек. Шаг дискретизации определяет угловые размеры восстанавливаемого изображения. Для того чтобы угловые размеры изображения были порядка десяти градусов или больше при длине волны источника света, используемого для восстановления голограммы, порядка 0,5 мкм, $\Delta\xi$ и $\Delta\eta$ должны быть меньше 3 мкм. В настоящее время наилучшие устройства имеют шаг дискретизации 1–10 мкм [16, 49], а общее

количество экспонируемых ячеек от 10^3x10^3 до 10^4x10^4. На ранних этапах развития работ по цифровой голографии для записи голограмм использовались стандартные графопостроители и символьные печатающие устройства для ЭВМ. Поэтому для получения приемлемых значений шага дискретизации требовалось фотографическое уменьшение получавшихся с помощью этих устройств «бумажных» голограмм.

Известные в настоящее время методы записи синтезированных голограмм на амплитудных, фазовых, бинарных и комбинированных средах, ориентированные на описанные типы устройств, можно классифицировать по различным признакам. Ниже приведена классификация по способу представления комплексных чисел, описывающих отсчеты математической голограммы. Другие варианты классификации читатель может найти в обзорах [16] и [59].

Возможны два способа представления комплексных чисел: экспоненциальное и аддитивное (рис. 10.2). С физической точки зрения наиболее естественным является *экспоненциальное представление* вида $A\exp(i\varphi)$, где A – модуль комплексного числа, φ– его фаза. Этому способу лучше всего соответствуют комбинированные среды, например трехслойная цветная фотопленка

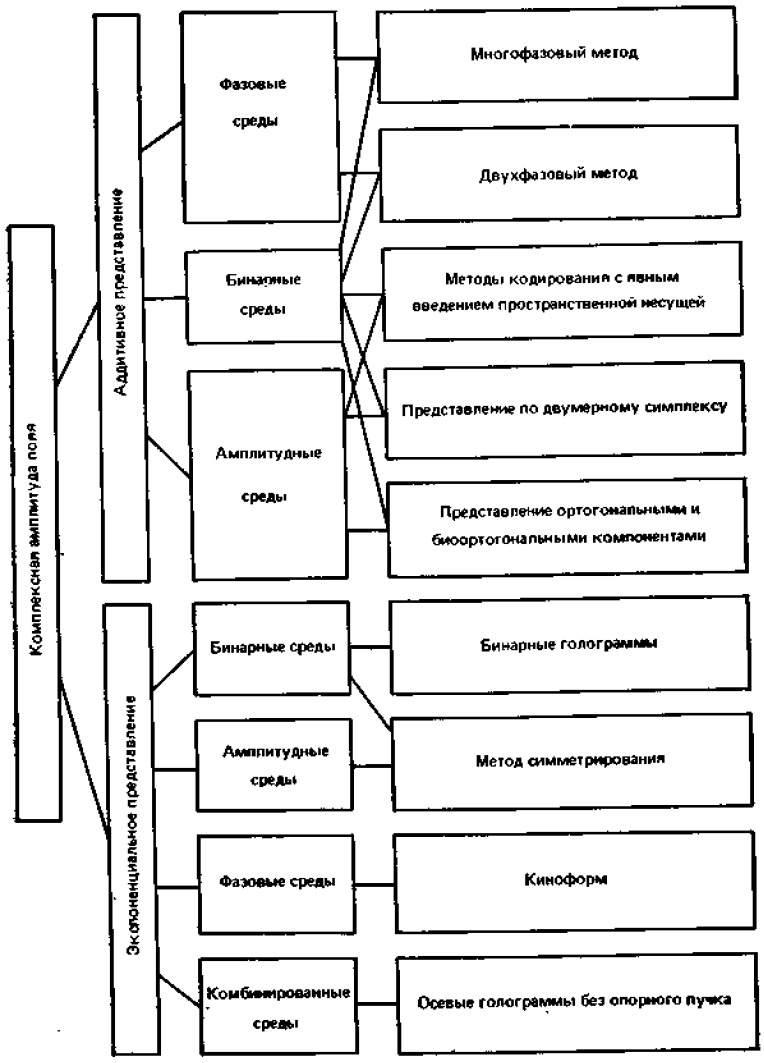


Рис. 10.2. Классификация методов записи синтезированных голограмм

(см. также [16]). Леземом и другими предложено при записи синтезированных голограмм игнорировать амплитудную информацию и записывать только фазу отсчетов голограммы. Это позволяет применять для записи чисто фазовые среды. Хотя такие голограммы, получившие название «киноформ», восстанавливают волновое поле с искажениями (некоторые данные о них можно найти в [49]), они очень выгодны энергетически, так как вся энергия пучка света, осуществляющего восстановление голограммы, не поглощаясь в голограмме и почти не рассеиваясь в посторонние порядки, переходит в энергию восстановленного волнового поля.

Кроме того, специальным выбором диффузной компоненты фазы поля на объекте можно до определенной степени уменьшить эти искажения [16].

В известном смысле дуальным методу киноформа является метод симметрирования [49]. Этот метод заключается в том, что посредством симметрирования объекта перед синтезом голограммы добиваются, чтобы его фурье-голограмма содержала только вещественные отсчеты и могла быть таким образом записана на чисто амплитудной среде. Поскольку вещественные числа могут принимать как положительные, так и отрицательные значения, которые невозможно передать с помощью амплитудной среды, значения голограммы должны записываться с постоянным положительным смещением, делающим все записываемые величины положительными.

Идея этого метода основана на хорошо известном свойстве интегрального преобразования Фурье, которое формально в обозначениях § 10.1 может быть записано следующим образом.

Если

$$\overline{A}_0(x,\ y) = \overline{A}_0^*(x,\ y) = \overline{A}_0(-x,\ y), \quad (10.35)$$

то $\Gamma_\Phi(\xi,\ \eta) = \Gamma_\Phi^*(\xi,\ \eta) = \Gamma_\Phi(-\xi,\ \eta),$

где * – знак комплексной сопряженности. Для СДПФ как дискретного представления интегрального преобразования Фурье аналогичное свойство требует симметрирования объекта по правилу, которое зависит от параметров сдвига *u, v, p, q* (§ 3.3). Так, при $2u$, $2v$ – целых и $p=q=0$ это правило таково:

$$A_c(k,\ l) = \begin{cases} \overline{A}_0(k,\ l), & 0 \leqslant k \leqslant N_1 - 1;\ 0 \leqslant l \leqslant N_2 - 1, \\ \overline{A}_0(2N_1 - kl), & N_1 \leqslant k \leqslant 2N_1 - 1;\ 0 \leqslant l \leqslant N_2 - 1. \end{cases} \quad (10.36)$$

Оно означает симметрирование удвоением объекта. При этом количество отсчетов на объекте и соответственно на его голограмме Фурье вдвое больше количества отсчетов исходного объекта. Такая двойная избыточность и позволяет обойтись без записи фазовой составляющей голограммы. Возможно также симметрирование учетверением, которое состоит в симметричном дополнении объекта по правилу (10.36), но выполняемому по обоим индексам *k* и *l*. При этом избыточность голограммы возрастает до 4. Голограммы симметрированных объектов также оказываются симметричными и восстанавливают удвоенный или учетверенный, в зависимости от способа симметрирования, объект (рис. 10.3).

Симметрирование является примером оптимального кодирования информации об объекте для согласования его голограммы со свойствами регистрирующей среды.

При *аддитивном представлении* комплексное число, рассматриваемое как вектор на комплексной плоскости, выражается в виде суммы нескольких компонент. При записи на амплитудных средах эти компоненты должны иметь стандартное направление (фазовый угол) и управляемую длину (амплитуду). Простейший случай – представление вектора Г его ортогональными компонентами, например вещественной Γ_{re} и мнимой T_{im} частями (рис. 10.5,а):

$$\Gamma = \Gamma_{re}\vec{e}_{re} + \overline{\Gamma}_{im}\vec{e}_{im}, \quad (10.37)$$

где $\vec{e}_{re}$ и $\vec{e}_{im}$ –ортогональные векторы единичной длины.

Кодировать фазовый угол между ортогональными компонентами можно методом фазового набега, записывая величины Γ_{re} и T_{im} в соседние по строкам растра ячейки разрешения голограм-

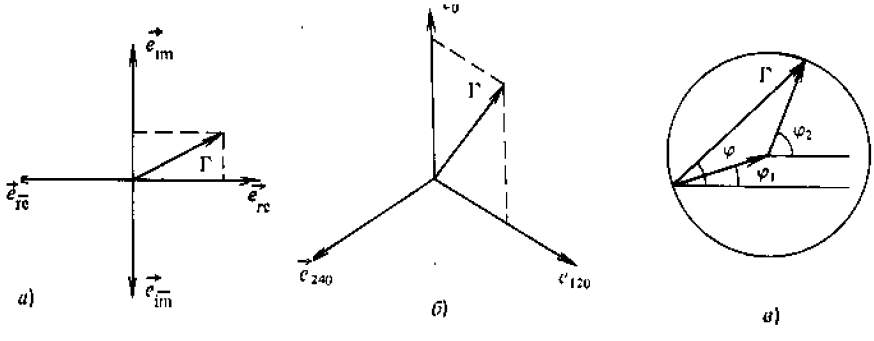


Рис. 10.5. Аддитивное представление комплексных чисел:
а – по биортогональному базису; б – по двумерному симплексу; в –суммой двух векторов одинаковой длины

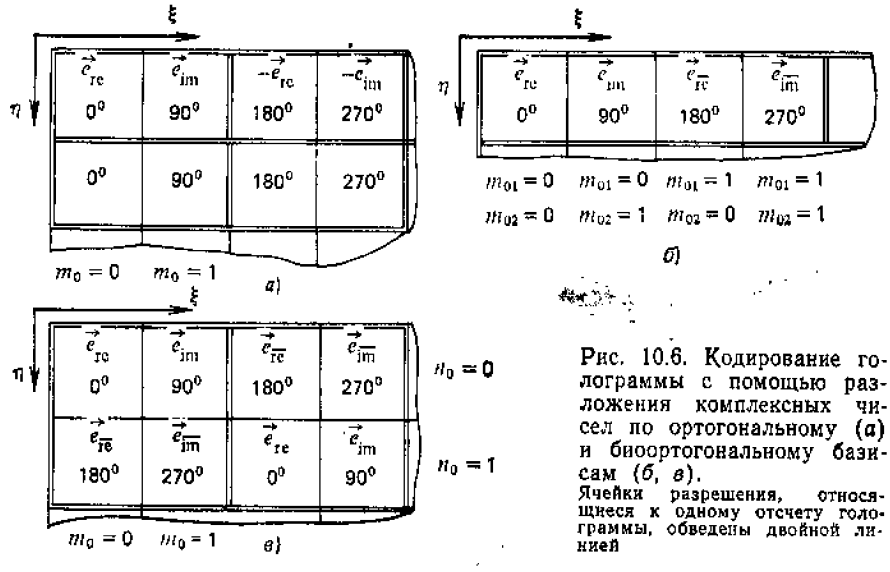


Рис. 10.6. Кодирование голограммы с помощью разложения комплексных чисел по ортогональному (*а*) и биортогональному базисам (*б*, *в*).
Ячейки разрешения, относящиеся к одному отсчету голограммы, обведены двойной линией

мы (см. рис. 10.6, с) [49]. При этом восстановленное изображение будет наблюдаться под углом Θ_ξ к оси ξ, определяемым соотношением:

$$\Delta_\varepsilon \cos \Theta_\xi = \lambda/4, \quad (10.38)$$

где λ–длина волны пучка света, осуществляющего восстановление голограммы. Для записи отрицательных значений Γ_{re} и Γ_{im} можно вводить постоянное положительное смещение.
Поскольку здесь для записи одного отсчета голограммы затрачивается два элемента разрешения среды, такая голограмма обладает двойной избыточностью, как и в методе симметрирования с удвоением. При таком кодировании и записи голограммы необходимо также учитывать, что каждой паре ячеек разрешения голограммы соответствует оптическая разность хода $\lambda/2$ и $3\lambda/4$ (см. рис. 10.6, а), т.е. значения Γ_{re} и Γ_{im} каждого нечетного отсчета математической голограммы следует записывать с противоположным знаком.

Формально этот способ кодирования можно описывать следующим образом. Пусть (r, s) – индексы, характеризующие номер отсчета математической голограммы $\Gamma_{r,s}$, r=0, 1, ..., N_1–1; s=0, 1,..., N_2–1; m, n – индексы номера ячейки разрешения среды; $\tilde{\Gamma}(m, n)$ – голограмма, закодированная для записи. Тогда закодированная голограмма записывается как

$$\tilde{\Gamma}(m, n) = \frac{1}{2}(-1)^r (1)^{m_0}[(-1)^{m_0}\Gamma(r, s) + \Gamma^*(r, s)] + c, \quad (10.39)$$

где $m=2r+m_0$, m_0=0; 1; n=s; c – положительная константа смещения. Действительно,

при $m_0 = 0$, $\tilde{\Gamma}(m, n) = (-1)^r \mathrm{Re}\{\Gamma(r, s)\}$;

при $m_0 = 1$, $\tilde{\Gamma}(m, n) = (-1)^r \mathrm{Im}\{\Gamma(r, s)\}$.

Чтобы избежать необходимости введения постоянного смещения в значения Γ_{re} и Γ_{im} и связанной с этим потери энергетической эффективности, можно [59] отводить для записи одного отсчета голограммы четыре соседние по строкам растра ячейки разрешения среды (рис. 10.6,*б*), При угле восстановления, определяемом (10.38), им соответствуют набеги фазы 0; $\pi/2$; π; $3\pi/2$. Поэтому запись в эти ячейки должна производиться в следующем порядке: $(\Gamma_{re} + |\Gamma_{re}|)/2$; $(\Gamma_{im} + |\Gamma_{im}|)/2$; $(|\Gamma_{re}| - \Gamma_{re})/2$; $(|\Gamma_{im}| - \Gamma_{im})/2$.

Нетрудно видеть, что при этом все записываемые величины неотрицательны, и метод означает представление вектора на комплексной плоскости по биортогональному базису $(\vec{e}_{re}, \vec{e}_{im}, \vec{e}_{\overline{re}}, \vec{e}_{\overline{im}})$ (рис. 10.5, с):

$$\Gamma = \frac{1}{2}(\Gamma_{re} + |\Gamma_{re}|)\vec{e}_{re} + \frac{1}{2}(\Gamma_{im} + |\Gamma_{im}|)\vec{e}_{im} + \frac{1}{2}(|\Gamma_{re}| - \Gamma_{re})\vec{e}_{\overline{re}} + \frac{1}{2}(|\Gamma_{im}| - \Gamma_{im})\vec{e}_{\overline{im}}. \quad (10.40)$$

Формально, в обозначениях формулы (10.39) этот метод можно описать следующим образом:

$$\tilde{\Gamma}(m, n)=\frac{1}{4}\{|(i)^{m_{01}}[(-1)^{m_{02}}\Gamma(r, s)+\Gamma^*(r, s)]|+$$
$$+(-1)^{m_{01}}(i)^{m_{02}}[(-1)^{m_{02}}\Gamma(r, s)+\Gamma^*(r, s)]\}, \quad (10.41)$$

где $m=4r+2m_{01}+m_{02}$; $m_{01}=0;\ 1$; $m_{02}=0;\ 1$; $n=s$, а вертикальная черта означает операцию взятия модуля числа.

Для сохранения пропорций восстановленного изображения при кодировании голограммы в этом случае требуется, чтобы размеры ячейки разрешения среды в одном направлении были вчетверо меньше, чем в перпендикулярном направлении. Если при кодировании отводить для записи каждого отсчета математической голограммы по две ячейки разрешения в двух соседних строках растра (рис. 10.6,*в*), т.е. выполнять запись, например, в соответствии со следующим соотношением:

$$\tilde{\Gamma}(m, n)=\frac{1}{4}\{|(i)^{m_0}[(-1)^{m_0}\Gamma(r, s)+\Gamma^*(r, s)]|+$$
$$+(-1)^{n_0}(i)^{m_0}[(-1)^{m_0}\Gamma(r, s)+\Gamma^*(r, s)]\}, \quad (10.42)$$

где $m=2r+m_0$; $n=2s+n_0$; $m_0, n_0=0, 1$, то изображение будет восстанавливаться в направлении, составляющем углы Θ_ξ и Θ_η с направлениями вдоль осей ξ и η в плоскости голограммы, такие, что

$$\Delta\xi\cos\Theta_\xi=\lambda/2;\ \Delta\eta\cos\Theta_\eta=\lambda/4. \quad (10.43)$$

Представление вектора по биортогональному базису (10.40) избыточно. Это проявляется в том, что из четырех компонент две всегда равны нулю. Избыточность можно уменьшить, если производить разложение комплексного числа по двумерному симплексу $(\vec{e}_0, \vec{e}_{120}, \vec{e}_{240})$ (рис.10.5,5):

$$\Gamma=\Gamma_0\vec{e}_0+\Gamma_1\vec{e}_{120}+\Gamma_2\vec{e}_{240}. \quad (10.44)$$

Этот базис, так же как и биортогональный, не является линейно независимым, поскольку

$$\vec{e}_0+\vec{e}_{120}+\vec{e}_{240}=0, \quad (10.45)$$

и содержащуюся в нем избыточность можно использовать для того, чтобы гарантировать неотрицательность компонент Γ_0, Γ_1 Γ_2. Известны два варианта кодирования голограмм по двумерному симплексу. Идея Бэркхардта состоит в том, что произвольный вектор на плоскости может быть представлен в виде суммы двух компонент, направленных вдоль тех двух из трех векторов $\vec{e}_0$, $\vec{e}_{120}$ и $\vec{e}_{240}$, которые ограничивают треть плоскости, содержащую данный вектор. Это значит, что из трех чисел $\mathbf{\Gamma}_0$, $\mathbf{\Gamma}_1$ и $\mathbf{\Gamma}_2$, определяющих данное комплексное число $\mathbf{\Gamma}$ по (10.44), два всегда положительны и являются проекциями вектора $\mathbf{\Gamma}$ на соответствующие базисные векторы, а третье равно нулю. Из этого условия вытекают следующие соотношения для $\mathbf{\Gamma}_0$, $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$

$$\Gamma_0=\frac{1}{2\sqrt{3}}[(1+\operatorname{sign}A)(|B|-B)+(1-\operatorname{sign}A)(|C|+C)];$$
$$\Gamma_1=\frac{1}{2\sqrt{3}}[(1+\operatorname{sign}C)(|A|-A)+(1-\operatorname{sign}C)(|B|+B)]; \quad (10.46)$$
$$\Gamma_2=\frac{1}{2\sqrt{3}}[(1+\operatorname{sign}B)(|C|-C)+(1-\operatorname{sign}B)(|A|+A)],$$

где

$$A = \frac{1}{2}(\Gamma + \Gamma^*);$$
$$B = \frac{1}{2}\left(\Gamma \exp\left(i\frac{2\pi}{3}\right) + \Gamma^* \exp\left(-i\frac{2\pi}{3}\right)\right);$$
$$C = \frac{1}{2}\left(\Gamma \exp\left(-i\frac{2\pi}{3}\right) + \Gamma^* \exp\left(i\frac{2\pi}{3}\right)\right), \quad (10.47)$$

а $\operatorname{sign}(A, B, C)$ –знак A, B и C соответственно.

Шавел и Югонен заметили, что если прибавить к компонентам Γ_0, Γ_1 Γ_2 произвольную константу V, то вследствие (10.45) сумма (10.44) не изменится. Поэтому они предложили отыскивать Γ_0, Γ_1 Γ_2 в виде

$$\Gamma_k = \Gamma_k^0 + V, \quad k = 0, 1, 2, \quad (10.48)$$

связав Γ_k^0 следующим условием:

$$\sum_{k=0}^{2} \Gamma_k^0 \vec{e}_{k\cdot 120} = 0$$

и выбирая V так, чтобы Γ_k были неотрицательными. В этом случае Γ_k^0 определяются следующими соотношениями:

$$\Gamma_0^0 = \frac{1}{3}(\Gamma + \Gamma^*) = \frac{2}{3}A;$$
$$\Gamma_1^0 = \frac{1}{3}\left(\Gamma \exp\left(-i\frac{2\pi}{3}\right) + \Gamma^* \exp\left(i\frac{2\pi}{3}\right)\right) = \frac{2}{3}C;$$
$$\Gamma_2^0 = \frac{1}{3}\left(\Gamma \exp\left(i\frac{2\pi}{3}\right) + \Gamma^* \exp\left(-i\frac{2\pi}{3}\right)\right) = \frac{2}{3}B. \quad (10.49)$$

Очевидно, смещение V может иметь различные значения для различных отсчетов голограммы. Ясно, что выбор этих значений сказывается на качестве восстановленного изображения. Шавел и Югонен предложили несколько вариантов выбора V, обеспечивающих высокое качество восстановленных изображений.

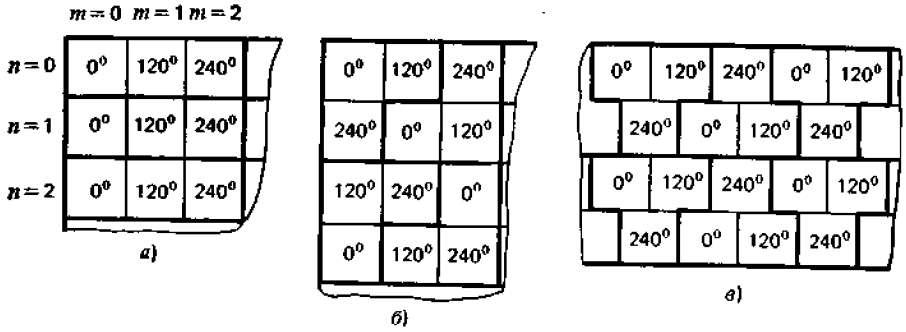


Рис. 10.7. Кодирование голограммы с помощью разложения комплексных чи сел по симплексу:

а — метод Бэркхардта; *б* — вариант метода с укладкой ячеек, передающих векторы $\vec{e}_0$ $\vec{e}_{120}$ и $\vec{e}_{240}$ по ортогональному растру; *в* — вариант с укладкой по гексагональному растру Тройки ячеек разрешения среды, используемых для кодирования одного отсчета голо граммы, обведены жирной линией

Для кодирования при записи голограммы фазового угла, соответствующего векторам $\vec{e}_0$, $\vec{e}_{120}$ и $\vec{e}_{240}$, можно использовать упоминавшийся уже метод фазового набега, записывая компоненты Γ_0, Γ_1 и Γ_2 в соседние на растре ячейки разрешения голограммы. Если, как предложил Бэркхардт, размещать три компоненты разложения вектора по симплексу в трех соседних по строкам растра ячейках разрешения, нумеруемых по m_0 от 0 до 2 (см. рис. 10.7,а), изображение восстанавливается в направлении, составляющем угол

$$\Theta_\xi = \arccos \lambda/3\Delta\xi \quad (10.50)$$

к оси ξ, совпадающей с направлением строк растра голограммы.

Пользуясь (10.46), нетрудно показать, что методу кодирования Бэркхардта соответствует следующая формула, связывающая величины, записываемые в ячейках *m*, *n* среды с отсчетами математической голограммы **Γ** (*r, s*):

$$\tilde{\Gamma}(m,\ n)=\frac{1}{2\sqrt{3}}\left[\left(1+\frac{\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i\frac{2\pi}{3}m_0\right)\right\}}{\left|\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i\frac{2\pi}{3}m_0\right)\right\}\right|}\right)\times\right.$$
$$\times\left(\left|\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i\frac{2\pi}{3}(m_0-1)\right)\right\}\right|-\right.$$
$$\left.-\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i\frac{2\pi}{3}(m_0-1)\right)\right\}\right)+$$
$$+\left(1-\frac{\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i\frac{2\pi}{3}m_0\right)\right\}}{\left|\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i\frac{2\pi}{3}m_0\right)\right\}\right|}\right)\times$$
$$\times\left(\left|\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(i\frac{2\pi}{3}(m_0+1)\right)\right\}\right|+\right.$$
$$\left.\left.+\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(i\frac{2\pi}{3}(m_0+1)\right)\right\}\right)\right],\qquad(10.51)$$

где $\operatorname{Re}\{z\}$ –вещественная часть числа *z*; $m=3r+m_0$; $m_0=0,\ 1,\ 2$; $n=s$.

Аналогично можно записать голограмму, закодированную методом Шавела к Югонена.

При записи компонент симплекса вдоль строк растра используется втрое больше ячеек разрешения голограммы, чем вдоль столбцов, поэтому масштаб восстановленного изображения вдоль горизонтали и вертикали также будет отличаться в три раза. Чтобы уравнять масштабы, можно каждую строку голограммы повторять три раза. Но это означает излишний расход ячеек разрешения. Уменьшить его можно, прибегнув к двумерной укладке ячеек, передающих векторы $\vec{e}_0,\ \vec{e}_{120}$ и $\vec{e}_{240}$. На рис. 10.7 б, в показаны два способа такой укладки: по ортогональному растру и по гексагональному растру. В этом случае отношение масштабов по координатам становится соответственно $2:1{,}5$ и $3\sqrt{3}:4$ вместо 3:1. Формулы, описывающие эти модификации метода кодирования по симплексу, опускаем из-за их громоздкости.

Очевидно, что описанные методы применимы для записи голограмм не только на амплитудных, но и на бинарных средах. В этом случае для передачи проекций комплексного числа на базовые векторы следует менять размеры прозрачной апертуры в каждой из ячеек разрешения, отвечающих за передачу соответствующего базисного вектора.

Однако специфически бинарным методом, реализующим идею аддитивного представления комплексного числа, является метод *импульсно-кодовой модуляции* (ИКМ). Он заключается в том, что для передачи одного отсчета математической голограммы отводится группа, например, KxL соседних по растру ячеек разрешения (К вдоль строк, L вдоль столбцов растра), причем каждая ячейка может быть только полностью прозрачной (отражающей) или непрозрачной. Полное число возможных состояний такой группы равно, очевидно, 2^{KL}. Таким образом, каждая группа может кодировать 2^{KL} различных векторов. Все эти векторы могут быть заранее рассчитаны. При записи голограммы требуется только для каждого отсчета математической голограммы найти ближайший к нему вектор из 2^{KL} возможных и записать соответствующую этому вектору комбинацию прозрачных и непрозрачных ячеек. Поиск может осуществляться путем перебора, который может потребовать до 2^{KL} шагов сравнения, или путем процедуры взвешивания, сокращающей это число до KL. Метод ИКМ существенно более эффективно, чем описанные выше методы бинарной записи для экспоненциального представления комплексных чисел, использует степени свободы среды. Действительно, при том же числе KL ячеек разрешения среды, отводимом для передачи одного отсчета математической голограммы, методом ИКМ можно закодировать не K.L различных векторов (например,

векторов, имеющих K значений амплитуды и L значений фазы), а 2^{KL}.

Перейдем теперь к методам записи на фазовой среде, опирающимся на аддитивное представление комплексных чисел. В этом случае составляющие векторы должны иметь стандартную длину и управляемое направление (фазовый угол). Простейший вариант– представление векторов в виде суммы двух компонент (рис. 10.5, в):

$$\Gamma = A_0 \exp(i\varphi_1) + A_0 \exp(i\varphi_2) \quad (10.52)$$

Из рис. 10.5,*в* легко видеть, что фазовые углы φ_1 и φ_2 составляющих векторов определяются следующей формулой:

$$\varphi_1 = \varphi - \arccos(|\Gamma|/2A_0);\ \varphi_2 = \varphi + \arccos(|\Gamma|/2A_0), \quad (10.53)$$

где $|\Gamma|$–мбдуль, а φ–фазовый угол кодируемого отсчета голограммы.

Этот метод назван методом *двухфазного кодирования*. Его можно использовать для записи голограмм как на фазовых, так и на бинарных средах.

При записи на фазовых средах для передачи двух составляющих векторов можно отводить две соседние ячейки разрешения среды, как на рис. 10.6,*а*. Формально это можно записать следующим образом:

$$\tilde{\Gamma}(m,\ n) = A_0 \exp\left\{i\left[\varphi(r, s) - (-1)^{m_0}\arccos\left(\frac{|\Gamma(r, s)|}{2A_0}\right)\right]\right\}, \quad (10.54)$$

где $m=2r+m_0$; $m_0 = 0, 1$; $n=s$.

Изображение в этом случае будет восстанавливаться в направлении, нормальном к плоскости голограммы, так как в этом направлении оптическая разность хода лучей, проходящих через соседние ячейки разрешения голограммы, равна нулю. Однако это справедливо только для центра изображения, через который проходит оптическая ось системы восстановления. На периферийных участках изображения между этими лучами накапливается некоторый набег фазы, вследствие чего на периферии изображение искажается. Подробнее эти искажения будут рассмотрены в следующем разделе.

Шмарев предложил записывать отдельно две голограммы для каждого из составляющих векторов (10.52) и суммировать в специальной оптической схеме восстанавливаемые с них изображе-

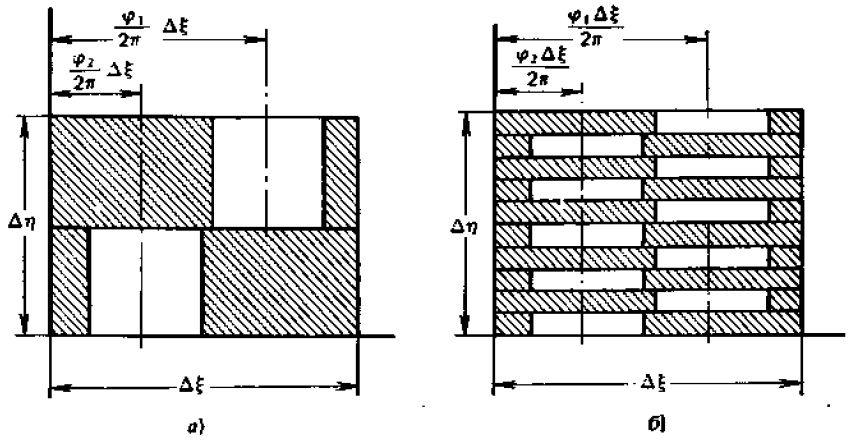


Рис. 10.8. Двухфазовый метод кодирования при записи голограмм на бинарной среде:
а – вариант с двумя разделевнымн ячейками; б – вариант с разбиением ячеек на подъ-ячейки и перемежением подъячеек

ния. Хсу и Савчук исследовали два варианта использования этого метода для записи на бинарных средах с кодированием фазы составляющих векторов методом фазового набега. В первом варианте на передачу каждого из двух векторов-компонент отводятся две элементарные ячейки бинарной среды, а фазы векторов кодируются смещением прозрачной (или полностью отражающей) апертуры вдоль перпендикуляра к линии, соединяющей центры ячеек (рис.10.8,а).

Такой записи присущи те же искажения, связанные с пространственным смещением элементарных ячеек друг относительно друга, что и для записи на фазовых средах. Для уменьшения этих искажений Хсу и Савчук предложили разбить каждую элементарную ячейку на n подъячеек и расположить их, как показано на рис. 10.8,6. Очевидно, что при этом

относительное смещение элементарных ячеек будет уже равно не $\Delta\eta$, а $\Delta\eta/n$, и искажения изображения на периферии могут быть существенно уменьшены. Перемежение подъячеек – эффективный способ борьбы с искажениями, но, как и любой способ бинарной записи, он требует повышенного расхода пространственных степеней свободы среды и устройств для записи голограмм.

Двухфазовый метод очевидным образом обобщается на случай многофазового кодирования разложения вектора по К компонентам равной длины [49]:

$$\Gamma = \sum_{k=0}^{K-1} A_0 \exp(i\varphi_k) \qquad (10.55)$$

Так как **Г**–комплексное число, равенство (10.55) представляет собой два уравнения для **К** неизвестных значений фаз φ_k . Эти уравнения имеют единственное решение (10.53) только при K=2. При K>2 φ_K могут выбираться достаточно произвольно. Например, при нечетных K удобно выбирать φ_k так, чтобы фазовые углы φ_k составляли арифметическую прогрессию:

$$\varphi_{k+1} - \varphi_k = \varphi_k - \varphi_{k-1} = \theta. \qquad (10.56)$$

В этом случае получим следующие уравнения для приращения θ:

$$(\sin K\theta/2)/\sin(\theta/2) = |\Gamma|/A_0 \qquad (10.57)$$

и фазового угла φ_k;

$$\varphi_k = \varphi + \left(k - \frac{K-1}{2}\right)\theta; \quad k = 0, 1, \ldots, K-1 \qquad (10.58)$$

При нечетных K уравнение (10.57) сводится к алгебраическому уравнению степени (K–1)/2 относительно $\sin^2\theta/2$. Так, при K=3 получаем

$$\theta = 2\arcsin\frac{\sqrt{3}}{2}\sqrt{1 - \frac{|\Gamma|}{3A_0}}. \qquad (10.59)$$

При четных K целесообразно разделить все составляющие векторы на две группы векторов с одинаковыми фазовыми углами φ_1 и φ_2, определяемыми по аналогии с (10.53) равенствами:

$$\varphi_1 = \varphi - \arccos(|\Gamma|/KA_0); \quad \varphi_2 = \varphi + \arccos(|\Gamma|/KA_0). \qquad (10.60)$$

Выбор K>2 позволяет увеличить динамический диапазон возможных значений амплитуды голограммы, так как максимальная воспроизводимая амплитуда равна KA_0.

Из случаев K>2 наибольший интерес представляют K=3 и K=4. Для них можно более эффективно, чем для K=2, использовать двумерные пространственные степени свободы среды и устройств для записи голограмм, размещая составляющие векторы по схемам, показанным на рис. 10.7,*б, в* и 10.6,*в*.

Описанные методы кодирования голограмм, основанные на аддитивном представлении комплексного числа, объединяет еще одно важное свойство. Все они в том или ином виде используют неявное введение пространственной несущей и нелинейное преобразование сигнала с пространственной несущей, аналогичные тому, как это делается в классическом методе записи оптических голограмм [60].

Действительно, легко проверить, что, например, формулы (10.39), (10.41), (10.42), (10.51), (10.54) можно переписать в следующей эквивалентной форме, содержащей в явном виде отсчеты голограммы, умноженные на отсчеты пространственной несущей по одной или обеим координатам:

$$\tilde{\Gamma}(m,\ n)=\frac{1}{4}\operatorname{Re}\left\{\Gamma(r,\ s)\exp\left(-i2\pi\frac{m}{2}\right)\right\}+c, \qquad (10.61)$$

$$m=2r+m_0;\ m_0=0,\ 1;\ n=s;$$

$$\tilde{\Gamma}(m,\ n)=\frac{1}{2}\operatorname{rctf}\left\{\operatorname{Re}\left[\Gamma(r,\ s)\exp\left(-i2\pi\frac{m}{2}\right)\right]\right\}; \qquad (10.62)$$

$$m=4r+2m_{01}+m_{02};\ m_{01},\ m_{02}=0,\ 1;\ n=s;$$

$$\tilde{\Gamma}(m,\ n)=\frac{1}{2}\operatorname{rctf}\left\{\operatorname{Re}\left[\Gamma(r,\ s)\exp\left(-i2\pi\frac{m+2n}{4}\right)\right]\right\}, \qquad (10.63)$$

$$m=2r+m_0;\ n=2s+n_0;\ m_0,\ n_0=0,\ 1;$$

$$\tilde{\Gamma}(m,\ n)=\frac{1}{\sqrt{3}}\sum_{p=0}^{1}\operatorname{hctf}\left\{\operatorname{Re}\left[\Gamma(r,\ s)\exp\left(-i2\pi\frac{m+p}{3}\right)\right]\right\}\times$$

$$\times\operatorname{rctf}\left\{\operatorname{Re}\left[\Gamma(r,\ s)\exp\left(-i2\pi\frac{m+p+1/2}{3}\right)\right]\right\}; \qquad (10.64)$$

$$m=3r+m_0;\ m_0=0,\ 1,\ 2;\ n=s;$$

$$\tilde{\Gamma}(m,\ n)=A_0\exp\left\{i\left[\varphi(r,\ s)-\cos\left(2\pi\frac{m}{2}\right)\arccos\left(|\Gamma(r,\ s)|/2A_0\right)\right]\right\}; \qquad (10.65)$$

$$m=2r+m_0;\ m_0=0,\ 1,\ n=s,$$

где $\operatorname{rctf}(z)$ – функция выпрямителя:

$$\operatorname{rctf}(z)=\begin{cases}z, & z\geqslant 0;\\ 0, & z<0;\end{cases} \qquad (10.66)$$

$\operatorname{hctf}(z)$ – функция жесткого ограничителя:

$$\operatorname{hctf}(z)=(1+\operatorname{sign} z)/2=\begin{cases}1, & z\geqslant 0;\\ 0, & z<0.\end{cases} \qquad (10.67)$$

Как видно из этих выражений, пространственная несущая имеет период по каждой координате, по крайней мере вдвое меньший периода следования отсчетов голограммы, т.е. на один отсчет голограммы приходится не менее двух отсчетов пространственной несущей для того, чтобы по промодулированному сигналу пространственной несущей можно было восстановить амплитуду и фазу отсчетов голограммы. Эта избыточность означает, что для осуществления модуляции голограммой пространственной несущей необходимо кроме основных отсчетов математической голограммы иметь от одного до нескольких промежуточных отсчетов. Их можно находить с помощью того или иного способа интерполяции отсчетов голограммы.

Простейший способ интерполяции – повторение отсчетов. Именно такая интерполяция и подразумевается в приведенных способах записи. Например, согласно (10.39) каждый отсчет голограммы повторяется дважды на два отсчета пространственной несущей, в формуле (10.41)– четыре раза на четыре отсчета и т.д. Конечно, такая интерполяция нулевого порядка, характерная для всех способов кодирования, использующих метод фазового набега, дает грубое приближение значений промежуточных отсчетов. Связанные с ней искажения восстановленного изображения, проявляющиеся в наложении на него мешающих изображений, иллюстрируются в § 10.4. Для частичной коррекции этих искажений был предложен ряд итерационных алгоритмов расчета голограмм, в основном для бинарных способов кодирования. Идея всех таких алгоритмов состоит в том, чтобы итеративным путем определять значение фазы отсчета голограммы, расположенного в элементарной ячейке, там, где в соответствии с методом фазового набега должна располагаться прозрачная апертура.

Для определений точных значений требуемых промежуточных отсчетов голограммы необходимо при синтезе голограммы дополнительно выполнить СДПФ ($u,\ v,\ p,\ q$) столько раз, сколько требуется дополнительных отсчетов на один основной отсчет, изменяя каждый раз соответствующим образом параметры сдвига p и q (см. § 4.2). Отметим, что метод симметрирования можно рассматривать как аналог метода (10.39) с идеальной интерполяцией

промежуточных отсчетов. При методе симметрирования такая интерполяция осуществляется автоматически и, как увидим в § 10.4, не происходит искажения восстановленного изображения мешающими изображениями.

Кроме рассмотренных методов с неявным введением пространственной несущей, известен ряд методов, основанных на явном введении в голограмму пространственной несущей. Большинство из них было предложено еще в первых работах по цифровой голографии из соображений имитации схем оптических голограмм и аналогии между голограммами и интерферограммами.

Из методов, ориентированных на использование амплитудных сред, упомянем методы Бэрча, предложившего записывать голограмму в виде

$$\tilde{\Gamma}(m,\ n)=0{,}5\left\{1+|\ \Gamma(m,\ n)\ |\cos\left[\varphi(m,\ n)+\frac{2\pi}{a}\,m\right]\right\} \qquad (10.68)$$

и Хуанга – Прасады, предложивших для увеличения контраста полезной компоненты голограмм (второго слагаемого в сумме (10.68)) использовать, постоянное смещение, равное $|\Gamma(m,\ n)|$

$$\tilde{\Gamma}(m,\ n)=0{,}5\ |\ \Gamma(m,\ n)\ |\left\{1+\cos\left[\varphi(m,\ n)+\frac{2\pi}{a}\,m\right]\right\}. \qquad (10.69)$$

Из методов, ориентированных на бинарные среды, можно назвать метод, описанный в обзоре [59]:

$$\tilde{\Gamma}(m,\ n)=\operatorname{hctf}\left\{\cos\left[\arcsin\left(\frac{|\ \Gamma(m,\ n)\ |}{A_0}\right)\right]+\right.$$
$$\left.+\cos\left[\varphi(m,\ n)+\frac{2\pi}{a}\,m\right]\right\}. \qquad (10.70)$$

В выражениях (10.69), (10.70) a – период пространственной несущей.

Из методов, ориентированных на фазовые среды, назовем метод Кирка, Джонса, предложивших записывать на фазовой среде функцию

$$\tilde{\Gamma}(m,\ n)=A_0\exp\{1[\varphi(m,\ n)-g(m,\ n)\cos 2\pi m/a]\}, \qquad (10.71)$$

где $g(m,\ n)$ определенным образом зависит от $|\Gamma(m,n)|$ и номера порядка дифракции, в котором требуется получить восстановленное изображение. Этот метод в известном смысле эквивалентен методам многофазового кодирования, а при a=2

$$g(m,\ n)=\arccos(\ |\ \Gamma(r,\ s)|\ /2A_0),$$
$$m=2r+m_0;\ m_0=0,\ 1;\ n=s, \qquad (10.72)$$

так что он совпадает с методом двухфазового кодирования (10.54).

10.4. ВОССТАНОВЛЕНИЕ СИНТЕЗИРОВАННЫХ ГОЛОГРАММ

Процесс восстановления и визуального наблюдения синтезированных голограмм можно описать схемами, показанными на рис. 10.9,*а, б*. В этих схемах синтезированная дискретная голограмма, рассчитанная с помощью дискретных представлений преобразований поля, работает как аналоговый физический элемент, и создаваемое ею поле подвергается непрерывным преобразованиям. При этом эффекты дискретизации и способ преобразования цифрового сигнала в физическую голограмму (способ записи голограммы) непосредственно сказываются на результатах восстановления голограмм.

Ниже приведен анализ процесса восстановления синтезированных фурье-голограмм в аналоговой схеме, осуществляющей преобразование Фурье, для трех методов записи синтезированных голограмм: симметрирования, ортогонального кодирования (10.39), двухфазовой записи на фазовой среде.

Метод симметрирования. В этом случае процесс превращения матрицы чисел математической голограммы $\tilde{\Gamma}_\Phi(r,\ s)$ в физиче–

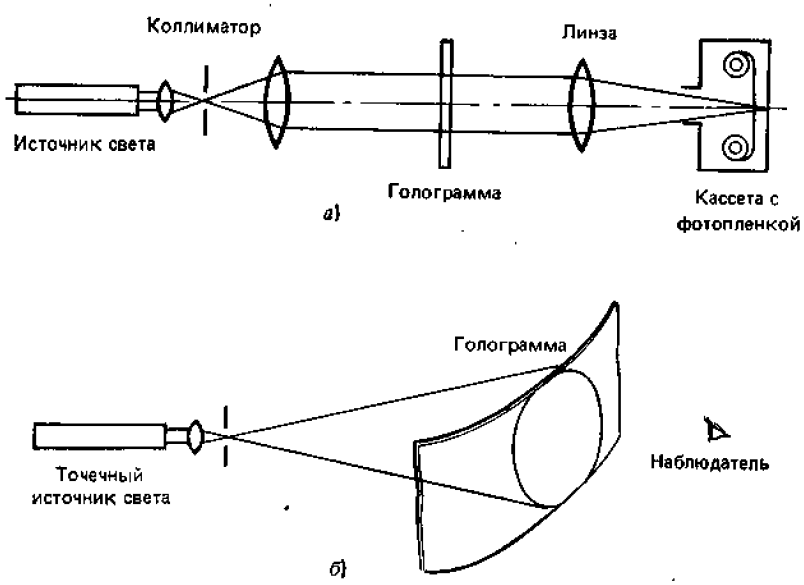


Рис. 10.9. Схемы восстановления синтезированных голограмм для документирования (а) и визуального наблюдения (б) восстановленных изображений

скую голограмму $\Gamma(\xi,\ \eta)$ в записывающем устройстве можно описать следующим образом:

$$\Gamma(\xi,\ \eta) = W(\xi,\ \eta)\left\{\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}[\hat{\Gamma}_{\Phi}(r,\ s)+c]\,H_{3}(\xi+\xi_{0}-r\Delta\xi;\right.$$
$$\left.\eta+\eta_{0}-s\Delta\eta)\right\}, \qquad (10.73)$$

где с – константа постоянного смещения, необходимая для передачи отрицательных значений $\Gamma_{\phi}(r,\ s)$; $\Delta\xi,\ \Delta\eta$ –размеры шага дискретизации по осям $\xi,\ \eta$ в устройстве записи голограмм; ξ_0, η_0 – константы, зависящие от геометрии расположения голограммы относительно оптической оси схемы; $H_3(\xi,\ \eta)$–функция, описывающая апертуру устройства записи, которое осуществляет аналоговую интерполяцию отсчетов дискретной голограммы; $W(\xi,\ \eta)$ – маскирующая функция, которая определяет физические размеры записанной голограммы (аподизирующая функция). Формулу (10.73) можно переписать в виде

$$\Gamma(\xi,\ \eta) = W(\xi,\ \eta)\,[H_{3}(\xi,\ \eta) * \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}[\hat{\Gamma}_{\Phi}(r,\ s)+c]\times$$
$$\times\,\delta(\xi+\xi_{0}-r\Delta\xi)\,\delta(\eta+\eta_{0}-s\Delta\eta)], \qquad (10.74)$$

где * –знак операции свертки.

В соответствии с теоремой о свертке результат обратного непрерывного преобразования Фурье голограммы $\Gamma(\xi,\ \eta)$, выполняемого при ее восстановлении, можно записать так:

$$A_в(x,\ y) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \Gamma(\xi,\ \eta)\exp\left[-i\frac{2\pi}{\lambda D}(x\xi + y\eta)\right]d\xi d\eta =$$

$$= w(x,\ y) * \left[h_з(x,\ y)\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\left(\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}[\hat{\Gamma}_Ф(r, s) + c]\times\right.\right.$$

$$\left.\left.\times\delta(\xi + \xi_0 - r\Delta\xi)\,\delta(\eta + \eta_0 - s\Delta\eta)\right)\exp\left[-i\frac{2\pi}{\lambda D}(x\xi + y\eta)\right]dxdy\right] =$$

$$= w(x,\ y) * \left\{h_з(x,\ y)\exp\left[i\frac{2\pi}{\lambda D}(\xi_0 x + \eta_0 y)\right]\times\right.$$

$$\left.\times\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}[\hat{\Gamma}_Ф(r,\ s) + c]\exp\left[-i\frac{2\pi}{\lambda D}(r\Delta\xi x + s\Delta\eta y)\right]\right\}, \tag{10.75}$$

где

$$w(x,\ y) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} W(\xi,\ \eta)\exp\left[-i\frac{2\pi}{\lambda D}(\xi x + \eta y)\right]d\xi d\eta, \tag{10.76}$$

$$h_з(x,\ y) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} H_з(\xi,\ \eta)\exp\left[-i\frac{2\pi}{\lambda D}(\xi x + \eta y)\right]d\xi d\eta. \tag{10.77}$$

Подставив в (10.75) выражение (10.22) для $\hat{\Gamma}_Ф(r,\ s)$, заменив в ней для случая симметрирования удвоением N_1 на $2N_1$, получим

$$A_в(x,\ y) = w(x,\ y) * \left\{h_з(x,\ y)\exp\left[i\frac{2\pi}{\lambda D}(\xi_0 x + \eta_0 y)\right]\times\right.$$

$$\times\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\left\{\sum_{k=0}^{2N_1-1}\sum_{l=0}^{N_2-1}A_0(k,\ l)\exp\left\{i2\pi\left[\frac{(k+u)(r+p)}{2N_1} + \right.\right.\right.$$

$$\left.\left.\left. + \frac{(l+v)(s+q)}{N_2}\right]\right\} + c\right\}\exp\left[-i\frac{2\pi}{\lambda D}r\Delta\xi x + s\Delta\eta y)\right]\Bigg\}, \tag{10.78}$$

или после преобразований:

$$A_в(x,\ y) = w(x,\ y) * \left\{h_з(x,\ y)\exp\left[i\frac{2\pi}{\lambda D}(\xi_0 x + \eta_0 y)\right]\times\right.$$

$$\times\left\{\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\sum_{k=0}^{2N_1-1}\sum_{l=0}^{N_2-1}A_0(k,\ l)\exp\left\{i2\pi\left[\frac{(k+u)p}{2N_1} + \right.\right.\right.$$

$$\left.\left.\left. + \frac{(l+v)q}{N_2}\right]\right\}\delta\left(\frac{k+u}{2N_1} - \frac{\Delta\xi x}{\lambda D} + m\right)\delta\left(\frac{l+v}{N_2} - \right.\right.$$

$$\left.\left. - \frac{\Delta\eta y}{\lambda D} + n\right)\right\} + c\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\delta\left(\frac{\Delta\xi x}{\lambda D} + m\right)\delta\left(\frac{\Delta\eta y}{\lambda D} + n\right)\Bigg\}. \tag{10.79}$$

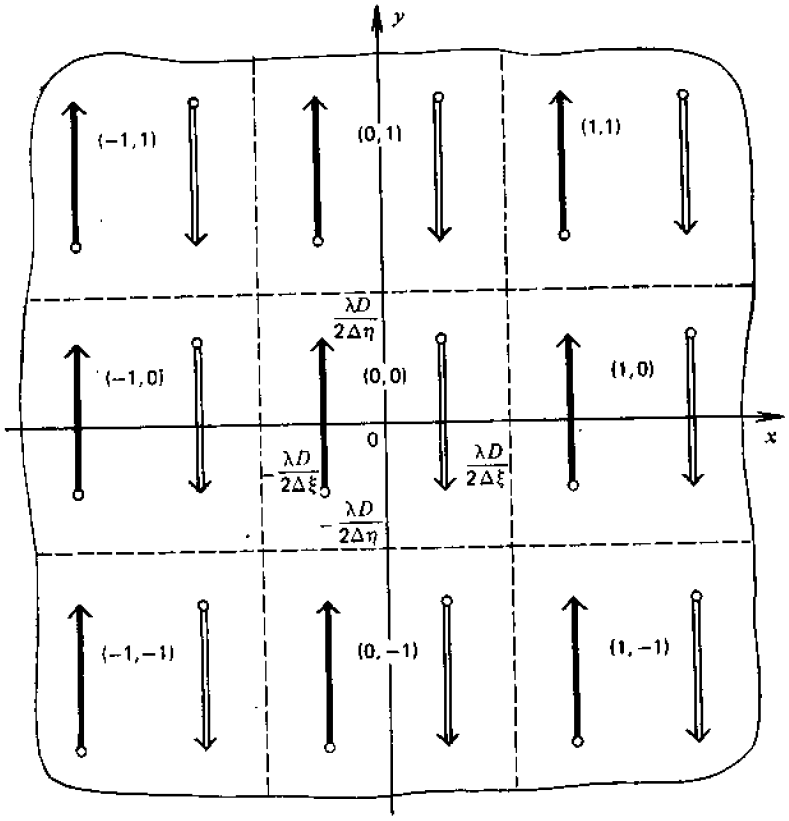


Рис. 10.10. Схема расположения дифракционных порядков при восстановлении голограммы, синтезированной методом симметрирования удвоением

Формула (10.75) показывает, что при синтезе и записи фурье-голограммы следует выбирать $p=q=0,\ \xi_0=\eta_0=0$. Из нее вытекает также, что голограмма, помещенная в оптическую схему Фурье, будет восстанавливать отсчеты исходного распределения поля на объекте в нескольких порядках дифракции (их номер определяется числами тип). Отсчеты маскируются функцией $h_3(x,\ y)$, являющейся преобразованием Фурье апертурной функции $H_3(x,\ y)$ записывающего элемента устройства записи голограмм, и интерполируются функцией, являющейся преобразованием функции окна голограммы $W(\xi,\ \eta)$. Второе слагаемое в фигурных скобках (10.79) описывает так называемое центральное пятно, возникающее вследствие наличия в записанной голограмме постоянной составляющей.

Схема расположения дифракционных порядков на восстановленной картине показана на рис. 10.10 для параметров сдвига $u=-N_1;\ v=-N_2/2$. В скобках на этом рисунке указаны номера дифракционных порядков $(m,\ n)$ для этих параметров сдвига, сплошные и полые стрелки обозначают исходное симметрированное изображение.

Пример изображения, восстановленного с голограммы, записанной по этому методу, показан на рис. 10.3, а.

Метод ортогонального кодирования. В этом случае в соответствии с формулой (10.39)

$$\Gamma(\xi,\ \eta) = W(\xi,\ \eta)\left\{\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\sum_{m_0=0}^{1}\{\mathrm{Re}\{(-1)^r(-\mathrm{i})^{m_0}\hat{\Gamma}_\Phi(r,\ s)\}+c\}H_3(\xi+\xi_0-r\Delta\xi,\ \eta+\eta_0-s\Delta\eta)\right\} = W(\xi,\ \eta)\{H_3(\xi,\ \eta)*$$
$$*\left\{\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\sum_{m_0=0}^{1}\{\mathrm{Re}\{(-1)^r(-\mathrm{i})^{m_0}\hat{\Gamma}_\Phi(r,\ s)\}+c\}\delta(\xi+\xi_0-\right.$$
$$\left.-r\Delta\xi)\,\delta(\eta+\eta_0-s\Delta\eta)\right\}. \qquad (10.80)$$

Действуя так же, как и при выводе формулы (10.79), после достаточно простых выкладок, которые из-за их громоздкости опустим, можно получить, что такая голограмма восстанавливает в схеме непрерывного преобразования Фурье следующую функцию:

$$A_3(x,\ y)=w(x,\ y)*\left\{h_3(x,\ y)\exp\left[i\frac{2\pi}{\lambda D}(\xi_0 x+\eta_0 y)\right]\right\}\times$$
$$\times\left\{\cos\pi\left(\frac{1}{4}+\frac{\Delta\xi x}{\lambda D}\right)\exp\left[-i\pi\left(\frac{1}{4}+\frac{\Delta\xi x}{\lambda D}\right)\right]\times\right.$$
$$\times\left\{\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}A_0(k,\ l)\exp\left\{i2\pi\left[\frac{(k+u)p}{N_1}+\right.\right.\right.$$
$$\left.\left.\left.+\frac{(l+v)q}{N_2}\right]\right\}\delta\left(\frac{k+u}{N_1}+\frac{1}{2}-\frac{2\Delta\xi x}{\lambda D}+m\right)\delta\left(\frac{l+v}{N_2}-\right.\right.$$
$$\left.\left.-\frac{\Delta\eta y}{\lambda D}+n\right)\right\}+\sin\pi\left(\frac{1}{4}+\frac{\Delta\xi x}{\lambda D}\right)\exp\left[i\pi\left(\frac{1}{4}+\frac{\Delta\xi x}{\lambda D}\right)\right]\times$$
$$\times\left\{\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}A_0^*(k,\ l)\exp\left\{-i2\pi\left[\frac{(k+u)p}{N_1}-\right.\right.\right.$$
$$\left.\left.\left.-\frac{(l+v)q}{N_2}\right]\right\}\delta\left(\frac{k+u}{N_1}+\frac{1}{2}+\frac{2\Delta\xi x}{\lambda D}+m\right)\times\right.$$
$$\left.\times\delta\left(\frac{l+v}{N_2}+\frac{\Delta\eta y}{\lambda D}+n\right)\right\}+c\cos\left(\frac{\pi}{\lambda D}\Delta\xi x\right)\times$$
$$\times\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\delta\left(\frac{2\Delta\xi x}{\lambda D}-m\right)\delta\left(\frac{2\Delta\eta y}{\lambda D}-n\right).\qquad(10.81)$$

Очевидно, здесь так же, как и в предыдущем случае, при записи и синтезе фурье-голограммы следует выбирать $\xi_0=\eta_0=0$; $p=q=0$. Восстановленное изображение содержит в нескольких

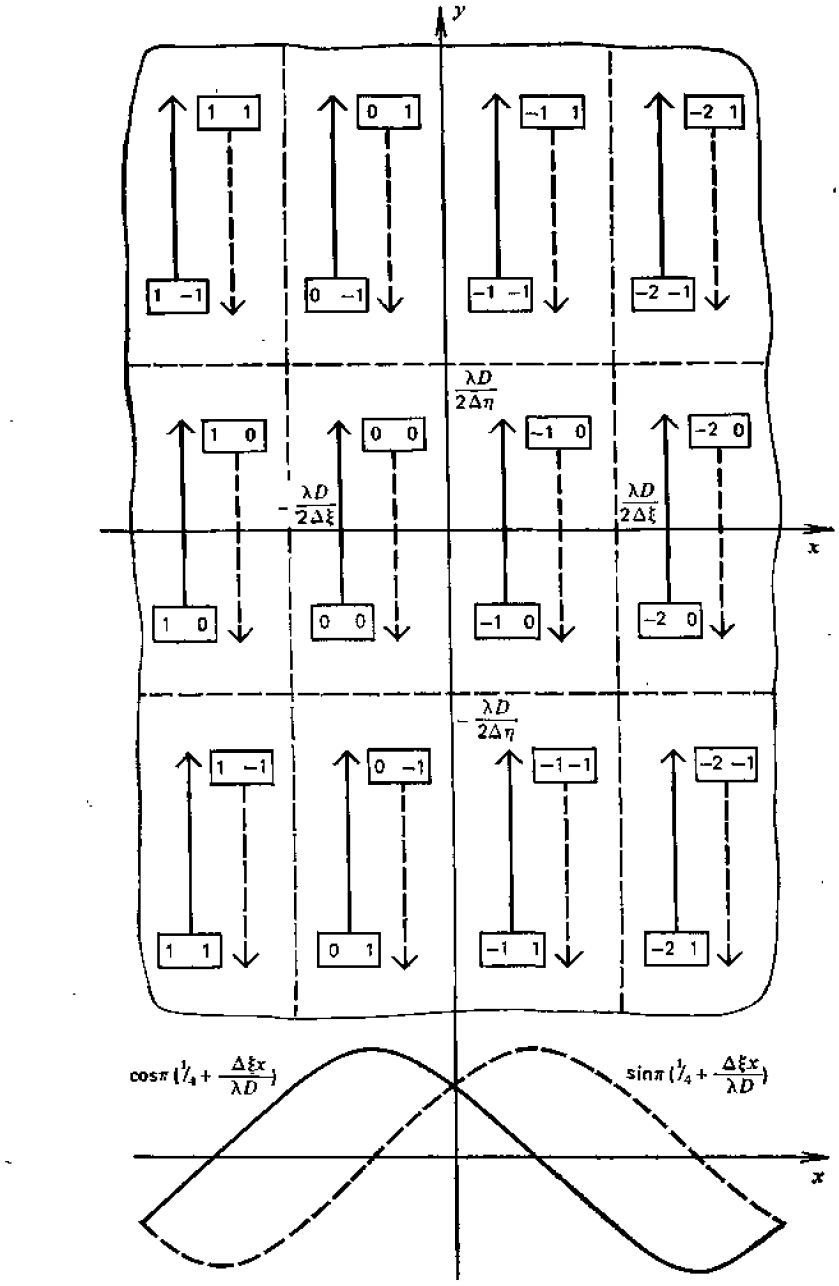


Рис. 10.11. Расположение дифракционных порядков для метода ортогонального кодирования:

сплошные и штриховые стрелки — накладывающиеся друг на друга прямое и сопряженное изображения. В рамке — номер порядка дифракции. Внизу показаны весовые функции прямого и сопряженного изображений

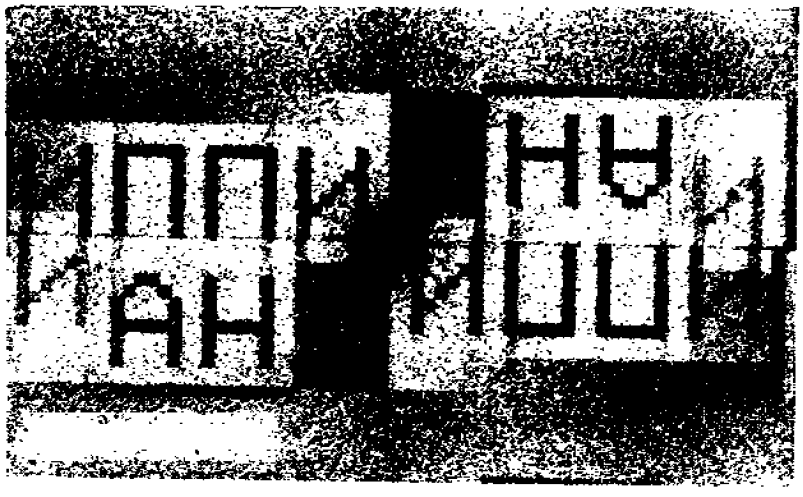


Рис 10.12. Пример восстановления голограммы, записанной методом ортогонального кодирования

дифракционных порядках центральное пятно за счет постоянной составляющей в голограмме [сумм с коэффициентом с в формуле (10.81)] и маскируется функцией $h_3(x, y)$– преобразованием Фурье апертуры записывающего элемента устройства записи голограммы. Но в отличие от предыдущего случая здесь каждый порядок дифракции содержит два наложенных друг на друга изображения объекта: прямое и сопряженное, повернутое на 180° относительно прямого. Каждое из них дополнительно маскируется: прямое – функцией $\cos\pi((1/4)+\Delta\xi x/\lambda D)$, сопряженное – функцией $\sin((1/4)+\Delta\xi x/\lambda D)$. Поэтому в центральной части прямого изображения сопряженное изображение подавлено, но на периферии прямого изображения помеха за счет сопряженного изображения сравнивается с интенсивностью прямого изображения. Этот недостаток метода по сравнению с методом симметрирования объясняется тем, что здесь дополнительные отсчеты голограммы, необходимые для передачи пространственной несущей, получаются ступенчатой интерполяцией отсчетов математической голограммы, тогда как в методе симметрирования интерполяция автоматически получается идеальной.

Картина расположения дифракционных порядков прямого и сопряженного изображения представлена на рис. 10.11 при $u=-N_1/2$; $v=-N_2/2$. На рис. 10.12 показан пример восстановленного изображения. На этом рисунке хорошо заметны прямое и мешающее сопряженное изображение, контраст которого возрастает от центра к периферии по горизонтали.

Двухфазовая запись на фазовой среде. При записи амплитуды и фазы голограммы на фазовой среде по способу, описываемому формулой (10.54):

$$\Gamma(\xi, \eta) = w(\xi, \eta)\left\{H_3(\xi, \eta) * \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\sum_{m_0=0}^{1} \exp\{i[\varphi(r, s) - \right.$$
$$- (-1)^{m_0}\arccos(|\hat{\Gamma}_\Phi(r, s)|/2|\Gamma_0|)]\}\delta(\xi + \xi_0 -$$
$$\left. - (2r + m_0)\Delta\xi)\,\delta(\eta + \eta_0 - s\Delta\eta)\right\}. \qquad (10.82)$$

После преобразования Фурье такая голограмма восстанавливает следующую функцию:

$$A_3(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Gamma(\xi, \eta)\exp\left[-i\frac{2\pi}{\lambda D}(\xi x + \eta y)\right]d\xi d\eta =$$
$$= w(x, y) * \left\{h_3(x, y)\left\{\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\sum_{m_0=0}^{1}\exp\{i[\varphi(r, s) - \right.\right.$$
$$- (-1)^{m_0}\arccos(|\hat{\Gamma}_\Phi(r, s)|/2|\Gamma_0|)]\}\exp\left\{-i\frac{2\pi}{\lambda D}[((2r +$$
$$\left.\left. + m_0)\Delta\xi - \xi_0)x + (s\Delta\eta - \eta_0)y]\right\}\right\} = w(x, y) * \{2h_3(x, y) \times$$
$$\times \exp\left\{i\frac{2\pi}{\lambda D}[(\xi_0 - \Delta\xi/2)x + \eta_0 y]\right\}\left\{\cos(\pi\Delta\xi x/\lambda D) \times\right.$$
$$\times \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}(|\hat{\Gamma}_\phi(r, s)|/2|\Gamma_0|)\exp[i\varphi(r, s)] \times$$
$$\left.\times \exp[-i2\pi(2r\Delta\xi x + s\Delta\eta y)]\right\}\}. \qquad (10.83)$$

Подставив в (10.38) вместо $|\hat{\Gamma}_\Phi(r, s)|\exp[i\varphi(r, s)] = \hat{\Gamma}_\Phi(r, s)$ (10.22) и введя функцию $\tilde{A}_0(k, l)$, определяемую

$$\sqrt{1 - |\hat{\Gamma}_\phi(r, s)|^2/4|\Gamma_0|^2}\exp[i\varphi(r, s)] = \frac{1}{2|\Gamma_0|} \times$$
$$\times \sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}\tilde{A}_0(k, l)\exp\left\{i2\pi\left[\frac{(k+u)(r+p)}{N_1} + \right.\right.$$
$$\left.\left. + \frac{(l+v)(s+q)}{N_2}\right]\right\}, \qquad (10.84)$$

получим после преобразований:

$$A_в(x,\ y)=w(x,\ y)*\left\{2h_3(x,\ y)\exp\left\{-i\frac{2\pi}{\lambda D}\left[\left(\xi_0-\frac{\Delta\xi}{2}\right)x+\right.\right.\right.$$

$$\left.\left.\left.+\eta_0 y\right]\right\}\cos\left(\frac{\pi\Delta\xi x}{\lambda D}\right)\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}\bar{A}_0(k,\ l)\times\right.$$

$$\times\exp\left\{i2\pi\left[\frac{(k+u)p}{N_1}+\frac{(l+v)q}{N_2}\right]\right\}\delta\left(\frac{k+u}{N_1}-\right.$$

$$\left.-\frac{2\Delta\xi x}{\lambda D}+m\right)\delta\left(\frac{l+v}{N_2}-\frac{\Delta\eta y}{\lambda D}+n\right)+\sin\left(\frac{\pi\Delta\xi x}{\lambda D}\right)\times$$

$$\times\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\sum_{k=0}^{N_1-1}\sum_{l=0}^{N_2-1}\tilde{A}_0(k,\ l)\exp\left\{i2\pi\left[\frac{(k+u)p}{N_1}+\right.\right.$$

$$\left.\left.+\frac{(l+v)q}{N_2}\right]\right\}\delta\left(\frac{k+u}{N_1}-\frac{2\Delta\xi x}{\lambda D}+m\right)\delta\left(\frac{l+v}{N_2}-\right.$$

$$\left.\left.-\frac{\Delta\eta y}{\lambda D}-n\right)\right\}. \qquad (10.85)$$

Таким образом, результат восстановления голограммы при двухфазовой записи ее на фазовой среде похож на тот, который получается для голограммы, записанной по методу ортогонального кодирования [см. формулу (10.81)]. Здесь также имеется несколько дифракционных порядков изображения, маскированных функцией $h_3(x, y)$, к наблюдается эффект наложения на исходное изображение (функция $\bar{A}_0(k, l)$) мешающего изображения (функция $\tilde{A}_0(k, l)$). Исходное и мешающее изображения дополнительно маскируются функциями $\cos(\pi\Delta\xi x/\lambda D)$ и $\sin(\pi\Delta\xi x/\lambda D)$, в результате чего помеха в центре изображения ослаблена, а на периферии может достигать той же интенсивности, что и основное изображение. Однако в отличие от записи по методу ортогонального кодирования мешающее изображение не является сопряженным к исходному, а является его искаженной копией. Согласно (10.84) оно имеет тот же фазовый спектр, но искаженный амплитудный спектр. Кроме того, в отличие от голограмм, синтезированных по методу симметрирования или по методу ортогонального кодирования, рассматриваемый двухфазовый способ записи не дает центральных пятен в дифракционных порядках восстановленного изображения, так как голограмма записывается на фазовой среде без постоянной амплитудной составляющей.

Картина расположения дифракционных порядков на восстановленном изображении для этого случая представлена на рис. 10.13 для значений $u=-N_1/2$, $v=-N_2/2$. На рис. 10.14 показан пример восстановления изображения в порядках дифракции (0,0) и (0, 1).

10.5. ПРИМЕНЕНИЕ СИНТЕЗИРОВАННЫХ ГОЛОГРАММ -ДЛЯ ВИЗУАЛИЗАЦИИ ИНФОРМАЦИИ

Главная область практического применения синтезированных голограмм в настоящее время – оптическая обработка информации, где синтезированные голограммы широко используются в качестве элементов оптических процессоров: пространственных фильтров, фокусаторов, дефлекторов, специальных дифракционных решеток и линз [16. 30, 49, 59]. Но имеется и другая, не менее важная, а в известной смысле и более заманчивая область применения синтезированных голограмм – визуализация информации,

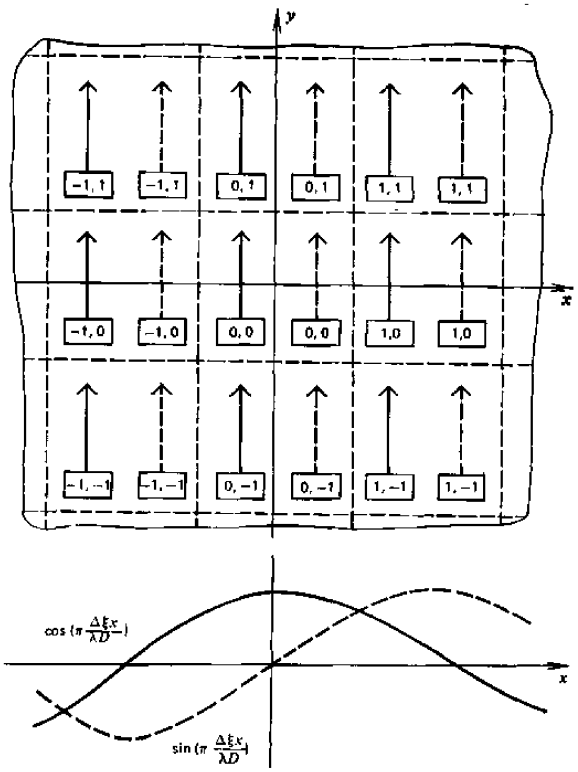


Рис. 10.13. Схема расположения порядков для метода двухфазовой записи:
сплошная и штриховая стрелки — основное и мешающее изображения; цифры в рамках обозначают порядок дифракции. Внизу показаны весовые функции основного и мешающего изображений

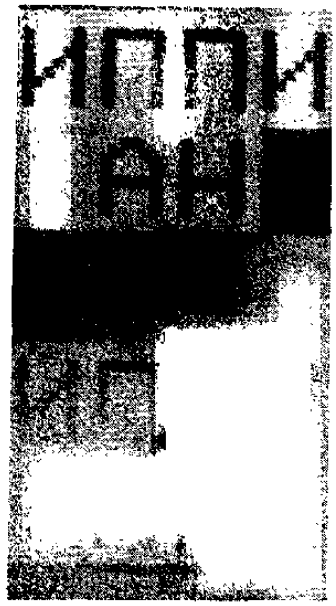


Рис. 10.14. Пример восстановления голограммы, записанной двухфазовым методом

создание голографических дисплеев. Идея создания голографических трехмерных дисплеев на основе синтеза голограмм была высказана уже в первых работах по цифровой голографии. Но только сейчас можно сказать, что имеются реальные предпосылки воплощения этой идеи.

К настоящему времени сложились три основных метода расчета голограмм для визуализации информации: методы многоплановых голограмм, композиционных стереоголограмм и программируемого диффузора.

Метод многоплановых голограмм [16] состоит в том, что трехмерный объект представляется как совокупность нескольких планов (сечений), расположенных на различном удалении от наблюдателя, и расчет голограммы объекта сводится к расчету голограмм отдельных сечений. Известны два варианта этого метода. В первом варианте голограммы отдельных сечений рассчитывают в плоскости наблюдения и затем суммируют для получения полной голограммы. В этом варианте предполагается, что близко расположенные сечения не закрывают более удаленных. В более совершенном втором варианте сначала рассчитывается голограмма Френеля самого удаленного первого сечения в плоскости второго сечения, ближе расположенного к наблюдателю, затем эта голограмма умножается на функцию прозрачности второго сечения, учитывающую, в частности, возможное затенение первого сечения вторым. Для полученного распределения поля рассчитывается голограмма Френеля в третьем сечении и так далее до плоскости наблюдения. Этот вариант метода многоплановых голограмм получил название пинг-понг-метод [16].

Метод композиционных стереоголограмм [49] основан на использовании механизма стереоскопического зрения как одного из главных зрительных механизмов восприятия объема. В соответствии с этим методом голограммы синтезируются отдельно для разных ракурсов тела – разных участков поверхности наблюдения, как это описано в § 10.1. Из этих голограмм можно набрать мозаику, получив тем самым составную, или композиционную голограмму. Тогда, сканируя глазами по этой мозаике, можно наблюдать плавный переход планов, как при наблюдении самого объекта через окно, размеры которого равны размерам композиционной голограммы. Чем больше площадь мозаичной голограммы, тем больше угол обзора объекта.

Очевидно, чтобы такие голограммы было удобно рассматривать, они должны иметь размеры, по крайней мере, в несколько раз превышающие межцентровое расстояние глаз. В то же время размер голограммы, соответствующей одному ракурсу рассматривания, по-видимому, может бъгть выбран равным размеру зрачка, т.е. с учетом движения глаза 10–20 мм. Физический размер элементарной голограммы, необходимой для восстановления объекта с требуемым разрешением, определяется числом элементов раз-

Рис. 10.15. Макет синтезированного кругового голографического фильма

решения на объекте и требуемым углом его наблюдения, т.е. максимальной пространственной частотой на синтезированной голограмме. При числе элементов разрешения 512x512 и максимальной пространственной частоте 100 линий на 1 мм элементарная голограмма будет иметь размер 5x5 мм. Это значит, что голограмму для одного ракурса рассматривания можно получить мозаичным повторением элементарной голограммы. Полная же макроголограмма получается укладкой в нужном порядке мозаик, построенных для каждого ракурса рассматривания. При этом, если для восприятия используется только горизонтальный параллакс, в вертикальном направлении голограммы можно повторять столько раз, сколько это нужно для получения макроголограммы удобных размеров.

На этих принципах были созданы круговые композиционные макроголограммы – синтезированные голографические фильмы [49] (рис. 10.15). Наблюдатель, глядя двумя глазами сквозь голограмму, образующую поверхность цилиндра, видит парящий в пространстве трехмерный объект.

Замечательным свойством композиционных стереоголограмм является то, что они позволяют воспроизводить не только объем, но и движение тела в пространстве. Так, если привести круговую голограмму во вращение, наблюдатель будет видеть вращение тел. Опыты показывают, что иллюзия непрерывного движения не нарушается даже при малой скорости вращения голограмм. Это позволяет говорить о возникновении кинематографического эффекта. Возможность непрерывной подачи кадров-голограмм Фурье, обусловленная инвариантностью преобразования Фурье относительно сдвига, является большим преимуществом голографических фильмов.

В тех случаях, когда воспроизводимые объекты не содержат контуров или деталей, способных создать стереоэффект, для синтеза голограмм можно воспользоваться *методом программируемого диффузора* [49], основанным на моделировании бликов (игры светотени) на диффузных поверхностях тел. Блики возникают при освещении тел направленным светом благодаря особому свойству диффузно отражающих объектов рассеивать падающий на них свет неравномерно по разным направлениям. Вследствие этого интенсивность света, отраженного некоторым участком поверхности объекта в данном направлении, зависит от угла между этим направлением и нормалью к данному участку поверхности, а также от направления на источник освещения.

Чтобы моделировать этот эффект при синтезе голограмм, можно воспользоваться тем, что в задаче визуализации важна передача только макроформы объекта, т.е. неровностей, значительно больших длины волны освещения. Задавшись этой макроформой, можно определить расстояние от каждой точки поверхности объекта до касательной к нему плоскости, перпендикулярной к направлению рассматривания (см. § 10.1). Это расстояние определяет набег фазы волны вследствие макроформы объекта, т.е. «регулярную» составляющую фазы коэффициента отражения, пересчитанного на касательную плоскость. Чтобы передать диффузные свойства поверхности, необходимо дополнить эту «регулярную» составляющую

«случайной», которая описывала бы микроформу поверхности, ее шероховатость. Для того чтобы имитировать неравномерное рассеивание света в разных направлениях, эта «случайная» компонента фазы должна представлять собой коррелированный процесс: ее энергетический спектр (квадрат модуля преобразования Фурье) должен совпадать с угловым распределением интенсивности отраженного света в данном месте диффузной поверхности. Некоррелированная компонента, или диффузор с некоррелированными отсчетами, который используется в большинстве методов синтеза голограмм для имитации диффузной подсветки, соответствует равномерному во всех направлениях рассеиванию света. Для синтеза коррелированного диффузора можно использовать алгоритм, описанный в § 6.2.

Программируемый диффузор открывает принципиальную возможность синтеза голограмм Фурье, содержащих информацию сразу о всех ракурсах объекта и тем самым о его форме. На рис. 10.16 показаны несколько результатов экспериментов с голограммами, синтезированными методом программируемого диффузора.

Для удобства рассматривания из голограммы с программируемым диффузором также целесообразно изготовлять макроголограммы. Но в отличие от композиционных стереоголограмм, здесь мозаичному размножению следует подвергать не всю голограмму, а отдельные ее фрагменты, соответствующие разным ракурсам. Отметим также, что, меняя кратность повторения таких фрагментов, можно произвольно менять масштаб рассматривания объекта по углу обзора.

Все описанные выше способы синтеза голограмм давали голограммы, которые восстанавливаются «на просвет» и в монохроматическом свете. Для целей визуализации удобнее иметь голограммы, которые могли бы восстанавливаться в белом свете я быть отражательными. Для этого они должны быть гибридными, оптико-цифровыми.

Общая идея синтеза гибридных голограмм состоит в том, чтобы производить запись синтезированных голограмм на носитель, который уже содержит сформированную заранее аналоговую голограмму, предназначенную для согласования записываемой голограммы с условиями ее наблюдения и освещения. В процессе записи синтезированная голограмма осуществляет модуляцию этой аналоговой голограммы, так что восстановленный волновой фронт определяется произведением амплитуд волновых фронтов от каждой голограммы. В частности, волновой фронт от аналоговой голограммы может служить в качестве восстанавливающего фронта цифровой голограммы. В этом случае аналоговая голограмма должна быть голограммой точечного источника света.

Если аналоговую голограмму записать во встречных пучках по методу Ю. Н. Денисюка [17], то полученная в результате гибридная голограмма будет восстанавливаться в белом свете и может быть сделана отражательной. Таким образом, гибридная голограмма может сочетать достоинства оптических голограмм – простоту и удобство наблюдения в естественном освещении – с возможностью визуализации объектов, заданных математическим описанием или сигналом.

Для изготовления гибридных голограмм по этому способу необходимы, вообще говоря, специальные устройства записи синтезированных голограмм, способные использовать предварительно экспонированные фотографические материалы. Дело в том, что требования к фотографическим материалам для записи оптических и синтезированных голограмм значительно отличаются. Первые должны иметь очень высокую разрешающую способность (несколько тысяч линий на 1 мм), но могут иметь низкую чувствительность. Вторые могут иметь невысокую разрешающую способность (достаточно несколько сотен линий на 1 мм), но должны быть высокочувствительными, чтобы время записи голограмм было небольшим. Сочетать высокую разрешающую способность и высокую чувствительность в одном материале трудно.

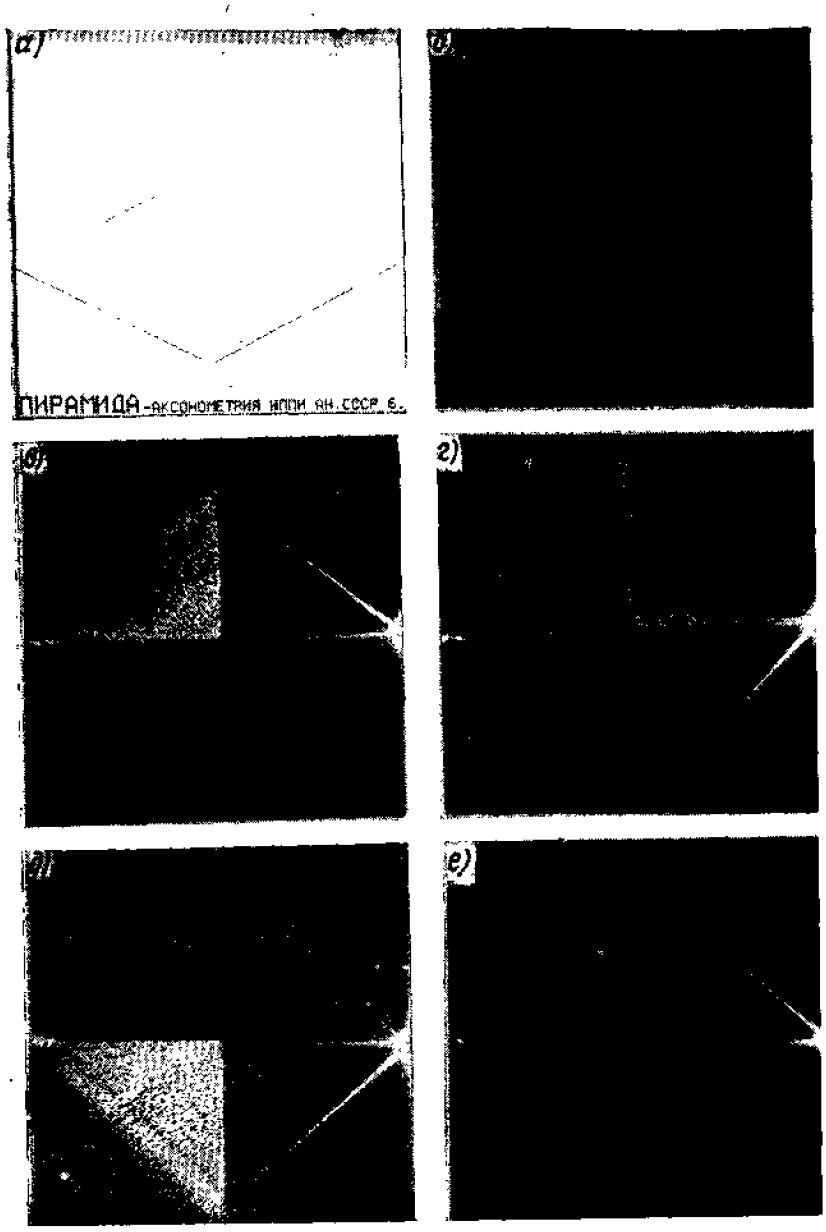


Рис. 10.16. Примеры восстановления изображений с голограммы, синтезированной методом программируемого диффузора:

а — объект (равномерно окрашенная пирамида); *б* — голограмма; *в* — результат восстановления левого верхнего фрагмента голограммы: вид на пирамиду слева сверху; *г* — результат восстановления правого верхнего фрагмента голограммы: вид на пирамиду справа сверху; *д* и *е* — соответственно вид слева снизу и справа снизу

Поэтому для записи оптических и цифровых голограмм в настоящее время приходится использовать разные фотографические материалы. Учитывая это, можно предложить следующие два метода изготовления гибридных голограмм.

Метод фотографической пересъемки. Он заключается в том, что синтезированная голограмма копируется на фотопластинке или фотопленке, предварительно экспонированной оптической голограммой, после чего пластинка (пленка) подвергается фотохимической обработке. Этот метод сопряжен с необходимостью тщательного подбора режимов экспозиции, но он полнее всего реализует идею гибридных голограмм.

Метод сэндвич-голограммы. Этот метод значительно проще в технологии и заключается в том, что синтезированная и оптическая голограммы изготавливаются отдельно, после чего складываются вместе, образуя так называемую *сэндвич-голограмму*. Метод позволяет сочетать киноформ и отбеленную оптическую голограмму, обладающие наивысшей дифракционной эффективностью. Однако эти два метода пока не реализованы из-за трудностей получения хороших оптических голограмм точечного источника, свободных от хроматических аберраций.

Более практичной является *голографическая пересъемка*, при которой получают

оптическую голограмму изображения, восстанавливаемого с синтезированной голограммы.

В [57] описан метод голографической пересъемки применительно к радужным голограммам. Он позволил получить голограмму, содержащую несколько кадров голографического фильма [49], т.е. несколько ракурсов объемного тела. Метод пересъемки на радужную голограмму хорошо согласован с методом композиционных стереоголограмм, так как и в том и в другом игнорируется вертикальный параллакс.

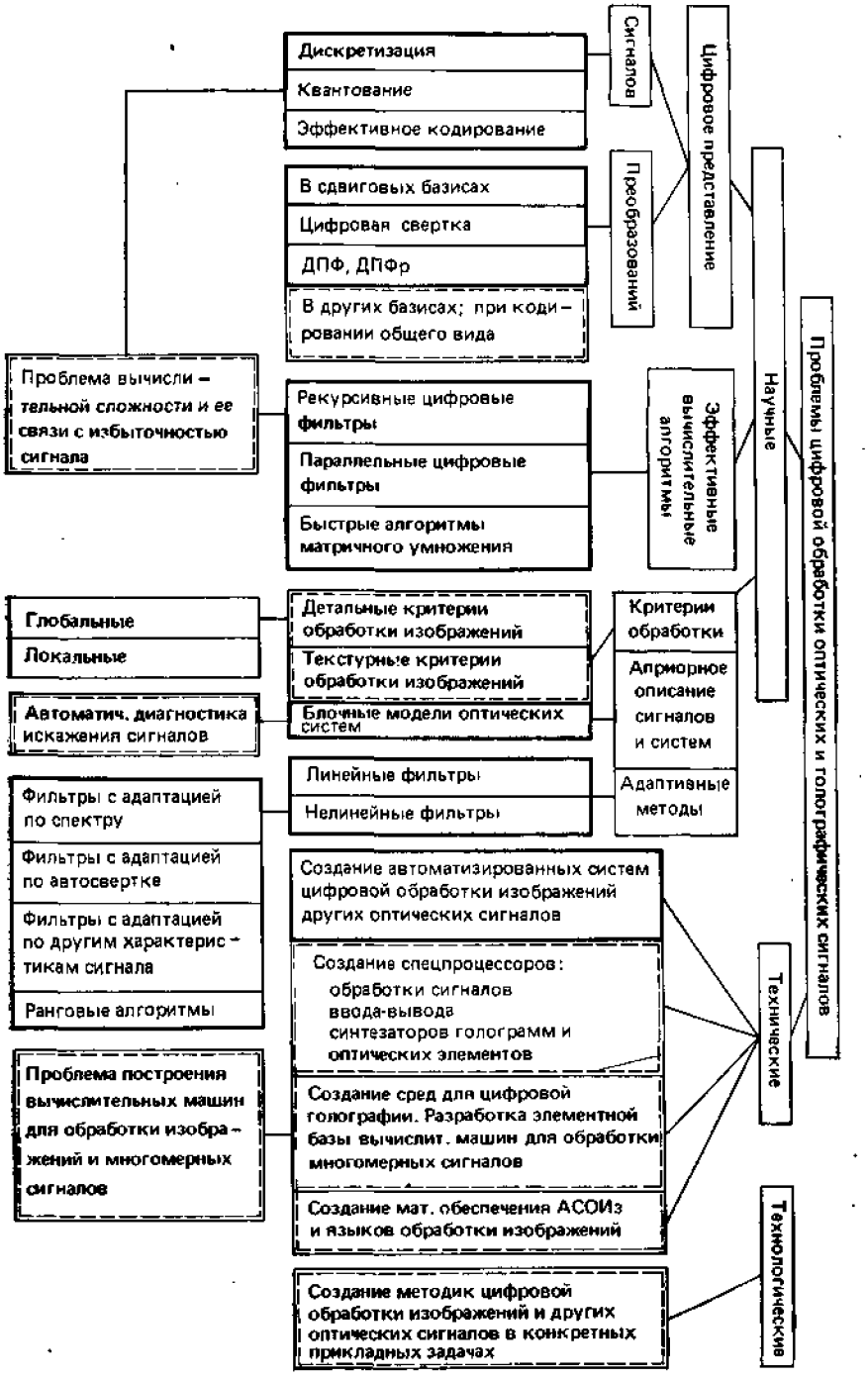
Но наиболее интересную, хотя и наиболее трудоемкую модификацию метода голографической пересъемки предложил Мак-Квигг. Он построил установку, в которой последовательно тремя экспозициями снимаются на объемную среду со сдвигом фазы предметного пучка в 120° друг относительно друга три голограммы транспаранта, на которых записаны три компоненты голограммы, синтезированной по методу Бэркхардта (см. § 10.3). После трех экспозиций полученная объемная голограмма подвергается фотохимической обработке. Изготовленная таким образом гибридная объемная голограмма восстанавливает в белом свете изображение так же, как голограмма Бэркхардта делает это в когерентном свете.

Все известные способы синтеза и записи цветных голограмм предполагают расчет трех отдельных голограмм, соответствующих красному, зеленому и синему цветам объекта, и различаются только записью таких голограмм. Так, в [16] описан способ синтеза трех бинарных голограмм для красного, синего и зеленого цветов с введением трех разных пространственных несущих. Восстанавливаются такие голограммы с помощью трех лазеров (красного, синего и зеленого), освещающих голограмму под соответствующими разными углами так, чтобы восстановленные цветоделенные изображения совместились друг с другом, давая цветное изображение. Фьенуп и Гудмен предложили для изготовления цветных синтезированных голограмм фотографировать на цветную фотопленку через красный, синий и зеленый светофильтры бинарные голограммы с экрана электронно-лучевой трубки. При этом каждая голограмма благодаря светофильтрам записывается в свой слой фотопленки. Для восстановления цветных изображений с таких голограмм используется лазер, имеющий три линии излучения – красную, синюю и зеленую, а цветное изображение формируется в фокальной плоскости линзы, выполняющей преобразование Фурье. При освещении таким лазером каждый слой фотопленки избирает свою компоненту луча и восстанавливает свое изображение. Поскольку спектральная избирательность красителей цветных фотографических эмульсий невысока, возможны искажения цветов вследствие влияния слоев. Эти искажения можно уменьшить, если кроме записи голограмм в разные слои производить также пространственное разделение цветоделейных голограмм либо путем их сдвига друг относительно друга, либо путем пространственного чередования элементов голограмм. Кроме того, это взаимное влияние может быть скомпенсировано соответствующей коррекцией значений амплитуды и фазы голограммы, записываемой в каждый слой.

Для упрощения подбора режимов экспонирования цветной пленки, а также для изготовления цветных голограмм, содержащих большое количество отсчетов (макроголограмм), можно контактным способом фотографически копировать на цветную фотопленку через цветные светофильтры три цветоделенные голограммы, записанные одним из известных способов на черно-белую фотопленку [49]. Получающиеся таким образом голограммы можно использовать также для непосредственного визуального наблюдения цветного объекта, если в качестве осветителя взять точечный источник белого света. Благодаря появлению устройств для фотозаписи цветных изображений (см. [49]) запись цветных макроголограмм можно осуществить непосредственно на цветной фотопленке без промежуточной записи цветоделенных голограмм на черно-белой пленке.

ЗАКЛЮЧЕНИЕ

В данной книге изложны основы цифровой оптики и на отдельных примерах показаны возможности ее практического применения. В заключение полезно попытаться сформулировать основные научные, технические и технологические проблемы этого направления, как они вырисовываются автору, и представить пути его дальнейшего развития. Эти проблемы и их взаимосвязь иллюстрируются рисунком на с. 287. Проблемы, в той или иной мере уже решенные и нашедшие отражение в книге, указаны в сплошных



прямоугольниках. Нерешенные проблемы указаны в штриховых прямоугольниках. Рассмотрим некоторые из них.

Научные проблемы – это проблемы цифрового представления сигналов и их преобразований, а также проблема эффективных вычислительных алгоритмов. В настоящее время цифровое представление сигналов и преобразований основано на двухступенчатой процедуре дискретизации – поэлементном квантовании, причем для дискретизации используются сдвиговые базисы. Из работ по кодированию изображений известно, что, применяя иные базисы, можно добиться более эффективного цифрового представления сигналов. Способы же представления преобразований сигналов, в частности интегральных преобразований, в других дискретных базисах, а также в общем случае эффективного кодирования сигналов, не разработаны.

В области разработки эффективных вычислительных алгоритмов обработки изображений и других многомерных сигналов главные усилия были направлены на преодоление «проклятия размерности». Выход найден в рекурсивных и параллельных цифровых фильтрах, быстрых алгоритмах преобразований, полиномиальных алгоритмах двумерных ДПФ и свертки. Однако до сих пор неизвестно, какова минимальная

вычислительная сложность того или иного вида обработки данного класса сигналов, как эта сложность связана с характеристиками обработки и с избыточностью цифрового представления сигнала, хотя имеется много конкретных примеров использования избыточности сигнала для сокращения числа операций, необходимых для его обработки.

Главное направление в разработке методов цифровой обработки изображений, оптических и голографических сигналов – разработка адаптивных методов обработки. Сформулированы и продолжают развиваться критерии адаптивной обработки, модели сигналов и систем, методы автоматической диагностики искажений сигналов, линейные фильтры с адаптацией параметров по спектру наблюдаемого сигнала, по значениям его автосвертки, по различным другим характеристикам сигнала. Из нелинейных адаптивных методов перспективными являются методы ранговой фильтрации, основанные на измерении характеристик локальных гистограмм значений сигнала.

Технические проблемы – это проблемы создания аппаратурного и математического обеспечения цифровой обработки оптических сигналов. Автоматизированные системы обработки изображений уже стали реальностью и в разных модификациях производятся промышленностью, выпускающей информационную технику. Для их совершенствования требуется разработка специализированных языковых средств обработки изображений и других многомерных сигналов, а также создание специализированных цифровых процессоров ввода-вывода оптических сигналов, синтезаторов голограмм и оптических элементов, процессоров обработки сигналов. Для расширения практического применения методов цифровой голографии необходимо создание специальных сред для записи синтезированных голограмм, фильтров и других оптических элементов.

Накопленный опыт разработки автоматизированных систем обработки изображений, создания эффективных вычислительных алгоритмов обработки приводит еще к одному важному выводу: многие, если не все трудности создания систем и эффективных вычислительных алгоритмов обработки изображений и аналогичных двумерных и многомерных сигналов, связаны с тем, что структура и набор элементарных операций (система команд) универсальных цифровых вычислительных машин, используемых в качестве центрального процессора таких систем, плохо согласована со структурой обрабатываемых данных (сигналов) и особенностями задач обработки. Отсюда вытекает необходимость разработки и создания нового класса цифровых вычислительных машин для обработки изображений и других многомерных сигналов.

Можно сформулировать следующие принципы построения таких вычислительных машин.

1. Память команд и .память данных этих машин должны быть разделены ввиду различной природы упорядоченности и структуры данных (сигналов) и команд обработки.
2. Память данных должна быть организована как двумерная (многомерная) векторная память с параллельным произвольным доступом. Параллельный доступ – для параллельной обработки локальных окрестностей сигнала. Произвольный доступ – для сканирования по сигналу параллельной апертурой.
3. Вычислительный процесс обработки изображений должен строиться по принципу параллельно-последовательной обработки: параллельная обработка локальных окрестностей и последовательное сканирование по всему изображению.
4. Арифметическое устройство должно строиться как иерархия специализированных процессоров с параллельно-последовательной адресацией к памяти данных.

Возможный состав процессоров является следующим.

Сверточный процессор – для вычисления двумерной цифровой свертки векторного двумерного (многомерного) сигнала. В режиме параллельно-последовательной обработки может быть организован как рекурсивный.

Спектральный процессор – для выполнения двумерной (многомерной) цифровой векторной фильтрации в спектральной области. Может быть построен как универсальный процессор преобразования сигналов по произвольному ортогональному базису с быстрым алгоритмом на основе представления матриц преобразований как поэтажно-кронекеровских.

Гистограммный процессор – определяющий локальные гистограммы скалярного (компоненты векторного) видеосигнала и его ранговые статистики и реализующий на этой

основе ранговые алгоритмы обработки.

Процессор бинарного сигнала – выполняющий логические операции над бинарным сигналом для обработки бинарных препаратов и изображений и принятия решений.

Процессор ввода сигнала – гибридный аналого-цифровой процессор, осуществляющий предварительную аналоговую обработку сигналов изображений и полей (пространственную фильтрацию, подавление помех, коррекцию искажений, преобразование координат для изменения вида проекции изображения), а также преобразование полученного сигнала в цифровой сигнал.

Дисплейный процессор – гибридный цифро-аналоговый процессор, осуществляющий визуализацию черно-белых, цветных, объемных и подвижных изображений, преобразование вида проекции и формы представления изображений для удобства визуализации, а также поддерживающий диалог пользователя и связь пользователя с вычислительной машиной.

Выходной синтезатор – гибридный цифро-аналоговый процессор для преобразования цифрового сигнала в документальные изображения, голограммы и оптические элементы.

Процессор базы данных изображений и полей – осуществляющий кодирование и декодирование сигнала изображений и полей для их экономного хранения в базе изображений и обеспечивающий доступ вычислительной машины обработки изображений к базе данных.

Управляющий процессор – связывающий воедино всю систему, задающий режим работы и связи с памятью остальных процессоров.

Наконец, актуальной проблемой является проблема создания технологии обработки изображений и полей, т.е. методик цифровой обработки в конкретных прикладных задачах экспериментальных научных исследований, исследований природных ресурсов Земли, неразрушающего контроля промышленных изделий, медицинской диагностики.