МФТИ

Г. Р. Локшин

ДИФРАКЦИЯ. ПРОСТРАНСТВЕННАЯ ФИЛЬТРАЦИЯ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

Г. Р. Локшин

Дифракция. Пространственная фильтрация

Издание 2-е, исправленное и дополненное

Рекомендовано
Учебно-методическим объединением
высших учебных заведений Российской Федерации
по образованию в области прикладных математики и физики
в качестве учебного пособия для студентов вузов
по направлению подготовки «Прикладные математика и физика»

МОСКВА
МФТИ
2016

УДК 535(075)
ББК 22.343я73
Л73

Рецензенты:
доктор физико-математических наук, профессор *С. Г. Каленков*
доктор физико-математических наук, профессор *Г. И. Соломахо*

Локшин, Г. Р.
Л73 Дифракция. Пространственная фильтрация: учебное пособие по курсу Общая физика. Изд. 2-е, испр. и доп. — М. : МФТИ, 2016. — 156 с.
ISBN 978-5-7417-0590-2

Изучение волновых оптических явлений и законов преобразования света в оптических системах основано на использовании закономерностей, управляющих работой линейных колебательных систем. Вопросы дифракции и формирования оптического изображения, а также принципы пространственной фильтрации и оптической обработки информации рассмотрены на основе естественного обобщения и развития принципов линейной фильтрации электрических сигналов. Обсуждаются методы улучшения качества изображения и наблюдения фазовых объектов, согласованная фильтрация и задача распознавания образов, способы получения голограмм без опорного пучка и другие задачи, основанные на принципах пространственной фильтрации.

Пособие предназначено как для студентов старших курсов, изучающих основы фурье-оптики, так и для студентов II курса, желающих глубже ознакомиться с рядом общефизических вопросов оптики.

УДК 535(075)
ББК 22.343я73

ISBN 978-5-7417-0590-2

Оглавление

Светлой памяти
профессора Б.Н. Митяшева
посвящается эта книга

Предисловие

Книга является введением в радиооптику — науку, которая использует язык и методы теории линейных колебаний для описания оптических явлений. Рассмотрены вопросы дифракции и формирования оптического изображения, а также принципы пространственной фильтрации и оптической обработки информации. Для более детального и глубокого ознакомления с методами радиооптики рекомендуем читателю обратиться к монографиям [2–13]. Блестящее изложение затронутых в книге вопросов можно найти также в статье С.М. Рытова [1]. Для первоначального знакомства с некоторыми вопросами, рассмотренными в книге, отсылаем читателя к [14, §§ 5, 6].

Книга будет полезна не только студентам старших курсов, приступающим к изучению радиооптики, но и студентам II курса, желающим глубже ознакомиться с рядом общефизических вопросов (дифракция, разрешающая способность и т. д.), входящих в программу общего курса физики. Студентам III курса книга поможет при выборе вопроса для Госэкзамена по физике.

В течение ряда лет автор читал курс лекций по оптической обработке информации для студентов IV–V курсов ФРТК. Программа этого курса подробно обсуждалась с проф. Б.Н. Митяшевым, что в значительной мере помогло автору при написании настоящей книги.

Автор считает своим приятным долгом поблагодарить проф. С.М. Козела, доц. В.Е. Белонучкина и доц. Д.А. Александрова за полезные обсуждения и замечания, сделанные ими при чтении рукописи.

Глава I

Принципы линейной фильтрации. Спектральный анализ линейных систем

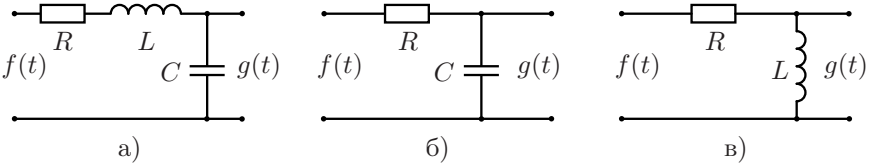
§ 1.1. Введение

Дифракция, оптическое изображение, разрешающая способность, пространственная фильтрация и голография — вот круг вопросов, рассматриваемых в настоящей книге. Из всего многообразия оптических явлений выделены лишь те, в которых особенно ярко проявляется волновая природа света.

В последние десятилетия интенсивно разрабатывается подход к изучению волновых явлений, основанный на понимании закономерностей, управляющих процессами в линейных колебательных системах (таких, как электрический колебательный контур, механический маятник и т. д.). Возникла новая дисциплина, называемая *радиооптикой* (или *фурье-оптикой*), которая использует язык и методы теории колебаний для решения волновых задач. Выявляющиеся при таком подходе колебательно-волновые аналогии позволяют глубже разобраться во многих оптических явлениях.

В силу сказанного выше, читателя не должно удивлять, что в первой главе обсуждаются вопросы, далёкие, на первый взгляд, от оптики. Прежде всего мы рассмотрим электрический колебательный контур, причём под таким углом зрения, чтобы это облегчило более глубокое понимание многих вопросов оптики, так, чтобы в дальнейшем выявились аналогии между задачей возбуждения колебаний в электрическом контуре и задачами дифракции, формирования изображения и оптической фильтрации.

§ 1.2. Линейные фильтры



а) б) в)

Рис. 1.1

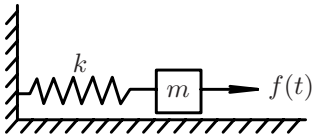


Рис. 1.2

На рис 1.1а изображён колебательный контур, содержащий сопротивление R, индуктивность L и ёмкость C. Контур находится под действием ЭДС, меняющейся со временем по закону $f(t)$. Нас интересует закон изменения напряжения на конденсаторе. Можно было бы исследовать RC-цепочку (рис. 1.1б) или LR-цепочку (рис. 1.1в), или какую-либо другую электрическую цепь, содержащую сопротивления, ёмкости и индуктивности. Мы могли бы, наконец, изучать механическую колебательную систему (рис. 1.2), находящуюся под действием внешней силы $f(t)$, исследуя зависимость координаты тела от времени.

Общим в приведённых примерах является то, что искомые функции определяются *линейными уравнениями*, т. е. уравнениями, в которых каждое слагаемое содержит неизвестную функцию и её производные только в первой степени. Действительно, напряжение на конденсаторе колебательного контура, согласно закону Ома, подчиняется уравнению

$$\ddot{g} + 2\gamma\dot{g} + \omega_0^2 g = \omega_0^2 f(t), \tag{1.1}$$

где $\gamma = \frac{R}{2L}$ — затухание, $\omega_0 = \frac{1}{\sqrt{LC}}$ — собственная частота. Координата тела в механическом осцилляторе (рис. 1.2) определяется вторым законом Ньютона:

$$m\ddot{x} + kx = f(t). \tag{1.2}$$

Читатель без труда может написать уравнения, определяющие искомые функции $g(t)$ для LR- и RC-фильтров, изображённых на рис. 1.1б, в, и убедиться, что и в этих примерах уравнения оказываются линейными.

Важно отдавать себе отчёт в том, в результате каких принятых идеализаций возникает то или иное уравнение, почему уравнения (1.1) и

(1.2) оказались линейными. При достаточно больших внешних ЭДС сопротивление контура перестаёт быть постоянной величиной; оно начинает зависеть от величины тока, при этом линейность уравнения (1.1) нарушается. Если внешняя сила, действующая на механический осциллятор, велика, то растяжение пружины оказывается большим — перестаёт выполняться закон Гука (линейность зависимости упругой силы от величины деформации), при этом нарушается линейность уравнения (1.2).

Отмеченный факт является общим для всех физических явлений: *линейность нарушается при достаточно сильных внешних воздействиях.*

Мы ограничимся сделанными замечаниями и будем полагать в дальнейшем, что внешние воздействия достаточно малы и не нарушают линейности рассматриваемых колебательных систем.

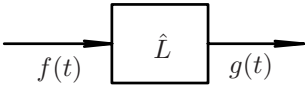


Рис. 1.3

При изучении общих свойств линейных систем (фильтров) обычно не интересуются их конкретным устройством и изображают с помощью блок-схемы (рис. 1.3). Внешнее воздействие $f(t)$ называют *входным сигналом фильтра*, а искомую зависимость $g(t)$ — *выходным сигналом* или *откликом фильтра*. Квадратик $\boxed{\hat{L}}$ представляет собой некоторое устройство, преобразующее входной сигнал $f(t)$ в выходной сигнал $g(t)$.

Тот факт, что $g(t)$ является откликом фильтра на входное воздействие $f(t)$, будем записывать в виде операторного равенства

$$g(t) = \hat{L}[f(t)]. \tag{1.3}$$

Фундаментальное свойство всех линейных физических явлений (т. е. явлений, описываемых линейными уравнениями) состоит в следующем: *результат нескольких одновременных воздействий можно найти, суммируя результаты, к которым приводит каждое отдельное воздействие.* Это — наиболее общая формулировка принципа линейной суперпозиции. Обратимся к приведённым выше примерам. Пусть $g_n(t)$ — напряжение на конденсаторе контура, находящегося под действием ЭДС $f_n(t)$ $(n = 1, \ \ldots \ N)$. Легко убедиться, что если внешняя ЭДС есть

$$f(t) = \sum_{n=1}^{N} c_n f_n(t),$$

то напряжение на конденсаторе будет меняться по закону

$$g(t) = \sum_{n=1}^{N} c_n g_n(t).$$

Точно так же зависимость координаты от времени при воздействии на механический осциллятор суммарной силы $f(t) = \sum c_n f_n(t)$ есть $x(t) = \sum c_n x_n(t)$, где $x_n(t)$ — отклонение под действием силы $f_n(t)$.

Можно сформулировать принцип линейной суперпозиции по отношению к произвольному фильтру следующим образом: если отклик фильтра на входной сигнал $f_1(t)$ есть $g_1(t)$, то есть $g_1(t) = \hat{L}[f_1(t)]$, а при входном воздействии $f_2(t)$ выходной сигнал есть $g_2(t)$: $g_2(t) = \hat{L}[f_2(t)]$, то отклик фильтра на линейную комбинацию $f(t) = c_1 f_1(t) + c_2 f_2(t)$ есть линейная комбинация откликов $g_1(t)$ и $g_2(t)$: $g(t) = c_1 g_1(t) + c_2 g_2(t)$. Таким образом, свойство линейности выражается равенством

$$\hat{L}[c_1 f_1(t) + c_2 f_2(t)] = c_1 \hat{L}[f_1(t)] + c_2 \hat{L}[f_2(t)]. \tag{1.4}$$

Можно рассмотреть линейную суперпозицию более общего вида, когда сигнал $f(t)$ выражается непрерывной суперпозицией функций $\varphi(t,\alpha)$ (α — непрерывно меняющийся параметр):

$$f(t) = \int c(\alpha)\varphi(t,\alpha)\, d\alpha.$$

Линейность фильтра означает тогда, что

$$\hat{L}[f(t)] = \int c(\alpha)\hat{L}[\varphi(t,\alpha)]\, d\alpha. \tag{1.5}$$

Итак, свойство линейности позволяет предложить следующий алгоритм решения любой (линейной) физической задачи. Произвольное (вообще говоря, сложное) внешнее воздействие нужно представить в виде линейной суперпозиции более простых воздействий и искать решение, соответствующее каждому слагаемому в этой суперпозиции. Искомое решение находится как линейная суперпозиция решений, соответствующих каждому слагаемому внешнего воздействия.

§ 1.3. Гармонические колебания в линейных системах

Мы видим, что все линейные задачи, независимо от физического содержания, имеют общую схему решения, которая включает в себя представление внешнего воздействия в виде суперпозиции некоторых

более простых элементарных воздействий. Конечно, такое представление неоднозначно. Выбор «базиса» (элементарных слагаемых) — это вопрос физической целесообразности, вопрос об отношении сигнала к той реальной физической системе, на которую сигнал воздействует. Особую роль по отношению к линейным системам играют гармонические функции

$$S(t) = a\cos(\omega t + \varphi), \tag{1.6}$$

где ω — частота колебания, a — амплитуда, $\Phi(t) = \omega t + \varphi$ — фаза колебания, φ — начальная фаза, равная значению $\Phi(t)$ при $t = 0$.

Будем пользоваться комплексной формой записи, рассматривая вместо реальных функций вида (1.6) комплексные функции

$$f(t) = ae^{i(\omega t+\varphi)} = ce^{i\omega t}. \tag{1.7}$$

Комплексное число $c = ae^{i\varphi}$ определяет и амплитуду колебания, и его начальную фазу. Представление (1.7) чрезвычайно упрощает все линейные операции (сложение и умножение на число, а также дифференцирование и интегрирование), поскольку они оставляют неизменным вид множителя $e^{i\omega t}$:

$$\frac{d}{dt}e^{i\omega t} = i\omega e^{i\omega t}, \qquad \int e^{i\omega t}\,dt = \frac{1}{i\omega}e^{i\omega t}.$$

Именно такие операции и производятся с функциями в линейных уравнениях. Таким образом, все линейные преобразования удобно проводить с функциями (1.7), а затем в конечном результате взять реальную часть полученного выражения.

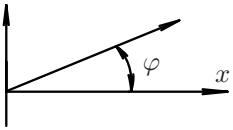


Рис. 1.4

В дальнейшем мы будем широко пользоваться векторными диаграммами, изображая гармоническое колебание (1.6) в виде вектора, длина которого равна a, а угол между действительной осью x и направлением вектора есть φ (рис. 1.4). Удобство и целесообразность векторного представления гармонических колебаний состоит в том, что сумма колебаний вида (1.6) может быть найдена по правилу сложения векторов, так что колебание $S(t) = S_1(t)+S_2(t)$ (где $S_1 = a_1\cos(\omega t+\varphi_1)$ и $S_2 = a_2\cos(\omega t + \varphi_2)$) изображается в виде вектора $\boldsymbol{S}$, равного сумме векторов $\boldsymbol{S}_1$ и $\boldsymbol{S}_2$ (рис. 1.5).

Действительно, проекции векторов $\boldsymbol{S}_1$ и $\boldsymbol{S}_2$ на ось x равны соответственно $a_1\cos\varphi_1$ и $a_2\cos\varphi_2$. При этом проекция суммарного вектора $\boldsymbol{S}$ равна, очевидно, сумме $a_1\cos\varphi_1 + a_2\cos\varphi_2$. Теперь представим, что векторы $\boldsymbol{S}_1$ и $\boldsymbol{S}_2$ вращаются с угловой скоростью ω против часовой

стрелки. При этом их проекции в момент времени t равны соответственно $s_1 = a_1\cos(\omega t + \varphi_1)$ и $s_2 = a_2\cos(\omega t + \varphi_2)$, а проекция суммарного вектора $\boldsymbol{S}$ есть $S(t) = a_1\cos(\omega t + \varphi_1) + a_2\cos(\omega t + \varphi_2)$.

Отметим, что комплексное число $Ae^{i\varphi}$ также можно изобразить в виде вектора на комплексной плоскости (x — действительная ось, y — мнимая ось), так что проекция вектора на ось x равна действительной части комплексного числа, а его проекция на ось y — мнимой части; модуль комплексного числа равен при этом длине вектора, а его аргумент φ — углу с действительной осью. Так же, как и гармонические колебания, комплексные числа можно складывать по правилу сложения векторов. Таким образом, комплексное представление гармонических колебаний и их векторное изображение полностью эквивалентны.

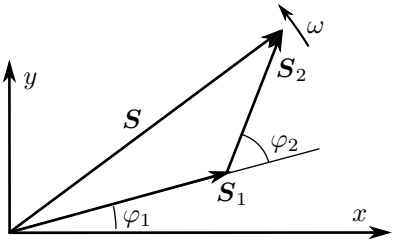


Рис. 1.5. Положение векторов $\boldsymbol{S}_1$ и $\boldsymbol{S_2}$ в момент времени $t = 0$

Итак, нас интересует, какова реакция линейной колебательной системы на внешнее гармоническое воздействие $f(t) = ae^{i\omega t}$. Обратимся к колебательному контуру. Напряжение на конденсаторе (искомый выходной сигнал $g(t)$) подчиняется уравнению

$$\ddot{g} + 2\gamma\dot{g} + \omega_0^2 g = a\omega_0^2 e^{i\omega t}, \tag{1.8}$$

которое описывает процесс вынужденных колебаний в контуре. Решение задачи (1.8) читателю хорошо известно: под действием внешней гармонической ЭДС частоты ω в контуре возникают «вынужденные» гармонические колебания той же частоты ω, но с амплитудой и фазой, отличными от амплитуда и фазы внешней ЭДС.

Чтобы убедиться в этом, попробуем решение уравнения (1.8) искать в виде

$$g(t) = aH(\omega)e^{i\omega t}. \tag{1.9}$$

Функция (1.9) представляет собой гармоническое колебание частоты ω, амплитуда этого колебания равна $|aH(\omega)|$, а сдвиг по фазе относительно внешнего воздействия есть $\arg H(\omega)$. Подставляя (1.9) в уравнение (1.8), получаем (так как $\dot{g}(t) = i\omega aH(\omega)e^{i\omega t}$; $\ddot{g}(t) = -\omega^2 aH(\omega)e^{i\omega t}$)

$$H(\omega) = \frac{\omega_0^2}{\omega_0^2 - \omega^2 + 2i\gamma\omega}. \tag{1.10}$$

Таким образом, функция (1.9) действительно представляет собой решение уравнения (1.8), т. е. $g(t)$ — выходной сигнал фильтра, если $H(\omega)$

определяется равенством (1.10). Функция $H(\omega)$ называется частотной характеристикой колебательного контура. (Найдите в качестве упражнения частотные характеристики RC- и LR-цепочек, изображённых на рис. 1.1б и в, а также частотную характеристику механического осциллятора (рис. 1.2).)

Смысл частотной характеристики в самом общем случае произвольного фильтра состоит в том, что колебание $g(t)$, определяемое равенством (1.9), является откликом на гармоническое внешнее воздействие $ae^{i\omega t}$. Функцию $|H(\omega)|$, определяющую амплитуду вынужденных колебаний, называют *амплитудной характеристикой фильтра*, а функцию $\arg H(\omega)$, определяющую фазовый сдвиг относительно внешнего воздействия, — *фазовой характеристикой*.

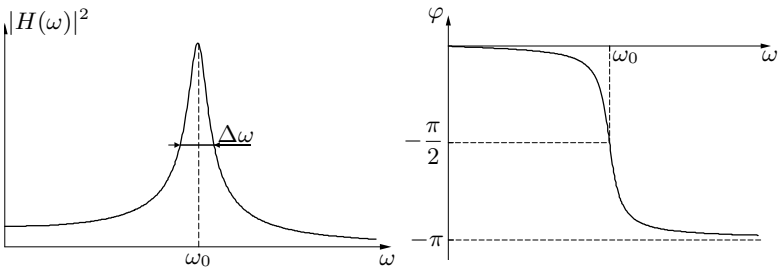


Рис. 1.6

На рис. 1.6 изображены резонансная кривая колебательного контура $|H(\omega)|^2$ (зависимость квадрата амплитуды вынужденных колебаний от частоты внешней ЭДС ω) и его фазовая характеристика. Как видно из рис. 1.6а, амплитуда вынужденных колебаний особенно велика, если частота внешней ЭДС лежит внутри интервала $\Delta\omega$, отмеченного на рисунке, — именно на эти частоты контур хорошо «резонирует» (как камертон на определённый звуковой тон). Читателю представляется возможность написать уравнения кривых $|H(\omega)|^2$ и $\arg H(\omega)$ для колебательного контура, а также для LR- и RC-фильтров.

Отметим, что к уравнению затухающего осциллятора (1.1), описывающего процессы в колебательном контуре, приводят множество разнообразных физических задач, в частности задача о колебаниях электрона в атоме, находящемся во внешнем электрическом поле; с помощью модели затухающего осциллятора, из которой следуют изображённые на рис. 1.6 зависимости, могут быть поняты такие оптические явления, как абсорбция и дисперсия.

Мы рассмотрели примеры линейных систем, колебания в которых возбуждаются внешним гармоническим воздействием, и можем теперь сформулировать общее правило.

Результат воздействия гармонического колебания $e^{i\omega t}$ на линейную колебательную систему есть $H(\omega)e^{i\omega t}$, т. е. гармоническое колебание той же частоты. Этот общий для всех линейных систем факт можно записать символически в виде равенства

$$\hat{L}[e^{i\omega t}] = H(\omega)e^{i\omega t}. \tag{1.11}$$

Функции $e^{i\omega t}$, удовлетворяющие равенству (1.11), математики называют *собственными функциями линейного фильтра.* Именно этот факт и определяет исключительную роль гармонических колебаний по отношению к линейным системам — реакция этих систем на гармоническое воздействие определяется наиболее простым образом: сигнал любой другой формы, отличный от гармонического колебания, возбуждает в линейной системе процессы, не совпадающие по форме с входным воздействием.

Вспоминая теперь основное свойство линейных систем — принцип линейной суперпозиции, — мы видим, что если бы произвольный входной сигнал $f(t)$ мог быть представлен в виде линейной суперпозиции гармонических колебаний с правильно подобранными амплитудами и фазами, то все проблемы были бы в принципе решены: суммируя отклики на каждую гармоническую составляющую суммарного сигнала $f(t)$, мы решили бы задачу о возбуждении линейного фильтра произвольным сигналом.

§ 1.4. Преобразование Фурье

Итак, вопрос состоит в том, можно ли произвольный сигнал $f(t)$ представить в виде линейной суперпозиции гармонических сигналов $e^{i\omega t}$. Оказывается, можно. Как доказывают математики, широкий класс функций может быть представлен в виде

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}\,\delta\omega. \tag{1.12}$$

Каков смысл соотношения (1.12)?

Рассмотрим выражение

$$\frac{1}{2\pi}F(\omega_n)e^{i\omega_n t}\Delta\omega.$$

Оно представляет собой гармоническое колебание частоты ω_n, амплитуда этого колебания равна $\frac{1}{2\pi}|F(\omega_n)|\Delta\omega$, начальная фаза — $\arg F(\omega)$, сумма таких гармонических колебаний

$$\sum_n \frac{1}{2\pi} F(\omega_n) e^{i\omega_n t} \Delta\omega$$

в пределе при $\Delta\omega \to 0$ переходит в интеграл Фурье (1.12). Формула (1.12) утверждает, что можно подобрать амплитуды и фазы гармонических колебаний разных частот (т. е. подобрать комплексную функцию $F(\omega)$ таким образом, что при суммировании (интегрировании) получится любой наперед заданный сигнал $f(t)$). Функция $F(\omega)$ определяет «вес», с которым каждая гармоника входит в суммарный сигнал $f(t)$, и называется *спектром* или *преобразованием Фурье* функции $f(t)$. Спектр $F(\omega)$ может быть отличен от нуля в дискретных точках ω_n, и тогда сигнал $f(t)$ состоит из дискретного набора гармоник $e^{i\omega_n t}$, а интеграл (1.12) превращается в сумму

$$f(t) = \sum c_n e^{i\omega_n t}. \tag{1.13}$$

В общем же случае функцию $f(t)$ можно представить с помощью непрерывного набора гармонических сигналов (с непрерывно меняющейся частотой), причём каждой гармонике $e^{i\omega t}$ соответствуют свои амплитуда и фаза, определяемая комплексной функцией $F(\omega)$. Этот факт и отражается соотношением (1.12). Связь между функцией $f(t)$ и её спектром $F(\omega)$ символически записывают в виде $f(t) \leftrightarrow F(\omega)$.

Условия, налагаемые на функцию $f(t)$, при выполнении которых возможно разложение (1.12) или (1.13), изучаются в курсах математики. Здесь же отметим замечательные математические свойства гармонических функций $e^{i\omega t}$.

Во-первых, гармоническое колебание частоты ω_0 — $e^{i\omega_0 t}$ не может быть представлено суперпозицией гармонических колебаний $\sum c_n e^{i\omega_n t}$ других частот $\omega_n \neq \omega_0$, какие бы коэффициенты $c_n = a_n e^{i\varphi_n}$, т. е. амплитуды и фазы слагаемых гармоник, мы ни старались подобрать. Математически это свойство называется ортогональностью: функция $e^{i\omega_0 t}$ не имеет «проекции» на любую другую функцию $e^{i\omega_n t}$ при $\omega_0 \neq \omega_n$, подобно тому как вектор, параллельный оси z, невозможно представить в виде суммы векторов, параллельных осям x и y.

Второе важнейшее математическое свойство — единственность представлений (1.13) или (1.12): существует единственный набор необходимых частот ω_n и единственный набор отвечающих этим частотам амплитуд a_n и фаз φ_n, обеспечивающий представление

функции $f(t)$ в виде суперпозиции гармонических функций. В случае интегрального представления существует единственная функция $F(\omega)$, обеспечивающая справедливость равенства (1.12).

Наконец, не вдаваясь в математические детали, отметим ещё одно важное обстоятельство: любой физически реализуемый колебательный процесс может быть представлен в виде суммы гармонических колебаний, т. е. либо в виде (1.13), либо в виде (1.12).

§ 1.5. Фильтрация электрических сигналов (спектральный подход)

Конечно, представить в виде линейной суперпозиции гармонических колебаний можно не только сигнал $f(t)$, поступающий на вход фильтра, но и выходной сигнал:

$$g(t) = \frac{1}{2\pi} \int G(\omega) e^{i\omega t}\, d\omega,$$

где $G(\omega)$ — спектр выходного сигнала $g(t)$.

Поставим теперь задачу так: пусть нам известны частотная характеристика фильтра $H(\omega)$ и спектр входного сигнала $F(\omega)$. Требуется определить отклик фильтра и его спектр $G(\omega)$. Решение поставленной таким образом задачи состоит в следующем. Мы знаем, что каждая гармоника входного сигнала $F(\omega_n)e^{i\omega_n t}\Delta\omega$, «возбуждая» фильтр, остаётся гармоникой той же частоты, «превращаясь» в гармонику выходного сигнала $G(\omega_n)e^{i\omega_n t}\Delta\omega$ (следовательно, спектр выходного сигнала на частоте ω зависит только от значения спектра входного сигнала на этой же частоте). При прохождении через фильтр амплитуда и фаза гармонического колебания изменяются: перед каждой гармоникой появляется комплексный множитель $H(\omega_n)$ (формула (1.11)), так что если на входе мы имели колебание

$$\frac{1}{2\pi} F(\omega_n) e^{i\omega_n t} \Delta\omega,$$

то на выходе получаем

$$\frac{1}{2\pi} H(\omega_n) F(\omega_n) e^{i\omega_n t} \Delta\omega.$$

Полный сигнал на выходе есть сумма таких колебаний по всем частотам, т. е.

$$g(t) = \frac{1}{2\pi} \int H(\omega) F(\omega) e^{i\omega t}\, d\omega,$$

и, следовательно, спектр выходного сигнала есть

$$G(\omega) = F(\omega) \cdot H(\omega). \tag{1.14}$$

Получено исключительной важности соотношение, на котором основаны принципы фильтрации в линейных системах. Проиллюстрируем основные положения линейной фильтрации на примере.

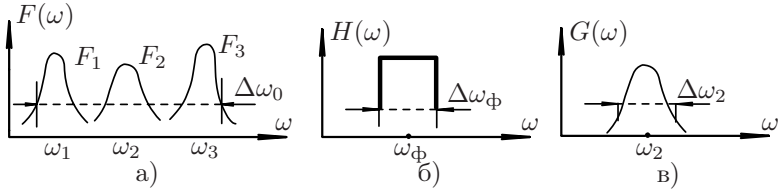


Рис. 1.7

Пусть на вход фильтра поступает сигнал $f(t)$, источником которого являются три различные радиостанции: $f(t)=f_1(t) + f_2(t) + f_3(t)$, ведущие передачи на разных «несущих» частотах ω_1, ω_2 и ω_3 так, что спектры составляющих сигналов $F_1(\omega)$, $F_2(\omega)$ и $F_3(\omega)$ не перекрываются (рис. 1.7а). Суммарный входной сигнал является «широкополосным», т. е. состоит из набора гармоник очень широкого интервала частот $\Delta\omega_0$. Наша задача состоит в том, чтобы с помощью фильтра выделить сигнал интересующей нас станции (например, $f_2(t)$) и отсеять сигналы $f_1(t)$ и $f_3(t)$, посылаемые другими станциями. Осуществим эту задачу с помощью фильтра, частотная характеристика которого изображена на рис. 1.7б. Какие требования должны быть предъявлены к характеристике $H(\omega)$? Эти требования вытекают из соотношения (1.14).

Мы должны решить задачу «селекции», т. е. из широкого спектра входного сигнала выделить интересующий нас сигнал. Для этого необходимо, чтобы фильтр был «настроен» на частоту полезного сигнала, т. е. $\omega_\Ф \simeq \omega_2$. При этом «полоса пропускания» фильтра $\Delta\omega_\Ф$ должна быть достаточно узкой ($\Delta\omega_\Ф < \Delta\omega_0$): выходной сигнал не должен содержать гармоник, принадлежащих сигналам $f_1(t)$ и $f_3(t)$. Наконец, чтобы «полезный» сигнал $f_2(t)$ (интересующая нас передача) был принят без искажений (т. е. $g(t) = f_2(t)$), нужно, чтобы все гармоники, входящие в состав сигнала $f_2(t)$, были пропущены фильтром, т. е. полоса пропускания фильтра должна быть больше полосы частот полезного сигнала ($\Delta\omega_\Ф \gtrsim \Delta\omega_2$). Если эти условия не будут выполнены, то на выходе фильтра появится искажённый сигнал: музыка или речь будут воспроизводиться неверно. На рис. 1.8 и 1.9 показана ситуация, когда

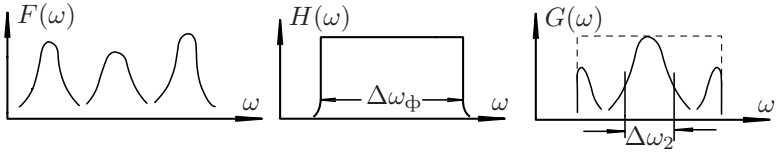


Рис. 1.8. Спектр выходного сигнала содержит лишние гармоники, так как полоса пропускания фильтра слишком велика

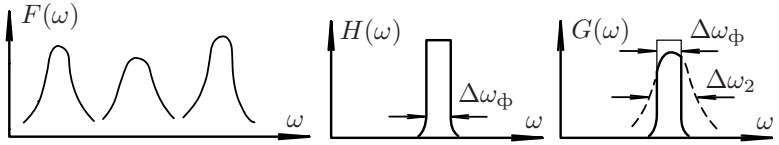


Рис. 1.9. Сигнал искажен, так как полоса пропускания фильтра меньше полосы частот полезного сигнала $\Delta\omega_\Phi < \Delta\omega_2$

ширина полосы фильтра подобрана неверно, в этом случае выходной сигнал или содержит «лишние» гармоники (принадлежащие $f_1(t)$ и $f_3(t)$, рис. 1.8), или в нём не хватает «полезных» гармоник (принадлежащих $f_2(t)$, рис. 1.9).

Приведённый пример поясняет принципы фильтрации в радиотехнике и смысл функции $H(\omega)$. Частотная характеристика даёт возможность не только определить отклик на гармоническое воздействие, с её помощью мы можем определить отклик на произвольный входной сигнал, если известен спектр сигнала на входе. *Равенство* (1.14) *является основой спектрального подхода к задачам линейной фильтрации.*

Одно важное замечание. Обратите внимание на то обстоятельство, что при нахождении частотной характеристики колебательного контура его параметры L, C, R полагались постоянными, не меняющимися во времени. Именно поэтому откликом на гармонический сигнал $e^{i\omega t}$ является также гармоническое колебание той же частоты. Действительно, в противном случае найденная нами функция H (формула (1.10)) зависела бы от времени. При этом выражение для отклика $g(t) = H(\omega,t)e^{i\omega t}$ уже не является гармоническим колебанием. Сказанное относится не только к колебательному контуру, но и к любому линейному фильтру. Очевидно, спектр выходного сигнала не может быть определён в этом случае соотношением (1.14). Фильтры, параметры которых со временем не меняются, называются *стационарными.* Таким образом, *равенство* (1.14) *справедливо для линейных стационарных фильтров.*

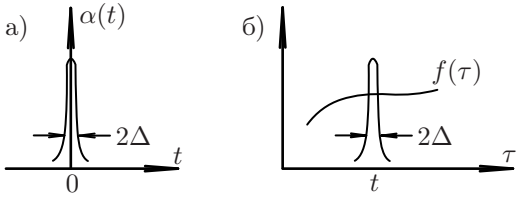


Рис. 1.10

§ 1.6. δ-импульс, временно́й подход к изучению линейных фильтров

Представление произвольного сигнала в виде суперпозиции гармонических колебаний не является единственно возможным. Для описания работы электрических фильтров часто сигнал представляют в виде суперпозиции δ-импульсов. δ-импульсы (или δ-функции Дирака) вообще играют огромную роль в современной физике, поэтому имеет смысл остановиться на них подробнее.

δ-импульс является математической идеализацией очень сильных и коротких реальных импульсов. Такой короткий и сильный сигнал $\alpha(t)$ единичной площади $\int \alpha(t)\,d\tau = 1$ изображён на рис. 1.10а. Импульс $\alpha(t)$ близок к нулю всюду, кроме небольшой окрестности 2Δ точки $t = 0$. Внутри этой окрестности функция $\alpha(t)$ принимает большие значения. Будем уменьшать длительность импульса 2Δ, одновременно увеличивая «силу» импульса так, что $\int \alpha(t)\,d\tau$ остаётся конечной величиной, равной 1. В пределе при $\Delta \to 0$ получаем идеализированный сигнал, называемый δ-импульсом:

$$\delta(t) = \lim_{\Delta\to 0} \alpha(t).$$

Тогда $\delta(t-\tau)$ есть функция, равная нулю всюду, кроме точки $\tau = t$ (в которой эта функция равна бесконечности), причём $\int \delta(t-\tau)\,d\tau = 1$. Рассмотрим выражение

$$\int_{-\infty}^{\infty} f(\tau)\alpha(t-\tau)\,d\tau,$$

где $f(\tau)$ — обычная «хорошая» функция (рис. 1.10б). Во-первых, ясно, что основной вклад в интеграл даёт лишь небольшая окрестность точки $\tau = t$, в которой импульс $\alpha(t-\tau)$ заметно отличен от нуля. Во-вторых, уменьшая длительность импульса, всегда можно сделать

его столь коротким, что внутри окрестности 2Δ функцию $f(\tau)$ можно считать константой, равной $f(t)$. Тогда имеем

$$\int\limits_{-\infty}^{\infty} f(\tau)\alpha(t-\tau)\,d\tau \approx \int\limits_{t-\Delta}^{t+\Delta} f(\tau)\alpha(t-\tau)\,d\tau \approx f(t)\int\limits_{t-\Delta}^{t+\Delta} \alpha(t-\tau)\,d\tau.$$

Поскольку импульс $\alpha(t-\tau)$ содержится целиком в окрестности 2Δ момента $\tau = t$, то пределы интегрирования в последнем выражении можно безболезненно считать бесконечными, значение интеграла при этом не изменится. Следовательно:

$$\int\limits_{-\infty}^{\infty} f(\tau)\alpha(t-\tau)\,d\tau \approx f(t)\int\limits_{-\infty}^{\infty} \alpha(t-\tau)\,d\tau \approx f(t). \tag{1.15}$$

В пределе, когда длительность импульса $\Delta \to 0$, приближённое равенство (1.15) становится точным, и мы получаем

$$\int\limits_{-\infty}^{\infty} f(\tau)\delta(t-\tau)\,d\tau = f(t). \tag{1.16}$$

Последнее соотношение можно считать определением δ-функции. Можно убедиться в том, что в качестве функции $\alpha(t)$, при уменьшении длительности которой происходит переход к δ-импульсу, можно взять, например, функцию $\frac{1}{T}p_T(t)$ (рис. 1.11), определяемую формулой

$$\frac{1}{T}p_T(t) = \begin{cases} \frac{1}{T} & \text{при} \quad |t| \leqslant \frac{T}{2}, \\ 0 & \text{при} \quad |t| > \frac{T}{2} \end{cases}$$

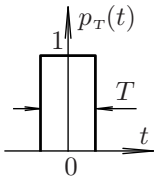


Рис. 1.11

или функцию

$$\frac{\sin at}{\pi t},$$

так что

$$\lim_{T\to 0}\frac{1}{T}p_T(t) = \delta(t), \qquad \lim_{a\to\infty}\frac{\sin at}{\pi t} = \delta(t) \tag{1.17}$$

(первое из этих равенств очевидно, второе будет доказано ниже). В дальнейшем мы воспользуемся замечательным соотношением, связывающим δ-импульс с гармоническими колебаниями:

$$\delta(t) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{i\omega t}\,d\omega. \tag{1.18}$$

Поменяв обозначение переменных, получаем аналогичное тождество:

$$2\pi\delta(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t}\, dt. \tag{1.19}$$

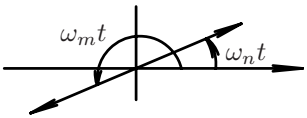


Рис. 1.12

Правая часть равенства (1.18) представляет собой сумму гармонических колебаний одинаковой амплитуды и всевозможных частот ω ($\approx \frac{1}{2\pi}\sum e^{i\omega_n t}\Delta\omega$). Каждое слагаемое можно изобразить в виде вектора длины $\frac{1}{2\pi}\Delta\omega$, составляющего угол $\omega_n t$ с действительной осью (рис. 1.12). При $t = 0$ вся цепочка векторов вытянута вдоль действительной оси (все колебания складываются синфазно), так что

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t}\, d\omega\bigg|_{t=0} = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega = \infty.$$

Для любого $t \neq 0$ в сумме гармонических колебаний для каждой гармоники частоты ω_n всегда найдётся гармоника частоты ω_m, имеющая противоположную фазу: $\omega_n t - \omega_m t = \pi$, так что $e^{i\omega_n t} + e^{i\omega_m t} = 0$. Следовательно,

$$\int_{-\infty}^{\infty} e^{i\omega t}\, d\omega\bigg|_{t\neq 0} = 0.$$

Мы привели качественные соображения, указывающие на то, что формула (1.18) не противоречит здравому смыслу. Сопоставляя (1.18) и (1.12), находим, что спектр δ-импульса есть константа, тождественно равная единице: $\delta(t) \leftrightarrow 1$.

Соотношение (1.16) показывает, что значение сигнала $f(t)$ (в момент времени $t = \tau$) может быть представлено в виде линейной суперпозиции δ-импульсов, действующих в разные моменты времени τ, причём весами-коэффициентами в этой суперпозиции являются значения самой функции $f(\tau)$ в разные моменты времени. Сказанное становится совершенно ясным, если вместо интеграла в левой части (1.16) написать сумму

$$f(t) \approx \sum_n f(t_n)\delta(t - t_n)\Delta t.$$

Представление сигналов в виде суперпозиции δ-импульсов используется при временно́м описании работы линейных фильтров. При

этом свойства фильтра характеризуются его откликом на входной δ-импульс. Итак, пусть $h(t,\tau)$ есть выходной сигнал фильтра (в момент времени t), на вход которого подаётся δ-импульс в момент времени τ, т. е.

$$\hat{L}[\delta(t-\tau)] = h(t,\tau). \tag{1.20}$$

Функция $h(t,\tau)$ называется импульсной реакцией фильтра. Отклик фильтра $g(t)$ на произвольный входной сигнал $f(t)$ можно определить, зная импульсную реакцию $h(t,\tau)$ и используя равенство (1.16). Мы имеем

$$g(t) = \hat{L}[f(t)] = L\left[\int_{-\infty}^{\infty} f(\tau)\delta(t-\tau)\,d\tau\right] = \int_{-\infty}^{\infty} f(\tau)\hat{L}[\delta(t-\tau)]\,d\tau.$$

Последнее равенство является следствием линейности фильтра (см. определение (1.5)). Используя (1.20), получаем окончательно

$$g(t) = \int_{-\infty}^{\infty} f(\tau)h(t,\tau)\,d\tau. \tag{1.21}$$

Если речь идёт о стационарных фильтрах, свойства которых со временем не меняются (например, контур с постоянными L, C, R), то импульсная реакция зависит только от промежутка времени $t-\tau$, прошедшего от начала действия входного δ-импульса до момента времени t, т. е. является функцией одной переменной $t-\tau$:

$$h(t,\tau) = h(t-\tau), \tag{1.22}$$

и для отклика фильтра на входной сигнал $f(t)$ имеем

$$g(t) = \int_{-\infty}^{\infty} f(\tau)h(t-\tau)\,d\tau. \tag{1.23}$$

Принцип причинности. Сейчас мы обсуждаем свойства временны́х (например, электрических) фильтров, в которых входными и выходными сигналами являются функции времени. Такие фильтры должны удовлетворять принципу причинности, суть которого состоит в следующем: выходной сигнал в момент времени t может зависеть от значения входного сигнала только в моменты времени, предшествующие

моменту t, сигнал на выходе не может появиться раньше входного сигнала, поэтому, хотя формально пределы интегрирования в (1.21) и (1.23) бесконечны, в действительности выходной сигнал определяется интервалом времени от $-\infty$ до t. В силу того же принципа причинности импульсная реакция должна быть равна нулю при $t < \tau$, т. е. реакции нет, пока не подействовал входной δ-импульс.

Формула (1.23) даёт решение задачи линейной фильтрации *на временно́м языке*: в этой формуле мы имеем дело не со спектрами, а с самими функциями времени, и свойства фильтра здесь характеризуются не частотной характеристикой, а импульсной реакцией.

Если на линейный фильтр в момент времени $t = 0$ подействовал входной δ-импульс, то при $t > 0$ система оказывается свободной от внешнего воздействия и в ней начинается так называемый «переходно́й процесс» (процесс «свободных» колебаний) — это и есть импульсная реакция фильтра.

Важной характеристикой импульсной реакции является её длительность, которая называется постоянной времени фильтра. Постоянная времени характеризует «инерционность» фильтра, его способность реагировать на быстрые изменения входного сигнала.

Рассмотрим два примера.

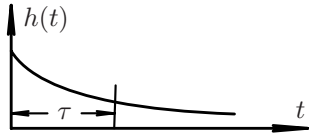


Рис. 1.13

1. Импульсный отклик RC-цепочки. Процесс разряда конденсатора через сопротивление R, начинающийся по завершении действия входного δ-импульса, описывается уравнением

$$\dot{q} + \frac{1}{RC} q = 0;$$

искомая реакция имеет вид

$$h(t) = q_0 e^{-t/RC}$$

(q_0 — начальный заряд, созданный входным δ-импульсом, рис. 1.13). Постоянная времени фильтра $\tau = RC$.

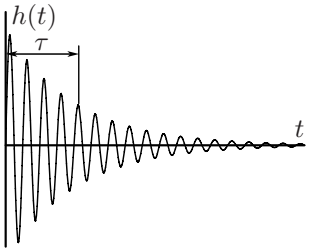


Рис. 1.14

2. Импульсный отклик колебательного контура — это процесс свободных колебаний, возникающих в контуре по окончании действия δ-импульса, он описывается уравнением

$$\ddot{h} + 2\gamma\dot{h} + \omega_0^2 h = 0$$

и имеет вид (рис. 1.14)

$$h(t) = \frac{1}{\omega} e^{-\gamma t} \sin \omega t,$$

где $\omega = \sqrt{\omega_0^2 - \gamma^2}$. Постоянная времени контура есть $\tau \simeq 1/\gamma$. Идеальный, безынерционный фильтр должен иметь импульсный отклик $h(t) = \delta(t)$. Фильтр с таким откликом «пропускает» входной сигнал без искажений: $g(t) = f(t)$, это следует из (1.23) при $h(t) = \delta(t)$. Выходной сигнал оказывается мало искажённым ($g(t) \approx f(t)$) и в том случае, когда постоянная времени фильтра τ существенно меньше характерного времени изменения входного сигнала $\Delta t_{\text{с}}$. Такая ситуация изображена на рис 1.15.

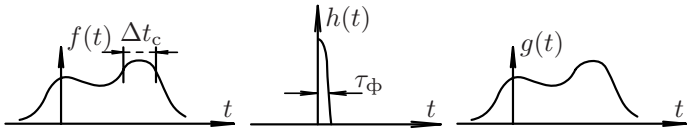


Рис. 1.15

В противном случае (при $\tau_{\text{Ф}} > \Delta t_{\text{с}}$) выходной сигнал заметно отличается по форме от входного сигнала (рис. 1.16); фильтр с большой инерционностью «не успевает» следить за быстрыми изменениями сигнала $f(t)$; согласно (1.23), за время $\tau_{\text{Ф}}$ происходит усреднение и сглаживание быстрых флуктуаций входного сигнала.

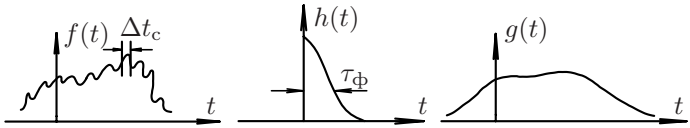


Рис. 1.16

Таким образом, при решении задачи фильтрации на временно́м языке характерными параметрами являются постоянная времени фильтра $\tau_{\text{Ф}}$ и характерное время изменения (или длительность) сигнала $\Delta t_{\text{с}}$. Результат фильтрации существенно зависит от соотношения этих времён.

Итак, с одной стороны, мы имеем дело с сигналами — функциями времени и импульсной реакцией, с другой стороны, — со спектрами сигналов и частотной характеристикой.

Каково соотношение между временны́м (формула (1.23)) и спектральным (формула (1.14)) подходами к исследованию линейных систем? Выясним этот вопрос, найдя связь между частотной характеристикой и импульсной реакцией линейного фильтра.

Воспользуемся соотношением (1.18); δ-импульс, поступающий на вход линейной системы, представим в виде суммы гармонических ко-

лебаний. Имеем, используя (1.18), свойство линейности и (1.11):

$$h(t) = \hat{L}[\delta(t)] = L\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t}\,d\omega\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} L\left[e^{i\omega t}\right]\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)e^{i\omega t}\,d\omega.$$

Таким образом, частотная характеристика $H(\omega)$ является преобразованием Фурье (спектром) импульсной реакции:

$$h(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)e^{i\omega t}\,d\omega. \tag{1.24}$$

Ещё раз обратим внимание на соотношения (1.23) и (1.14). Первое из них связывает функции времени $f(t)$, $h(t)$ и $g(t)$, второе — спектры этих функций $F(\omega)$, $H(\omega)$ и $G(\omega)$, причём связь между спектрами оказывается несравненно проще, чем связь самих функций. Функция $g(t)$, согласно (1.23), находится с помощью операции *свёртки* функций $f(t)$ и $h(t)$ (так называется интегральная операция в (1.23)). Спектр $G(\omega)$ находится как произведение спектров $H(\omega)$ и $F(\omega)$ — операция гораздо менее трудоёмкая. Это одна из причин, по которой спектральный подход к задаче линейной фильтрации оказывается предпочтительным.

Замечание: из изложенного выше следует, что если функция $f(t)$ является свёрткой двух функций $f_1(t)$ и $f_2(t)$:

$$f(t) = \int_{-\infty}^{\infty} f_1(\tau)f_2(t-\tau)\,d\tau,$$

то спектр этой функции $F(\omega)$ является произведением спектров $F_1(\omega)$ и $F_2(\omega)$:

$$F(\omega) = F_1(\omega)\cdot F_2(\omega).$$

Операцию свёртки символически изображают в виде, которым в дальнейшем мы будем пользоваться:

$$f(t) = f_1(t) \otimes f_2(t).$$

Сформулированное утверждение в теории преобразования Фурье называется теоремой о фурье-образе свёртки.

Сформулируем другую теорему, которую читатель легко докажет в качестве упражнения.

Фурье-образ произведения двух функций равен свёртке фурье-образов, т. е. если

$$f(t) = f_1(t) \cdot f_2(t), \qquad \text{то} \qquad F(\omega) = F_1(\omega) \otimes F_2(\omega). \tag{1.25}$$

§ 1.7. Сигналы и их спектры. Некоторые свойства преобразований Фурье. Соотношение неопределённостей. Примеры

Каким образом по заданной функции $f(t)$ находить её спектр, т. е. находить амплитуды и фазы слагаемых гармонических колебаний так, чтобы в сумме они дали заданную функцию? Рассмотрим вначале периодический процесс $f(t) = f(t+T)$. В этом случае сигнал $f(t)$ может быть представлен рядом Фурье — суммой гармонических колебаний с *кратными частотами* $\omega_n = n\omega_0$, где $T = \frac{2\pi}{\omega_0}$ — период процесса

$$f(t) = \sum_n c_n e^{in\omega_0 t}. \tag{1.26}$$

Действительно, для любого t функция (1.26) повторяет своё значение через интервал времени T, поскольку

$$e^{in\omega_0(t+T)} = e^{in\omega_0 T} e^{in\omega_0 t} = e^{i2\pi n} e^{in\omega_0 t} = e^{in\omega_0 t}.$$

Спектр $\{c_n\}$ (набор в общем случае комплексных коэффициентов) можно найти следующим образом. Умножим обе части равенства (1.26) на $e^{-im\omega_0 t}$ и проинтегрируем по t за время, равное периоду (от $-\frac{T}{2}$ до $\frac{T}{2}$). Получаем

$$\int\limits_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-im\omega_0 t}\, dt = \sum_n c_n \int\limits_{-\frac{T}{2}}^{\frac{T}{2}} e^{i(n-m)\omega_0 t}\, dt.$$

Легко проверить, что интеграл в правой части равенства есть

$$\int\limits_{-\frac{T}{2}}^{\frac{T}{2}} e^{i(n-m)\omega_0 t}\, dt = \begin{cases} 0 & \text{при} \quad n \neq m, \\ T & \text{при} \quad n = m. \end{cases}$$

Действительно, представив подынтегральную функцию с помощью формулы Эйлера:

$$e^{i(n-m)\omega_0 t} = \cos(n-m)\omega_0 t + i \sin(n-m)\omega_0 t,$$

заметим, что период функций $\cos(n-m)\omega_0 t$ и $\sin(n-m)\omega_0 t$ есть

$$T_{nm} = \frac{2\pi}{(n-m)\omega_0},$$

поэтому интеграл от cos и sin за время $T = \frac{2\pi}{\omega_0} = (n-m)T_{nm}$, равное целому числу периодов T_{nm}, равен нулю.

Следовательно, получаем

$$c_m = \frac{1}{T} \int\limits_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-im\omega_0 t}\, dt. \tag{1.27}$$

Формула (1.27) даёт правило нахождения коэффициентов разложения периодической функции $f(t)$ в ряд Фурье.

Рассмотрим теперь произвольный сигнал $f(t)$. Его спектр $F(\omega)$ может быть найден следующим образом. Умножим обе части равенства (1.12) на $e^{-i\omega' t}$ и проинтегрируем по t в бесконечных пределах. Мы имеем (меняя в правой части порядок интегрирования)

$$\int\limits_{-\infty}^{\infty} f(t) e^{-i\omega' t}\, dt = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} F(\omega) \left[\int\limits_{-\infty}^{\infty} e^{-i(\omega'-\omega)t}\, dt \right] d\omega.$$

Внутренний интеграл, согласно (1.19), есть $2\pi\delta(\omega'-\omega)$, и, следовательно, по определению δ-функции (1.16) правая часть последнего равенства есть $F(\omega')$. Обозначая затем ω' через ω, приходим к формуле

$$F(\omega) = \int\limits_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt. \tag{1.28}$$

Соотношение (1.28) принято называть *прямым преобразованием Фурье*, а формулу (1.12) — *обратным преобразованием Фурье*.

Обратим внимание на симметрию выражений, определяющих прямое и обратное преобразование Фурье. Следствием этой симметрии является приведённый ниже результат двух последовательных преобразований Фурье, применённых в функции $f(t)$.

Первое преобразование Фурье даёт (формула (1.28))

$$F_1(\omega) = \int\limits_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt.$$

Применим к полученной функции $F_1(\omega)$ ещё одно преобразование Фурье. Получаем

$$F_2(\omega) = \int_{-\infty}^{\infty} F_1(\omega') e^{-i\omega'\omega}\, d\omega'.$$

Далее, используя выражение для $F_1(\omega)$, получаем

$$F_2(\omega) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(t) e^{-i\omega' t}\, dt \right] e^{-i\omega'\omega}\, d\omega'.$$

Меняя порядок интегрирования, находим

$$F_2(\omega) = \int_{-\infty}^{\infty} f(t) \left[\int_{-\infty}^{\infty} e^{-i(\omega+t)\omega'}\, d\omega' \right] dt.$$

Внутренний интеграл (по переменной ω') равен $2\pi\delta(\omega + t)$. Следовательно:

$$F_2(\omega) = 2\pi \int_{-\infty}^{\infty} f(t)\delta(\omega + t)\, dt = 2\pi f(-\omega).$$

Последнее следует из определения δ-функции (1.16). Итак, два последовательных преобразования Фурье, применённых к функции $f(t)$, возвращают нас с точностью до множителя 2π и инверсии (изменения знака аргумента) к исходной функции. Символически полученный результат означает, что если $f(t) \leftrightarrow F(\omega)$, то

$$F(t) \leftrightarrow 2\pi f(-\omega). \tag{1.29}$$

Общее соотношение (1.28) можно использовать для нахождения спектра произвольного сигнала, в том числе и в случае, когда сигнал $f(t)$ является периодическим, с периодом T, т. е. представляется суммой (1.26). Мы получаем

$$F(\omega) = \int_{-\infty}^{\infty} \sum_n c_n e^{in\omega_0 t} e^{-i\omega t}\, dt = \sum_n c_n \int_{-\infty}^{\infty} e^{-i(\omega - n\omega_0)t}\, dt.$$

Интеграл в последнем выражении равен согласно (1.19)

$$2\pi\delta(\omega - n\omega_0),$$

поэтому находим

$$F(\omega) = 2\pi \sum_n c_n \delta(\omega - n\omega_0),$$

т. е. спектр периодического сигнала представляет собой «гребёнку» равноотстоящих по частоте δ-функций с весовым множителем c_n, определяемым с помощью (1.27).

Сравнивая (1.27) и (1.28), легко получить

$$c_n = \frac{1}{T} \int\limits_{-\frac{T}{2}}^{\frac{T}{2}} f_0(t) e^{-in\omega_0 t}\, dt = \frac{1}{T} F_0(n\omega_0). \tag{1.30}$$

(Напомним, что речь идёт о сигнале $f(t)$, являющимся периодическим с периодом T повторением сигнала $f_0(t)$, причём на интервале $[-\frac{T}{2}, \frac{T}{2}]$ функция $f(t)$ совпадает с $f_0(t)$.)

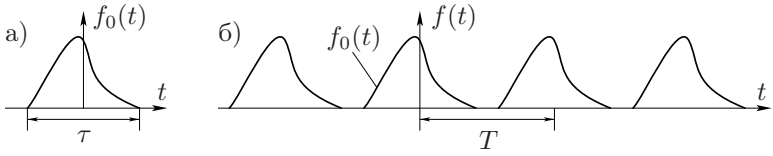


Рис. 1.17

Пусть сигнал

$$f(t) = \sum_n f_0(t - nT)$$

представляет собой периодическое повторение одиночного импульса $f_0(t)$, причём отдельные слагаемые сигнала $f(t)$ не перекрываются (рис. 1.17), т. е $T > \tau$. Тогда, используя последнее выражение для c_n, получаем полезное выражение для спектра периодического сигнала:

$$F(\omega) = \frac{2\pi}{T} \sum_n F_0(n\omega_0)\delta(\omega - n\omega_0) = \frac{2\pi}{T} F_0(\omega) \sum_n \delta(\omega - n\omega_0). \tag{1.31}$$

Таким образом, для нахождения спектра периодического сигнала

$$f(t) = \sum_n f_0(t - nT)$$

достаточно знать значения спектра одиночного «импульса» $f_0(t)$ в «отсчётных» точках $\omega_n = n\omega_0$. Рис. 1.18 поясняет полученный результат.

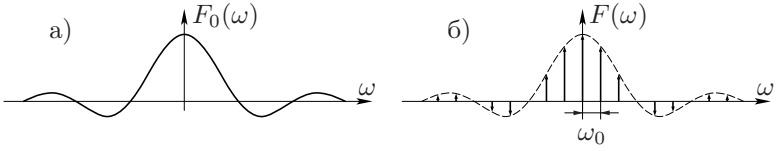


Рис. 1.18

Вместо непрерывного спектра одиночного импульса $F_0(\omega)$ (рис. 1.18а) получаем дискретный спектр $F(\omega)$ периодической последовательности одинаковых импульсов. Огибающей гребёнки δ-функций (показана пунктиром на рис. 1.18б) является спектр одиночного импульса — функция $\frac{1}{T}F_0(\omega)$.

Теорема сдвига. *Пусть сигнал $f_0(t)$ имеет спектр $F_0(\omega)$:*

$$f_0(t) \leftrightarrow F_0(\omega).$$

Найдём спектр сигнала $f_0(t-T)$, смещённого по времени (запаздывание) на T. Получаем

$$F(\omega) = \int_{-\infty}^{\infty} f_0(t-T)e^{-i\omega t}\,dt.$$

После замены переменной $t' = t - T$ имеем

$$F(\omega) = \int_{-\infty}^{\infty} f_0(t')e^{-i\omega(t'+T)}\,dt' = e^{-i\omega T}\int_{-\infty}^{\infty} f_0(t')e^{-i\omega t'}\,dt'.$$

Таким образом,

$$F(\omega) = F_0(\omega)e^{-i\omega T},$$

или символически

$$f_0(t-T) \leftrightarrow F_0(\omega)e^{-i\omega T}.$$

Сдвиг сигнала во времени (на T) приводит к линейному по частоте изменению фазы спектральных компонент сигнала — появлению множителя $e^{-i\omega T}$.

Используя теорему сдвига, можно получить следующее выражение для спектра периодической последовательности одинаковых импульсов. Если спектр одиночного импульса $f_0(t)$ есть $F_0(\omega)$, т. е. $f_0(t) \leftrightarrow F_0(\omega)$, то спектр сигнала

$$f(t) = \sum_n f_0(t-nT)$$

есть

$$F(\omega) = F_0(\omega) \sum_n e^{-in\omega T}. \tag{1.32}$$

Сравнивая (1.31) и (1.32), получаем полезное тождество:

$$\sum_n e^{-in\omega T} = \frac{2\pi}{T} \sum_n \delta(\omega - n\omega_0), \qquad \left(\omega_0 = \frac{2\pi}{T}\right). \tag{1.33}$$

Проиллюстрируем полученные выше общие соотношения рядом характерных примеров.

1. Спектр прямоугольного импульса

$$f_0(t) = p_\tau(t) = \begin{cases} 1 & \text{при} \quad |t| \leqslant \frac{\tau}{2}, \\ 0 & \text{при} \quad |t| > \frac{\tau}{2}. \end{cases}$$

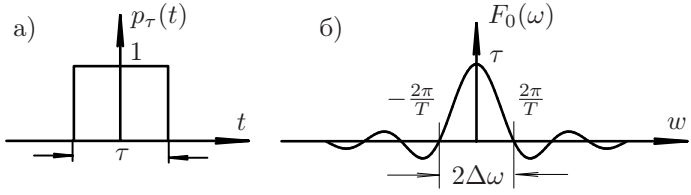


Рис. 1.19

Используя (1.28), находим

$$F_0(\omega) = \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} e^{-i\omega t}\, dt = 2\left(\frac{\sin\omega\frac{\tau}{2}}{\omega}\right) = \tau\left(\frac{\sin\omega\frac{\tau}{2}}{\omega\frac{\tau}{2}}\right). \tag{1.34}$$

Сигнал $p_\tau(t)$ и его спектр $F_0(\omega)$ изображены на рис. 1.19а, б.

«Ширину» спектра, т. е. интервал частот, в котором функция $F_0(\omega)$ заметно отличается от нуля, можно оценить по полуширине её *главного максимума*: $\Delta\omega = \frac{2\pi}{\tau}$. Длительность сигнала Δt есть $\Delta t = \tau$. Отметим, что $\Delta\omega \cdot \Delta t = 2\pi$: с увеличением длительности сигнала Δt ширина спектра $\Delta\omega$ уменьшается. Наоборот, уменьшение длительности импульса приводит к уширению его спектра.

Воспользовавшись равенствами (1.19) и (1.34), можно записать

$$\delta(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega t}\, dt = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} e^{-i\omega t}\, dt = \lim_{T\to\infty} \frac{\sin\omega T}{\pi\omega}. \tag{1.35}$$

Мы пришли ко второй формуле (1.17). Из (1.35) следует, что спектр функции $f(t) = 1$ есть $2\pi\delta(\omega)$: $1 \leftrightarrow 2\pi\delta(\omega)$.

2. Спектр $\{c_n\}$ периодической (с периодом T) последовательности прямоугольных импульсов длительности τ (рис. 1.20) найдём, используя общее соотношение (1.30), а также выражение (1.34) для спектра $F_0(\omega)$ одиночного прямоугольного импульса. Получаем

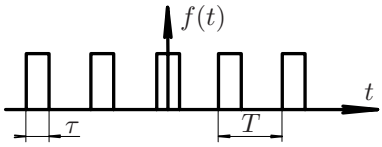


Рис. 1.20

$$c_n = \frac{\tau}{T}\left(\frac{\sin n\omega_0 \frac{\tau}{2}}{n\omega_0 \frac{\tau}{2}}\right). \tag{1.36}$$

Спектр $\{c_n\}$ показан на рис. 1.18б: «огибающей» гребёнки δ-функций является функция (1.34) — спектр одиночного импульса (с множителем $\frac{1}{T}$).

3. Пусть спектр сигнала $f_0(t)$ есть $F_0(\omega)$: $f_0(t) \leftrightarrow F_0(\omega)$. Найдём спектр сигнала $f(t) = f_0(t)e^{i\omega_0 t}$. Согласно (1.28), находим

$$F(\omega) = \int\limits_{-\infty}^{\infty} f_0(t)e^{i\omega_0 t}\cdot e^{-i\omega t}\,dt = \int\limits_{-\infty}^{\infty} f_0(t)e^{-i(\omega-\omega_0)t}\,dt = F_0(\omega-\omega_0). \tag{1.37}$$

Таким образом, $f_0(t)e^{i\omega_0 t} \leftrightarrow F_0(\omega-\omega_0)$: умножение сигнала $f_0(t)$ на «несущее» колебание $e^{i\omega_0 t}$ (модуляция) приводит к сдвигу спектра $F_0(\omega)$ по оси частот на величину ω_0 — частоту несущего колебания.

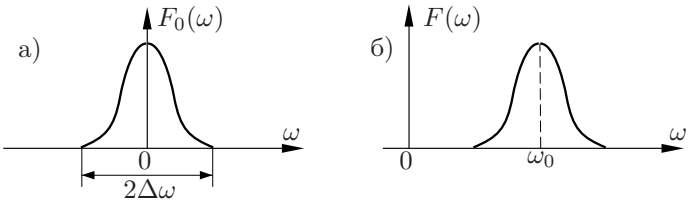


Рис. 1.21

Заметим, что если спектр сигнала $f_0(t)$ локализован в области частот $\Delta\omega$, т. е. $F_0(\omega) \equiv 0$ при $|\omega| > \Delta\omega$, то модулированный сигнал $f_0(t)e^{i\omega_0 t}$ не содержит отрицательных частот, если $\omega_0 > \Delta\omega$, т. е. $F(\omega) \equiv 0$ при $\omega < 0$ (рис. 1.21).

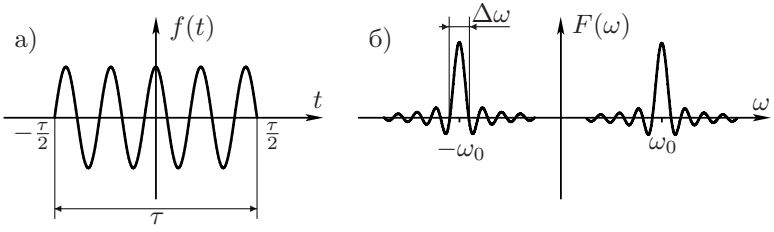


Рис. 1.22

4. Спектр обрывка косинусоиды (цуга) длительности τ (рис. 1.22):

$$f(t) = p_\tau(t)\cos\omega_0 t = \begin{cases} \cos\omega_0 t & \text{при} \quad |t| \leqslant \frac{\tau}{2}, \\ 0 & \text{при} \quad |t| > \frac{\tau}{2}. \end{cases}$$

Записав сигнал $f(t)$ в виде

$$f(t) = \frac{1}{2}p_\tau(t)e^{i\omega_0 t} + \frac{1}{2}p_\tau(t)e^{-i\omega_0 t},$$

получаем, используя (1.37):

$$F(\omega) = \frac{1}{2}F_0(\omega - \omega_0) + \frac{1}{2}F_0(\omega + \omega_0), \tag{1.38}$$

где $F_0(\omega)$ — спектр прямоугольного импульса длительности τ.

Итак, спектр цуга получается сдвигом спектра прямоугольного импульса по оси частот влево и вправо на ω_0 — частоту «несущего» колебания (рис. 1.22):

$$F(\omega) = \frac{\pi}{2}\left[\frac{\sin(\omega - \omega_0)\frac{\tau}{2}}{(\omega - \omega_0)\frac{\tau}{2}}\right] + \frac{\pi}{2}\left[\frac{\sin(\omega + \omega_0)\frac{\tau}{2}}{(\omega + \omega_0)\frac{\tau}{2}}\right]. \tag{1.39}$$

Как и в примере 1, частотный интервал гармоник $\Delta\omega \approx \frac{2\pi}{\tau}$ растёт при уменьшении длительности цуга τ и уменьшается при увеличении его продолжительности. В пределе при $\tau \to \infty$ сигнал становится бесконечной косинусоидой, а его спектр находим из (1.39) при $T \to \infty$:

$$F(\omega) = \pi\delta(\omega - \omega_0) + \pi\delta(\omega + \omega_0).$$

5. Импульсный отклик RC-цепочки:

$$f(t) = \begin{cases} e^{-t/\tau} & \text{при } t \geqslant 0, \\ 0 & \text{при } t < 0, \end{cases} \tag{1.40}$$

где $\tau = RC$ — постоянная времени цепочки (длительность отклика),

$$F(\omega) = \frac{1}{1/\tau + i\omega}. \tag{1.41}$$

Читатель может убедиться, что функция (1.41) (спектр импульсной реакции) совпадает с частотной характеристикой RC-цепочки, как и должно быть согласно общему соотношению (1.24).

На рис. 1.23 изображены сигнал (1.40) и квадрат модуля спектра (1.41).

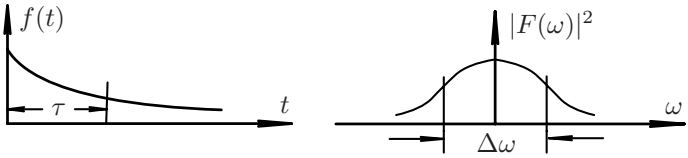


Рис. 1.23

Ширину спектра $\Delta\omega$ оценим из условия спадания функции $|F(\omega)|^2$ вдвое (т. е. из равенства $|F(\omega)|^2 = \frac{1}{2}|F(0)|^2$), откуда находим $\Delta\omega = 2/\tau$.

Снова можем убедиться, что увеличение постоянной времени τ приводит к уменьшению ширины спектра $\Delta\omega$.

6. Свободные колебания осциллятора с затуханием (импульсная реакция):

$$f(t) = \begin{cases} e^{-\gamma t}\sin\omega_0 t & \text{при } t \geqslant 0, \\ 0 & \text{при } t < 0. \end{cases} \tag{1.42}$$

Для определения спектра $F(\omega)$ введём функцию

$$F_1(\omega) = \int\limits_0^\infty e^{-\gamma t}\cos\omega_0 t\, e^{-i\omega t}\, dt$$

и образуем равенства

$$iF(\omega) + F_1(\omega) = \frac{1}{\gamma + i(\omega - \omega_0)},$$

$$iF(\omega) - F_1(\omega) = \frac{1}{\gamma + i(\omega + \omega_0)}.$$

Решая написанную систему уравнений относительно $F(\omega)$ и $F_1(\omega)$, находим

$$F(\omega) = \frac{1}{(\omega_0^2 - \omega^2) + 2i\gamma\omega}. \tag{1.43}$$

Полученное выражение, очевидно, совпадает с частотной характеристикой затухающего осциллятора (1.10). Длительность сигнала (1.42) и ширина его спектра определяются равенствами

$$\Delta t \simeq \frac{1}{\gamma}; \qquad \Delta\omega \simeq \gamma.$$

Процесс затухающих колебаний (1.42) и квадрат модуля спектра (1.10) изображены на рис. 1.14 и 1.6.

Соотношение неопределённостей. Рассматривая различные пары преобразований Фурье (примеры 1–6), мы видим, что длительность сигнала Δt и ширина его спектра $\Delta\omega$ связаны общим соотношением:

$$\Delta t \cdot \Delta\omega \simeq 2\pi. \tag{1.44}$$

Это соотношение является фундаментальным свойством преобразования Фурье и называется *соотношением неопределённостей*.

Согласно (1.44) уменьшение длительности сигнала Δt приводит к увеличению ширины спектра $\Delta\omega$ — частотному интервалу гармонических колебаний, суперпозиция которых образует данный сигнал.

Наоборот, с ростом длительности сигнала спектр сужается, т. е. уменьшается интервал частот $\Delta\omega$, вносящих заметный вклад в суммарный сигнал, — в этом смысл соотношения неопределённостей.

Часто сигналы $f(t)$ характеризуются различными постоянными времени: в примерах 6 и 4 одно характерное время — общая длительность сигнала Δt (τ в примере 4 и $1/\gamma$ в примере 6). Другое характерное время — это *минимальное* время $t_{\min}$, в течение которого происходит заметное изменение величины сигнала (в примерах 4 и 6 это время есть π/ω_0). Эти два временных масштаба Δt и $t_{\min}$ могут заметно различаться по величине: $\Delta t \gg t_{\min}$ (если «цуг» косинусоиды в примере 4 содержит большое число периодов или за время $\Delta t = 1/\gamma$ происходит большое число колебаний осциллятора — малое затухание в примере 6). При этом частота ω_0 — это максимальная частота гармоник, участвующих в образовании сигнала $f(t)$: $\omega_{\max} \simeq \omega_0$. Как видно из приведённых примеров, $t_{\min}$ и $\omega_{\max}$ связаны соотношением $t_{\min} \cdot \omega_{\max} \simeq 2\pi$, которое является другой формой соотношения неопределённостей.

§ 1.8. Теорема Котельникова (теорема отсчётов)

Для целей обработки информации и анализа часто удобно задавать функцию набором её выборочных значений, взятых в дискретной совокупности точек (так называемых отсчётных точках). Ясно,

что если выборочные значения взяты достаточно близко друг к другу, то функция достаточно точно аппроксимируется путём интерполирования по этим значениям. Но оказывается, что для определённого класса функций возможно *точное восстановление* функции по её выборочным значениям, если интервал между отсчётными точками не превышает определённого предельного значения. Это класс функций с *финитным спектром*.

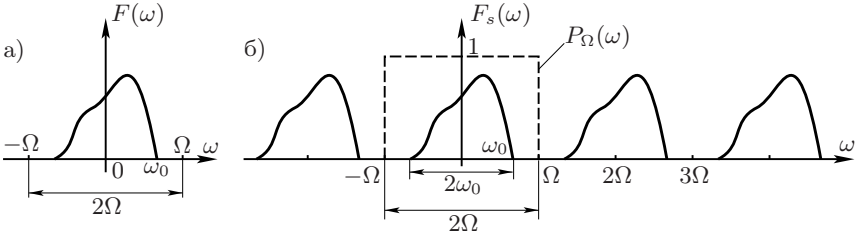


Рис. 1.24

Пусть функция $f(t)$ имеет спектр $F(\omega)$: $f(t) \leftrightarrow F(\omega)$, удовлетворяющий условию: $F(\omega) \equiv 0$ при $|\omega| \geqslant \Omega$ (функция с финитным спектром (рис. 1.24а)). Тогда имеет место следующее представление:

$$f(t) = \sum_{n=-\infty}^{\infty} f\left(n\frac{\pi}{\Omega}\right) \frac{\sin\Omega\left(t - n\frac{\pi}{\Omega}\right)}{\Omega\left(t - n\frac{\pi}{\Omega}\right)}, \tag{1.45}$$

где $\Omega \geqslant |\omega_{\max}| = \omega_0$, $\omega_{\max}$ — максимальное значение частоты, при которой спектр $F(\omega)$ отличен от нуля. Формула (1.45) представляет собой содержание теоремы отсчётов: функция $f(t)$ восстанавливается с помощью равенства (1.45) по своим значениям в отсчётных точках $t_n = n\frac{\pi}{\Omega}$, лишь бы интервал между отсчётами не превышал величины $\frac{\pi}{\omega_{\max}} = \frac{\pi}{\omega_0}$.

Рассмотрим функцию $f_s(t)$, спектр которой $F_s(\omega)$ представляет собой периодическое повторение финитного спектра $F(\omega)$ (с частотным интервалом 2Ω) так, что отдельные части спектра $F_s(\omega)$ не перекрываются (рис. 1.24б), т. е.

$$F_s(\omega) = 2\pi \sum_n F(\omega - n \cdot 2\Omega). \tag{1.46}$$

Собственно функция $f_s(t)$ — обратное преобразование Фурье функции $F_s(\omega)$ — имеет вид

$$f_s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_s(\omega)e^{i\omega t}\,d\omega = \sum_n \int_{-\infty}^{\infty} F(\omega - n \cdot 2\Omega)e^{i\omega t}\,d\omega.$$

Используя (1.37): $f(t)e^{in2\Omega t} \leftrightarrow F(\omega - n \cdot 2\Omega)$, получаем

$$f_s(t) = f(t) \sum_n e^{in2\Omega t}. \tag{1.47}$$

С помощью тождества (1.33) преобразуем (1.47) к виду

$$f_s(t) = \frac{\pi}{\Omega} f(t) \sum_n \delta\left(t - n\frac{\pi}{\Omega}\right) = \frac{\pi}{\Omega} \sum_n f\left(n\frac{\pi}{\Omega}\right) \delta\left(t - n\frac{\pi}{\Omega}\right). \tag{1.48}$$

Финитный спектр $F(\omega)$ можно получить из $F_s(\omega)$ с помощью очевидного равенства $F(\omega) = F_s(\omega)P_\Omega(\omega)$ ($F(\omega)$ представляет собой спектр сигнала на выходе фильтра с частотной характеристикой $H(\omega) = P_\Omega(\omega)$, на вход которого подаётся сигнал $f_s(t) \leftrightarrow F_s(\omega)$):

$$f_s(t) \to \boxed{H(\omega)} \to f(t),$$

где $P_\Omega(\omega)$ — единично-нулевая функция, показанная на рис. 1.24б пунктиром:

$$P_\Omega = \begin{cases} 1 & \text{при} \quad |\omega| \leqslant \Omega, \\ 0 & \text{при} \quad |\omega| > \Omega. \end{cases}$$

Согласно теореме свёртки, операции произведения спектров соответствует операция свёртки самих функций времени, поэтому получаем

$$f(t) = f_s(t) \otimes \frac{\sin \Omega t}{\pi t} \qquad \left(\text{поскольку} \quad \frac{\sin \Omega t}{\pi t} \leftrightarrow P_\Omega(\omega)\right).$$

Используя (1.48), находим

$$f(t) = \frac{\pi}{\Omega} \sum_n f\left(n\frac{\pi}{\Omega}\right) \int_{-\infty}^{\infty} \frac{\sin \Omega t'}{\pi t'} \delta\left(t - t' - n\frac{\pi}{\Omega}\right) dt',$$

откуда следует искомое равенство (1.45).

Ещё раз подчеркнём, что имеется произвол в выборе расстояния между отсчётными точками $\frac{\pi}{\Omega}$, которое лишь должно быть не больше $\frac{\pi}{\omega_0}$, где ω_0 — максимальная частота гармоники в спектре сигнала $f(t)$ (рис. 1.24а). Отметим замечательное свойство функций отсчётов $\frac{\sin \Omega\left(t - n\frac{\pi}{\Omega}\right)}{\Omega\left(t - n\frac{\pi}{\Omega}\right)}$. В каждой отсчётной точке $t_m = m\frac{\pi}{\Omega}$ «включается», т. е. отлична от нуля (и равна единице), одна из функций отсчётов в сумме (1.45), а именно слагаемое с $n = m$, все прочие функции отсчётов с $n \neq m$ обращаются в точке $t = t_m$ в нуль. В любой промежуточной

точке $t \neq m\frac{\pi}{\Omega}$ вклад в значение $f(t)$ дают все слагаемые в сумме, однако наибольший вклад обеспечивается соседними с точкой t отсчётными значениями.

Заключение. Исследуя в этой главе вопросы возбуждения колебаний в электрических системах, мы пользовались такими понятиями, как линейный фильтр, преобразование Фурье, частотная характеристика и импульсный отклик, полоса пропускания и постоянная времени фильтра и т. д. Эти понятия заимствованы из теории колебаний линейных систем. Изложение вопросов оптики в последующих главах будет построено таким образом, чтобы труд, потраченный читателем на чтение первой главы, не пропал даром. Введённые выше понятия будут использованы так, чтобы оптические явления могли быть изучены на основе естественного обобщения и развития принципов линейной фильтрации электрических сигналов. Фильтры, о которых пойдёт речь ниже, — это *пространственные фильтры*, в которых входными и выходными сигналами являются пространственные распределения оптических полей — функции координат. Само устройство фильтров не имеет ничего общего с устройством электрических фильтров (быть может, эти фильтры вообще ни из чего не устроены, состоят из «ничего»); тем не менее перенос понятий, возникающих в задачах электрической фильтрации, на рассматриваемые ниже оптические явления оказывается чрезвычайно полезным, а выявляющиеся при этом оптико-электрические аналогии делают изучение оптики более интересным занятием.

Глава II

Пространственная фильтрация и пространственный спектр

§ 2.1. Волны и волновое уравнение

Мы уже говорили о том, что настоящая книга посвящена изучению волновых оптических явлений. Прежде всего следует сказать о том, что мы называем волновым процессом, в каких случаях можно говорить о волновом характере какой-либо физической величины $E(\boldsymbol{r},t)$, являющейся функцией координат $\boldsymbol{r}(x,y,z)$ и времени t. Рассмотрим наиболее простой пример. Пусть созданное в некоторой точке пространства $\boldsymbol{r}_1(x_1,y_1,z_1)$ в момент времени t_1 возмущение $E(\boldsymbol{r},t)$ достигло точки $\boldsymbol{r}_2 = \boldsymbol{r}_1 + \Delta\boldsymbol{r}$ в момент времени $t_2 = t_1 + \Delta t$, т. е. $E(\boldsymbol{r}_1,t_1) = E(\boldsymbol{r}_2,t_2)$. Ясно, что зависимость функции E от аргументов $\boldsymbol{r}$ и t должна при этом иметь вид

$$E(\boldsymbol{r},t) = E(\boldsymbol{r} - \boldsymbol{v}t), \tag{2.1}$$

где $\boldsymbol{v} = \frac{\Delta\boldsymbol{r}}{\Delta t}$ — скорость распространения возмущения. Действительно, поскольку $\boldsymbol{r}_1 - \boldsymbol{v}t_1 = \boldsymbol{r}_2 - \boldsymbol{v}t_2$, то $E(\boldsymbol{r}_1,t_1) = E(\boldsymbol{r}_2,t_2)$. Таким образом, изменяющееся во времени поле $E(\boldsymbol{r},t)$ имеет характер волны — возмущения, распространяющегося с конечной скоростью $\boldsymbol{v}$ от одной точки пространства до другой.

Любая функция вида (2.1) описывает волновой процесс. Можно убедиться непосредственной подстановкой (упражнение для читателя), что любая функция вида (2.1) (т. е. любая функция от аргумента $\boldsymbol{r} - \boldsymbol{v}t$) удовлетворяет волновому уравнению

$$\nabla^2 E - \frac{1}{v^2}\frac{\partial^2 E}{\partial t^2} = 0. \tag{2.2}$$

Это уравнение описывает волны самой различной природы. Например, если речь идёт о звуковой волне, то $E(\boldsymbol{r},t)$ — давление в газе или

плотность в какой-либо упругой среде. Если изучаются волны на поверхности жидкости или волновые движения колеблющейся струны, то E имеет смысл координаты — отклонения поверхности жидкости или отклонения струны от положения равновесия. Возникает вопрос: носит ли характер волнового процесса также изменяющиеся во времени электрическое и магнитное поля? Оказывается, и это является важнейшим следствием законов электромагнетизма (уравнений Максвелла), что изменяющиеся во времени электрические и магнитные поля также являются волнами и удовлетворяют волновому уравнению.

Большое число экспериментальных фактов свидетельствует о том, что свет представляет собой электромагнитную волну, и, следовательно, уравнение (2.2) среди прочих волновых явлений описывает и законы распространения света.

Для явлений, о которых идёт речь ниже, несущественен векторный характер электромагнитных волн, поэтому мы будем интересоваться скалярной функцией $E(\boldsymbol{r},t)$, понимая под ней любую компоненту напряжённости электрического поля в световой волне. В дальнейшем для краткости функцию $E(\boldsymbol{r},t)$ будем называть *волной*. Если свет распространяется в диэлектрике с диэлектрической проницаемостью ε, то $v = c/\sqrt{\varepsilon}$, где $c \approx 3{\cdot}10^{10}$ см/с — скорость света в пустоте.

Какие самые общие заключения о характере распространения света могут быть сделаны на основании (2.2)? Наиболее важны для нас два обстоятельства. Первое — существует единственное решение $E(\boldsymbol{r},t)$ волнового уравнения (2.2), удовлетворяющее условию $E(\boldsymbol{r},t)\big|_{z=0} = \mathcal{E}_0(x,y,t)$. Это утверждение нужно понимать следующим образом: если в плоскости $z = 0$ задано изменяющееся во времени поле $\mathcal{E}_0(x,y,t)$, то волна в области $z > 0$, бегущая от границы вправо, определяется этим краевым полем единственным образом.

Второй момент — уравнение (2.2) линейно; отметим, линейно до тех пор, пока диэлектрическая проницаемость ε (и, следовательно, $v = c/\sqrt{\varepsilon}$) не зависит от величины поля. Как только начинает проявляться зависимость $\varepsilon = \varepsilon(E)$, линейность уравнения (2.2) нарушается. Надо сказать, что нелинейные эффекты в диэлектрике наблюдаются при фокусировке мощных лазерных пучков, когда напряжённость поля в волне становится сравнимой с внутриатомными полями (остаётся справедливым общее правило — линейность нарушается при достаточно сильных внешних воздействиях). Мы не будем касаться нелинейных явлений, полагая в дальнейшем, что диэлектрическая проницаемость не зависит от напряжённости поля в волне. Из линейности уравнения (2.2) следует, что если в области $z > 0$ имеем две волны $E_1(\boldsymbol{r},t)$ и

$E_2(\boldsymbol{r},t)$, удовлетворяющие на плоскости $z = 0$ условиям

$$E_1(\boldsymbol{r},t)\Big|_{z=0} = \mathcal{E}_1(x,y,t), \qquad E_2(\boldsymbol{r},t)\Big|_{z=0} = \mathcal{E}_2(x,y,t),$$

то решение волнового уравнения, удовлетворяющее на плоскости $z = 0$ условию

$$E(\boldsymbol{r},t)\Big|_{z=0} = c_1\mathcal{E}_1(x,y,t) + c_2\mathcal{E}_2(x,y,t),$$

есть

$$E(\boldsymbol{r},t) = c_1E_1(\boldsymbol{r},t) + c_2E_2(\boldsymbol{r},t). \tag{2.3}$$

Равенство (2.3) выражает принцип суперпозиции волн в линейной среде.

Одна из важнейших задач, решение которой нам предстоит найти, состоит в следующем: в плоскости $z = 0$ задано волновое поле $\mathcal{E}(x,y,t)$; это поле может иметь, вообще говоря, сложную конфигурацию — сложный характер зависимости от координат x, y. Требуется найти волну (решение волнового уравнения) $E(\boldsymbol{r},t)$ в области $z > 0$; в частности, ответить на вопрос: какое поле создаётся этой волной в некоторой плоскости $z = \text{const} > 0$?

Используя свойство линейности волнового уравнения и его следствие — принцип суперпозиции (2.3), — можно предложить следующую схему решения сформулированной выше задачи. Представим заданное волновое поле $\mathcal{E}(x,y,t)$ в плоскости $z = 0$ (входной сигнал линейного фильтра!) в виде линейной суперпозиции некоторых более простых волновых полей и найдём решения волнового уравнения, соответствующие каждому слагаемому в этой суперпозиции. Искомое решение (выходной сигнал фильтра!) находится как линейная суперпозиция решений, соответствующих каждому слагаемому «входного сигнала». Предложенный алгоритм решения полностью аналогичен схеме, по которой мы находим отклик линейного фильтра (такого, как электрический колебательный контур или механический маятник) на произвольное внешнее воздействие (внешнюю ЭДС в контуре или внешнюю силу в механической системе). Фильтр, о котором идёт речь в рассматриваемой волновой задаче, — это пространственный фильтр, причём входной и выходной сигналы — это функции координат: заданное волновое поле в плоскости $z = 0$ и искомое поле в плоскости $z = \text{const} > 0$. Поражает то, что, несмотря на разительную внешнюю несхожесть, рассматриваемая волновая задача может быть сформулирована и решена так же как задача возбуждения колебаний в механических или электрических системах. Так же, как в задаче фильтрации колебаний, необходимо «правильно» выбрать базис — те элементарные

волновые поля в плоскости $z = 0$, для которых решение волнового уравнения найти проще всего и суперпозиция которых даёт заданное «входное» поле. Решение этого вопроса мы отложим до § 2.5.

§ 2.2. Гармонические волны, комплексная амплитуда, уравнение Гельмгольца

Особый интерес представляют гармонические (или монохроматические) волны:

$$E(\boldsymbol{r},t) = a(\boldsymbol{r})\cos[\omega t - \varphi(\boldsymbol{r})]. \tag{2.4}$$

В фиксированной точке наблюдения волна (2.4) создаёт гармонические колебания с частотой ω, амплитудой $a(\boldsymbol{r})$ и начальной фазой $\varphi(\boldsymbol{r})$. Необходимость изучения волн вида (2.4) определяется тем, что любую немонохроматическую волну, так же, как и произвольный колебательный процесс, можно представить, согласно § 1.4, в виде суперпозиции монохроматических волн различных частот ω.

В главе I мы пользовались комплексной формой записи гармонических колебаний. Комплексная форма чрезвычайно удобна и для описания монохроматических волновых процессов. Формально переход к комплексной записи осуществляется следующим образом: наряду с функцией (2.4) рассмотрим функцию

$$E^{(i)}(\boldsymbol{r},t) = a(\boldsymbol{r})\sin[\omega t - \varphi(\boldsymbol{r})]. \tag{2.5}$$

Колебания, создаваемые волной (2.5) в любой точке $\boldsymbol{r}(x,y,z)$, отличаются от (2.4) только фазовым сдвигом в $\pi/2$ (иначе говоря, (2.5) получается из (2.4) только сдвигом начала отсчёта времени), так что если (2.4) является решением волнового уравнения, то и (2.5) также его решение. С помощью (2.4) и (2.5) образуем комплексную функцию

$$V(\boldsymbol{r},t) = E(\boldsymbol{r},t) - iE^{(i)}(\boldsymbol{r},t) = a(\boldsymbol{r})e^{-i[\omega t - \varphi(\boldsymbol{r})]}, \tag{2.6}$$

которую можно записать следующим образом:

$$V(\boldsymbol{r},t) = f(\boldsymbol{r})e^{-i\omega t}, \tag{2.7}$$

где

$$f(\boldsymbol{r}) = a(\boldsymbol{r})e^{i\varphi(\boldsymbol{r})}$$

— так называемая *комплексная амплитуда волны* (2.4) (зависящая только от координат, но не от времени!). Она содержит полную информацию о волне $E(\boldsymbol{r},t)$ (об амплитуде $a(\boldsymbol{r})$ и о фазе $\varphi(\boldsymbol{r})$). Достаточно уметь находить комплексную амплитуду, и тогда само реальное

волновое поле находится, как действительная часть функции $f(\boldsymbol{r})$, помноженной на $e^{-i\omega t}$:

$$E(\boldsymbol{r},t) = \mathrm{Re}[f(\boldsymbol{r})e^{-i\omega t}]. \tag{2.8}$$

Мы называем (2.4) волной, имея в виду, что амплитуда $a(\boldsymbol{r})$ и фаза $\varphi(\boldsymbol{r})$ не произвольны, но должны быть выбраны так, чтобы функция (2.4) (и, следовательно, (2.5) и их линейная комбинация (2.7)) удовлетворяла волновому уравнению. Подставив (2.7) в волновое уравнение, легко убедиться (упражнение для читателя), что комплексная амплитуда $f(\boldsymbol{r})$ должна удовлетворять уравнению

$$\nabla^2 f + k^2 f = 0, \tag{2.9}$$

где $k = \omega/c = 2\pi/\lambda$ — волновое число. Уравнение для комплексных амплитуд (2.9) называется *уравнением Гельмгольца*. Ещё раз подчеркнём, что комплексная амплитуда не содержит зависимости от времени, и если речь идёт о монохроматических волнах, то удобнее иметь дело с уравнением Гельмгольца и комплексной амплитудой, чем с волновым уравнением и реальной волной. В дальнейшем будем оперировать в основном с комплексной амплитудой, называя её для краткости полем.

Отметим два свойства решений уравнения (2.9), полностью аналогичных свойствам решений волнового уравнения.

1. Поле $f(x,y,z)$ в области $z > 0$ однозначно определяется заданием поля на плоскости $z = 0$:

$$f(x,y,z)\Big|_{z=0} = f_0(x,y).$$

2. Из линейности уравнения (2.9) следует, что если в области $z > 0$ мы имеем два решения $f_1(x,y,z)$ и $f_2(x,y,z)$, удовлетворяющие на плоскости $z = 0$ условиям

$$f_1(x,y,z)\Big|_{z=0} = f_1(x,y), \qquad f_2(x,y,z)\Big|_{z=0} = f_2(x,y),$$

то решение, удовлетворяющее на плоскости $z = 0$ условию

$$f(x,y,z)\Big|_{z=0} = c_1 f_1(x,y) + c_2 f_2(x,y),$$

есть

$$f(x,y,z) = c_1 f_1(x,y,z) + c_2 f_2(x,y,z). \tag{2.10}$$

Равенство (2.10) выражает принцип суперпозиции для комплексных амплитуд.

Задачу определения поля $g(x,y)$ (решения уравнения Гельмгольца) в некоторой плоскости $z = \text{const} > 0$ по заданному полю $f(x,y)$ в плоскости $z = 0$ можно рассматривать как задачу линейной фильтрации (ибо закон преобразования «входного сигнала» — поля в плоскости $z = 0$ в «выходной сигнал» — искомое поле в плоскости $z = \text{const} > 0$ определяется линейным уравнением Гельмгольца).

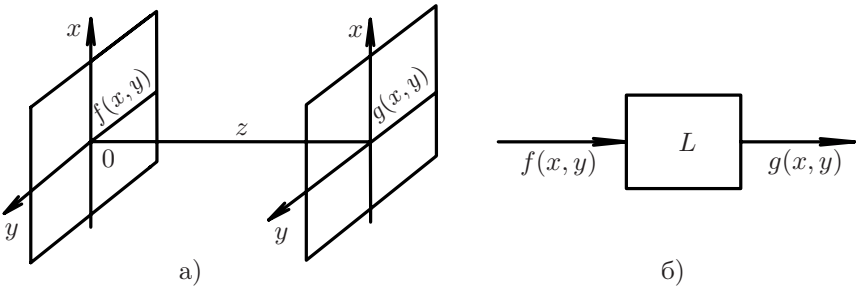


Рис. 2.1

Пространство между двумя плоскостями $z = 0$ и $z = \text{const} > 0$ — это простейший линейный пространственный фильтр, изображённый на рис. 2.1а. Входным сигналом фильтра является поле $f(x,y)$ в плоскости $z = 0$, а выходной сигнал фильтра — поле $g(x,y)$ в фиксированной плоскости $z = \text{const} > 0$. Эквивалентная блок-схема фильтра изображена на рис. 2.1б.

Операцию преобразования входного сигнала $f(x,y)$ в выходной $g(x,y)$ линейным фильтром можно, как и в главе I, записать в виде равенства

$$L[f(x,y)] = g(x,y). \tag{2.11}$$

Последовательность решения поставленной задачи традиционна: входной сигнал фильтра — поле $f(x,y)$ — необходимо представить в виде линейной комбинации некоторых элементарных, достаточно простых (для решения краевой задачи) полей. Искомое решение, согласно (2.10), является линейной суперпозицией решений, соответствующих каждому элементарному слагаемому «входного поля». Изучая временны́е линейные фильтры, мы выяснили, что в качестве элементарных слагаемых, суперпозицией которых представляется входной сигнал $f(t)$, целесообразно использовать гармонические колебания — функции $e^{i\omega t}$, ибо они являются собственными функциями линейного (стационарного) фильтра. Аналогичный вопрос необходимо поставить по отношению к исследуемому пространственному фильтру — найти

вид полей, которые в процессе распространения через свободное пространство не меняют своей пространственной конфигурации (так же, как гармонические колебания, «проходя» через временной фильтр, не изменяют своей «временной» структуры, оставаясь гармоническими колебаниями той же частоты).

Такими «собственными» волнами свободного пространства, или пространства, заполненного однородным диэлектриком, являются плоские волны.

§ 2.3. Плоские и сферические волны

Волновое уравнение и уравнение Гельмгольца допускают огромное разнообразие возможных волновых полей, с самыми разнообразными распределениями амплитуд и фаз колебаний в пространстве. Рассмотрим два наиболее простых решения — плоскую и сферическую волны.

1. Плоская бегущая волна

Так называется волна, определяемая равенством

$$V(\boldsymbol{r},t) = ae^{-i(\omega t - \boldsymbol{k}\cdot\boldsymbol{r})}, \tag{2.12}$$

где $a = a_0 e^{i\varphi_0}$ определяет и амплитуду волны a_0, и её начальную фазу φ_0 (в точке $\boldsymbol{r} = 0$ при $t = 0$). Здесь $\boldsymbol{r}(x,y,z)$ — радиус-вектор точки наблюдения P, $\boldsymbol{k}(k_x,k_y,k_z)$ — волновой вектор. Легко убедиться, что (2.12) удовлетворяет волновому уравнению, если $|\boldsymbol{k}| = \frac{\omega}{c} = \frac{2\pi}{\lambda}$. Как следует из определения (2.12), в любой фиксированный момент времени t для всех точек, координаты которых удовлетворяют уравнению плоскости

$$\boldsymbol{k}\cdot\boldsymbol{r} = k_x x + k_y y + k_z z = \text{const}, \tag{2.13}$$

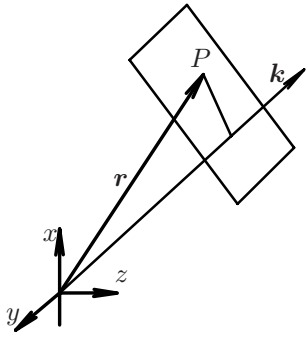


Рис. 2.2

фаза волны (т. е. фаза колебаний, создаваемых волной в этих точках) одна и та же, именно поэтому волна (2.12) называется *плоской*. Посмотрим, как перемещается со временем плоскость постоянной фазы. Из (2.13) следует, что вектор $\boldsymbol{k}$ перпендикулярен плоскостям постоянной фазы, а поскольку это справедливо для любого момента времени, вектор перемещения $d\boldsymbol{r}$ плоскости постоянной фазы совпадает с направлением вектора $\boldsymbol{k}$; очевидно, должно иметь место равенство

$$\omega t - \boldsymbol{k}\cdot\boldsymbol{r} = \omega(t+dt) - \boldsymbol{k}\cdot(\boldsymbol{r}+d\boldsymbol{r}) = \text{const},$$

из которого следует $\omega dt = \boldsymbol{k} \cdot d\boldsymbol{r} = kdr$; итак, скорость перемещения поверхности постоянной фазы равна

$$v_{\text{ф}} = \frac{dr}{dt} = \frac{\omega}{k} = v.$$

Таким образом, плоскости постоянной фазы перемещаются в пространстве по направлению вектора $\boldsymbol{k}$ со скоростью $v = c/\sqrt{\varepsilon}$.

Ясно, что в области $z > 0$ плоская волна (2.12) «убегает» от границы $z = 0$, если $k_z > 0$, и, наоборот, «бежит» к плоскости $z = 0$ при $k_z < 0$. На самой плоскости $z = 0$ две волны, отличающиеся только знаком k_z, создают одно и то же поле, т. е. удовлетворяют одному и тому же краевому условию

$$V\Big|_{z=0} = ae^{-i\omega t} \cdot e^{i(k_x x + k_y y)}.$$

Комплексная амплитуда плоской волны, по определению (2.7), есть

$$f(x,y,z) = ae^{i(k_x x + k_y y + \sqrt{k^2 - k_x^2 - k_y^2}\, z)}. \tag{2.14}$$

Непосредственной подстановкой можно убедиться, что (2.14) есть решение уравнения Гельмгольца, которое на плоскости $z = 0$ удовлетворяет условию

$$f(x,y,z)\Big|_{z=0} = f(x,y) = ae^{i(k_x x + k_y y)}. \tag{2.15}$$

две плоские волны в области $z > 0$, одна Сравнивая (2.14) и (2.15), мы видим, что поле плоской волны в некоторой плоскости $z = \text{const} > 0$ связано с полем этой волны в плоскости $z = 0$ равенством

$$g(x,y) = f(x,y,z)\Big|_{z=\text{const}} = f(x,y) \cdot e^{i\sqrt{k^2 - k_x^2 - k_y^2}\, z}, \tag{2.16}$$

т. е. поля отличаются множителем $e^{i\sqrt{k^2 - k_x^2 - k_y^2}\, z}$, определяющим набег фазы при распространении волны от одной плоскости до другой.

Простой факт, заключённый в равенстве (2.16), имеет интереснейшие следствия, если вновь обратиться к колебательно-волновым аналогиям. Воздействие пространственного фильтра на плоскую волну сводится, согласно (2.16), к умножению «входного поля» $e^{i(k_x x + k_y y)}$ на комплексное число

$$H(k_x,k_y) = e^{i\sqrt{k^2 - k_x^2 - k_y^2}\, z} \tag{2.17}$$

(при фиксированном z это число зависит от проекций волнового вектора плоской волны k_x, k_y). Следовательно, если входной сигнал фильтра есть плоская волна, то равенство (2.11) имеет вид

$$L\left[e^{i(k_x x+k_y y)}\right] = H(k_x,k_y)e^{i(k_x x+k_y y)}. \tag{2.18}$$

Соотношения (2.16) и (2.18) выражают один и тот же факт: плоские волны являются собственными функциями пространственного фильтра — пространства, заполненного однородным диэлектриком.

Другими словами, при распространении в однородном диэлектрике каждая плоская волна, характеризуемая волновым вектором $\boldsymbol{k}(k_x,k_y,k_z)$, остаётся плоской волной с тем же волновым вектором, возникает лишь набег фазы, определяемый комплексным числом $H(k_x,k_y)$. По существу, именно это и определяет исключительную роль плоских волн в изучении законов распространения оптических полей. Исследуя в главе I законы преобразования сигналов линейными колебательными системами, мы видели, что точно таким же свойством обладают гармонические колебания (равенство (1.11)).

Сопоставление соотношений (2.18) и (1.11) указывает на следующие колебательно-волновые аналогии.

Функция $H(k_x,k_y) = e^{i\sqrt{k^2-k_x^2-k_y^2}\,z}$ является аналогом частотной характеристики $H(\omega)$ линейной колебательной системы и может быть названа частотной характеристикой свободного пространства.

Поле плоской волны в плоскости $z = 0$: $e^{i(k_x x+k_y y)}$ является аналогом гармонического колебания $e^{i\omega t}$. Это одна из причин, по которой пару чисел k_x, k_y называют *пространственными частотами*.

Модуль функции $|H(k_x,k_y)| = 1$ можно называть амплитудной характеристикой, а $\arg H = \sqrt{k^2 - k_x^2 - k_y^2}\,z$ — фазовой характеристикой свободного пространства. Разумеется, на проекции вектора $\boldsymbol{k}(k_x,k_y)$ наложено принципиальное ограничение:

$$k_x^2 + k_y^2 \leqslant k^2 = \left(\frac{2\pi}{\lambda}\right)^2,$$

и все заключения о частотной характеристике свободного пространства справедливы при его выполнении. Подобного ограничения на частоту гармонических колебаний ω не существует в задачах электрической фильтрации (другой вопрос, что рост частоты ω может приводить к изменению свойств самого фильтра). Отметим также, что во временны́х задачах положительные и отрицательные частоты ω физически неразличимы. Отрицательные частоты возникли потому, что

мы использовали комплексную форму записи гармонических колебаний (если бы разложение сигналов проводилось по $\sin\omega t$ и $\cos\omega t$, то отрицательные частоты не возникали бы). Читатель может проверить (примеры на с. 30–34), что при разложении действительных функций $f(t)$ в спектр имеет место равенство $F(\omega) = F^*(-\omega)$, т. е. отрицательные частоты не несут никакой дополнительной информации о реальном сигнале $f(t)$. В то же время отрицательные *пространственные частоты* k_x, k_y имеют вполне понятный физический смысл: волна с волновым вектором $(-k_x, -k_y, k_z)$ отличается направлением распространения от волны с волновым вектором (k_x, k_y, k_z) и может реально существовать независимо от неё.

2. Сферическая волна

Эта волна определяется уравнением

$$V = \frac{a}{R} e^{-i(\omega t - kR)}, \tag{2.19}$$

где R — расстояние от источника $S(x_0,y_0,z_0)$ до точки наблюдения $P(x,y,z)$: $R = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}$.

Упражнение. Показать, что (2.19) удовлетворяет волновому уравнению, а комплексная амплитуда сферической волны

$$f(x,y,z) = a\frac{e^{ikR}}{R} \tag{2.20}$$

удовлетворяет уравнению Гельмгольца.

Упражнение. Показать, что если волна (2.20) бежит от источника (от точки S), то волна с комплексной амплитудой $f = a\frac{e^{-ikR}}{R}$ бежит к точке S — сходящаяся волна.

Зависимость фазы в (2.20) от координат имеет довольно сложный вид, поэтому часто пользуются приближением, заменяя сферический волновой фронт на небольшом участке параболическим. Поле, создаваемое источником S на плоскости $z = \text{const}$ (рис. 2.3), есть

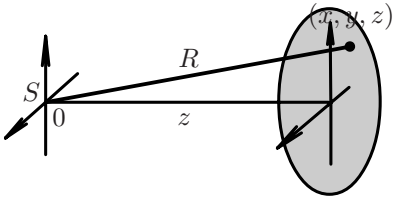


Рис. 2.3

$$f(x,y) = \frac{e^{ik\sqrt{x^2+y^2+z^2}}}{\sqrt{x^2+y^2+z^2}}.$$

Для небольшого участка волнового фронта вблизи начала координат (заштрихованная область на рис. 2.3) имеем

$$\sqrt{x^2+y^2+z^2} \approx z + \frac{x^2+y^2}{2z}. \tag{2.21}$$

С помощью (2.21) приближённо оценим фазу колебаний kR; что касается амплитудного множителя $1/R$, то достаточно положить для небольшой окрестности ($x^2+y^2 \ll z^2$): $1/R \approx 1/z$. Мы имеем

$$f(x,y) = \frac{e^{ikz}}{z} e^{i\frac{k}{2z}(x^2+y^2)}. \tag{2.22}$$

Приближение (2.22), называемое параболическим, справедливо, когда отброшенные члены разложения радикала в степенной ряд дают малую поправку в фазу.

Упражнение. Показать, что достаточное условие справедливости (2.22) есть $D^4/\lambda z^3 \ll 1$, D — размер заштрихованной области на плоскости $z = \text{const}$, для которой справедливо параболическое приближение.

§ 2.4. Произвольное оптическое поле как суперпозиция бегущих плоских волн

Имея в виду, что в дальнейшем нас будут интересовать законы распространения произвольной волны (с любым распределением амплитуд и фаз в пространстве), попытаемся ответить на вопрос: можно ли такую волну представить в виде суперпозиции бегущих плоских волн? Другими словами, можно ли в результате интерференции плоских волн с правильно подобранными амплитудами и фазами и различными волновыми векторами $\boldsymbol{k}(k_x,k_y,k_z)$ получить любое заранее заданное волновое поле $f(x,y)$. Точно так же при решении задачи о возбуждении электрического фильтра сигналом $f(t)$ мы представляли последний в виде суперпозиции гармонических колебаний $e^{i\omega t}$ и пришли при этом к преобразованию Фурье.

Без ограничения общности можно считать, что нас интересует результат интерференции бегущих плоских волн в некоторой плоскости $z = 0$, т. е. линейная суперпозиция функций $e^{i(k_x x + k_y y)}$. Линейная суперпозиция бегущих плоских волн в самом общем виде может быть записана так:

$$f(x,y) = \frac{1}{4\pi} \iint\limits_{k_x^2+k_y^2 \leqslant k^2} F(k_x,k_y) e^{i(k_x x + k_y y)}\, dk_x dk_y. \tag{2.23}$$

Подынтегральное выражение

$$\frac{1}{4\pi}F(k_x,k_y)dk_xdk_y\,e^{i(k_xx+k_yy)}$$

есть поле плоской волны с волновым вектором $\boldsymbol{k}(k_x,k_y,k_z)$, $(k_z = \sqrt{k^2-k_x^2-k_y^2})$. Амплитуда и начальная фаза этой волны определяются комплексным множителем $F(k_x,k_y)$. Амплитуда волны есть $a_0 = |F(k_x,k_y)|\,dk_xdk_y$, а начальная фаза $\varphi_0 = \arg F(k_x,k_y)$. Интеграл в правой части равенства означает, что результирующее поле получается суммированием всевозможных плоских волн (непрерывная суперпозиция со всеми возможными значениями k_x, k_y, лежащими в круге радиуса k: $k_x^2+k_y^2 \leqslant k^2$).

В случае дискретного набора плоских волн вместо интеграла получаем сумму

$$f(x,y) = \sum_n A_n e^{i(k_{xn}x+k_{yn}y)},$$

причём для всех n $k_{xn}^2+k_{yn}^2 \leqslant k^2$. Здесь амплитуду и фазу каждой плоской волны $\boldsymbol{k}_n(k_{xn},k_{yn})$ определяет комплексное число $A_n = a_n e^{i\varphi_n}$.

В выражении (2.23) каждой плоской волне с волновым вектором $\boldsymbol{k}(k_x,k_y,k_z)$ соответствует своё комплексное число $F(k_x,k_y)$ — тот вес, с которым данная волна входит в суммарное поле. Функция $F(k_x,k_y)$ называется спектром плоских волн суммарного поля $f(x,y)$. В случае дискретного набора плоских волн функция $F(k_x,k_y)$ отлична от нуля только в точках (k_{xn},k_{yn}) и спектр представляет собой дискретный набор чисел A_n.

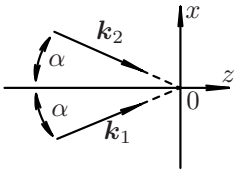


Рис. 2.4

Вопрос состоит в том, любая ли функция $f(x,y)$ может быть представлена равенством (2.23). Мы можем ответить на этот вопрос, рассмотрев простой пример. Найдём поле, образованное в плоскости $z = 0$ парой плоских волн с амплитудой a и волновыми векторами $\boldsymbol{k}_1(k_x,0,k_z)$ и $\boldsymbol{k}_2(-k_x,0,k_z)$, лежащими в плоскости $y = 0$ (рис. 2.4).

Волны f_1 и f_2 падают на плоскость $z = 0$ под углами $\pm\alpha$, причём $k_x = \frac{2\pi}{\lambda}\sin\alpha$, $k_z = \frac{2\pi}{\lambda}\cos\alpha$. Они образуют суммарное поле

$$f(x) = f_1(x) + f_2(x) = 2a\cos k_x x, \tag{2.24}$$

амплитуда которого изменяется по плоскости $z = 0$ по косинусоидальному закону, причём пространственный период равен

$$l = \frac{2\pi}{k_x} = \frac{\lambda}{\sin\alpha}. \tag{2.25}$$

Величина l является характерным масштабом, определяющим «быстроту» изменения функции $f(x)$. Для любой как угодно сложной функции $f(x)$ (или функции двух переменных $f(x,y)$) можно указать такой характерный масштаб l — расстояние, на котором происходит заметное изменение функции. Для поля (2.24) такой масштаб определяется формулой (2.25). Периодическую структуру со всё более мелким масштабом l можно получить, согласно (2.25), всё более увеличивая угол α. Минимальный период структуры получается при сложении пары плоских волн с углами $\alpha = \pm\pi/2$ (волны, бегущие навстречу друг другу).

Этой паре соответствует периодическая структура с $l_{\min} \simeq \lambda/2$. Можно составлять поле $f(x,y)$ в плоскости $z = 0$ из различного набора бегущих плоских волн, и из приведённого примера ясно, что характерный масштаб изменения такого поля всегда больше или порядка $\lambda/2$. Суперпозиция бегущих плоских волн самого общего вида — это поле $f(x,y)$, определяемое соотношением (2.23). Из сказанного выше следует, что, как бы мы ни подбирали спектр плоских волн (функцию $F(k_x,k_y)$), всегда образуется поле с «крупномасштабными» пространственными изменениями $l \gtrsim \lambda$.

§ 2.5. Преобразование Фурье функции двух переменных

Итак, мы убедились, что с помощью выражения (2.20) нельзя представить произвольное (например, обладающее «мелкомасштабными» неоднородностями $l < \lambda$) оптическое поле.

Расширим (пока чисто формально) пределы интегрирования в (2.23) до бесконечности, вводя для переменных интегрирования новые обозначения $k_x = u$, $k_y = v$:

$$f(x,y) = \frac{1}{4\pi^2} \iint\limits_{-\infty}^{\infty} F(u,v) e^{i(ux+vy)}\, du dv. \qquad (2.26)$$

Известная теорема фурье-анализа утверждает, что достаточно «хорошая» в математическом смысле функция $f(x,y)$ (в том числе и функция со сколь угодно мелкими пространственными неоднородностями) может быть представлена интегралом (2.26). Функция $F(u,v)$ называется *преобразованием Фурье* (или *спектром*) функции $f(x,y)$. Символически равенство (2.26) часто записывают в виде: $f(x,y) \Leftrightarrow F(u,v)$. Формула (2.26) является естественным обобщением на функции двух переменных выражения (1.12), связывающего сигнал $f(t)$ с его «временны́м» спектром $F(\omega)$, а переменные u и

v называются пространственными частотами. Таким образом, согласно (2.26), произвольная «хорошая» функция $f(x,y)$ может быть представлена суперпозицией экспоненциальных функций $e^{i(ux+vy)}$ (с коэффициентами-весами, определяемыми функцией $F(u,v)$). Образующие суммарное поле $f(x,y)$ экспоненциальные функции не обязательно имеют смысл поля бегущих плоских волн. Переменным u, v можно придать смысл проекций вектора $\boldsymbol{k}(u = k_x, v = k_y)$ только в том случае, если $u^2 + v^2 \leqslant k^2$ (тогда вектор $\boldsymbol{k}$ имеет z-компоненту — действительное число). В противном случае ($u^2 + v^2 > k^2$) пространственные частоты (u,v) уже не имеют смысла составляющих вектора $\boldsymbol{k}$ и соответствующие экспоненциальные функции уже не являются бегущими плоскими волнами. Физический смысл функций $e^{i(ux+vy)}$ при $u^2 + v^2 > k^2$ будет выяснен в гл. III.

По заданной функции $f(x,y)$ её спектр может быть найден по формуле

$$F(u,v) = \iint\limits_{-\infty}^{\infty} f(x,y)e^{-i(ux+vy)}\,dxdy, \tag{2.27}$$

вывод которой аналогичен выводу соотношения (1.28) — одномерного аналога (2.27).

§ 2.6. Оптические поля и их пространственные спектры. Соотношение неопределённости

Иллюстрацией формулы (2.27) являются рассмотренные ниже примеры преобразований Фурье, часто встречающиеся в оптических задачах. На этих примерах будет показано, как по заданной функции $f(x,y) = a(x,y)e^{i\varphi(x,y)}$ (комплексной амплитуде поля в плоскости $z = 0$) находить её спектр $F(u,v)$, т. е. весá, с которыми в суммарное поле $f(x,y)$ входят экспоненциальные функции $e^{i(ux+vy)}$.

Пример 1. Поле в плоскости $z = 0$ создано единственной плоской волной

$$f(x,y) = e^{i(u_0x+v_0y)},$$

где u_0 и v_0 — проекции волнового вектора $\boldsymbol{k}$ на оси x и y: $k_x = u_0$, $k_y = v_0$. Амплитуда поля есть константа $a(x,y) = 1$, а фаза линейно изменяется по плоскости $z = 0$: $\varphi(x,y) = u_0x + v_0y$. Если волна падает нормально на плоскость $z = 0$ (т. е. вектор $\boldsymbol{k}$ перпендикулярен этой плоскости), то $k_x = k_y = 0$ и, значит, $\varphi(x,y) = 0$, $f(x,y) = 1$.

Воспользуемся для нахождения спектра $F(u,v)$ общим соотношением (2.27), подставив вместо $f(x,y)$ функцию $e^{i(u_0x+v_0y)}$:

$$F(u,v)=\iint e^{i(u_0x+v_0y)}e^{-i(ux+vy)}\,dxdy=\iint e^{-i[(u-u_0)x+(v-v_0)y]}\,dxdy.$$

Здесь двойной интеграл есть просто произведение двух однократных интегралов:

$$F(u,v)=\int e^{i(u_0-u)x}\,dx\cdot\int e^{i(v_0-v)y}\,dy=4\pi^2\delta(u-u_0)\delta(v-v_0)$$

(мы воспользовались формулой (1.18), выведенной в § 1.6). Произведение $\delta(u-u_0)\delta(v-v_0)$ есть функция, равная нулю всюду, кроме точки (u_0,v_0), где эта функция равна бесконечности. Это произведение будем обозначать так: $\delta(u-u_0,v-v_0)$. Таким образом, спектр плоской волны есть δ-функция на частоте (u_0,v_0): $F(u,v)=\delta(u-u_0,v-v_0)$.

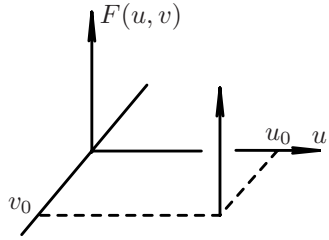


Рис. 2.5

Эта δ-функция изображена на рис. 2.5 в виде стрелки единичной длины, длина стрелки определяет величину интеграла

$$\int_{-\infty}^{\infty}\!\!\int\delta(u-u_0,v-v_0)\,dudv=1.$$

Полученный результат очевиден и мог бы быть написан сразу без использования общего соотношения. Действительно, ведь спектр $F(u,v)$ определяет, какие плоские волны дают вклад в суммарное поле $f(x,y)$. В данном случае суммарное поле $f(x,y)$ состоит из одной единственной плоской волны, поэтому спектр равен нулю всюду, за исключением точки (u_0,v_0). Эта точка определяет проекции волнового вектора той единственной волны, которая и образует суммарное поле. Если $u_0=v_0=0$, то спектр есть δ-функция в начале координат: $F(u,v)=\delta(u,v)$.

Пример 2.

$$f(x,y)=p_a(x)p_b(y)=\begin{cases}1 & \text{при} \quad |x|\leqslant a, \quad |y|\leqslant b,\\ 0 & \text{при} \quad |x|>a, \quad |y|>b.\end{cases}\tag{2.28}$$

(Если в плоскости $z=0$ установить непрозрачный экран с отверстием прямоугольной формы и осветить его слева нормально падающей

плоской волной единичной амплитуды (рис. 2.6), то на выходе из экрана в плоскости $z = 0_+$ образуется именно такое поле $f(x,y)$.) Подставляя (2.28) в (2.27), находим

$$F(u,v) = \int_{-\infty}^{\infty} p_a(x) e^{-iux}\, dx \cdot \int_{-\infty}^{\infty} p_b(y) e^{-ivy}\, dy.$$

Искомый спектр есть произведение двух одномерных преобразований. Спектр одномерной прямоугольной функции был найден ранее (1.34). Нужно только заменить аргумент ω на u или v:

$$F(u,v) = 4\frac{\sin au}{u} \cdot \frac{\sin bv}{v}. \qquad (2.29)$$

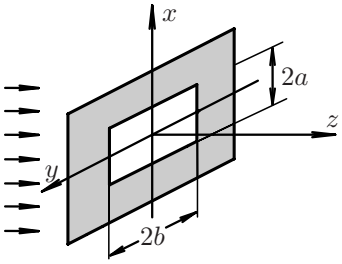


Рис. 2.6

В оптике часто имеют дело с полями, которые представляются в виде $f(x,y) = f(x)f(y)$. Очевидно, спектр $F(u,v)$ является в этом случае произведением двух одномерных преобразований: $F(u,v) = F(u)F(v)$; в частности, если поле зависит лишь от одной переменной $f(x,y) = f(x)$, то его спектр определяется одномерным преобразованием Фурье:

$$F(u) = \int_{-\infty}^{\infty} f(x) e^{-iux}\, dx. \qquad (2.30)$$

Ниже рассмотрен ряд примеров таких одномерных полей.

Пример 3. Синусоидальный закон изменения амплитуды поля (синусоидальная амплитудная решётка):

$$f(x) = 1 + \alpha \cos \Omega x, \qquad (2.31)$$

где $\alpha < 1$ — «глубина модуляции». Поле (2.31) — действительная положительная функция, следовательно, плоскость $z = 0$ является плоскостью постоянной (нулевой фазы): $\varphi(x) = 0$. Амплитуда $a(x)$ изменяется по синусоидальному закону.

Подставляя (2.31) в формулу (2.30), находим

$$F(u) = \int (1 + \alpha \cos \Omega x) e^{-iux}\, dx = \\ = \delta(u) + \alpha \int \frac{e^{i\Omega x} + e^{-i\Omega x}}{2} e^{-iux}\, dx = \\ = \delta(u) + \frac{\alpha}{2} \left[\int e^{-i(u+\Omega)x}\, dx + \int e^{-i(u-\Omega)x}\, dx \right], \quad (2.32)$$

или окончательно

$$F(u) = \delta(u) + \frac{\alpha}{2}[\delta(u + \Omega) + \delta(u - \Omega)]. \quad (2.33)$$

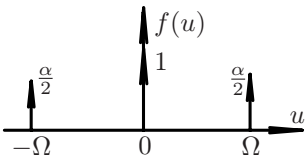


Рис. 2.7

Мы получили действительный спектр, а это означает, что спектральные компоненты $\delta(u+\Omega)$ и $\delta(u-\Omega)$ совпадают по фазе с компонентой $\delta(u)$.

Спектр (2.33) изображён на рис. 2.7. Три стрелочки — это три δ-функции: стрелка в начале координат единичной длины и две стрелки на частотах $-\Omega$ и $+\Omega$ длиной $\alpha/2$. Можно было бы получить спектр (2.33), не используя общее соотношение: единица в формуле (2.31) означает, что в состав поля $f(x)$ входит плоская волна единичной амплитуды, нормально падающая на плоскость $z = 0$ (и, следовательно, в спектре имеем δ-функцию на нулевой частоте, так как для этой волны $k_x = k_y = 0$). Слагаемое $\alpha \cos \Omega x$ образовано двумя плоскими волнами: $f_1 = \frac{\alpha}{2} e^{i\Omega x}$ и $f_2 = \frac{\alpha}{2} e^{-i\Omega x}$ (так что $f_1 + f_2 = \alpha \cos \Omega x$), которые дают в спектре δ-функции на частотах $\pm\Omega$. Таким образом, поле (2.31) состоит из трёх плоских волн: волны единичной амплитуды, нормально падающей на плоскость $z = 0$, и двух плоских волн с амплитудой $\alpha/2$, падающих под углами $\sin \alpha = \pm\Omega$ и синфазных с первой волной (колебания, создаваемые всеми волнами в начале координат совпадают по фазе).

Используя радиотехническую терминологию, можно сказать, что спектр состоит из «несущего колебания» (плоская волна, распространяющаяся вдоль оси z) и двух «боковых частот» (плоские волны под углами $\pm\alpha$ к оси z).

Пример 4. Синусоидальный закон изменения фазы поля:

$$f(x) = e^{i\alpha \cos \Omega x}. \quad (2.34)$$

Поле (2.34) представляет собой чисто фазовую структуру: амплитуда поля есть константа $a(x) \equiv 1$, а фаза изменяется по синусоидальному закону. Отметим здесь, что любую чисто фазовую структуру невозможно наблюдать «просто так», ведь все приёмники оптического излучения — это квадратичные приёмники, реагирующие на интенсивность света: глаз, фотоэлемент, фотопластинка регистрируют величину $I = ff^*$. Для любой фазовой структуры $f(x) = ae^{i\varphi(x)}$ имеем $ff^* = a^2 = \text{const}$, т. е. чисто фазовая структура невидима.

Найдём спектр поля (2.34) в предположении малой «глубины модуляции» фазы: $\alpha \ll 1$. Тогда $f(x) \approx 1 + i\alpha \cos \Omega x$, и для спектра, как и в примере 3, получим

$$F(u) = \delta(u) + i\frac{\alpha}{2}[\delta(u - \Omega) + \delta(u + \Omega)]. \qquad (2.35)$$

Таким образом, спектр (2.35) отличается от спектра (2.33) только наличием фазового множителя $e^{i\frac{\pi}{2}} = i$ перед квадратной скобкой. Интереснейший факт, к которому мы пришли, разбирая примеры 3 и 4, состоит в том, что спектры амплитудной структуры (2.31) и фазовой структуры (2.34) отличаются *только фазой боковых спектральных компонент*: для амплитудной структуры боковые спектры $\delta(u + \Omega)$ и $\delta(u - \Omega)$ синфазны с «несущим» колебанием $\delta(u)$. Другими словами, три плоские волны, составляющие поле (2.34) (волна $f_1(x) = 1$, распространяющаяся вдоль оси z, и волны $f_2 = \frac{\alpha}{2}e^{i\Omega x}$ и $f_3 = \frac{\alpha}{2}e^{-i\Omega x}$, составляющие углы $\pm\Omega/k$ с осью z), приходят в начало координат плоскости $z = 0$ синфазно, в то время как для фазовой структуры «боковые частоты» запаздывают по фазе на $\pi/2$.

Значит, стоит только изменить фазы боковых спектральных компонент (или фазу несущей) на $\pi/2$, как невидимая фазовая структура (2.34) превращается в видимую амплитудную структуру (2.31). Такой принцип используется в знаменитом «методе фазового контраста», используемом для наблюдения фазовых объектов (см. далее § 5.3).

Пример 5. Прямоугольная решётка (рис. 2.8):

$$f(x) = \sum_{n=-N}^{n=N} p_b(x - nd). \qquad (2.36)$$

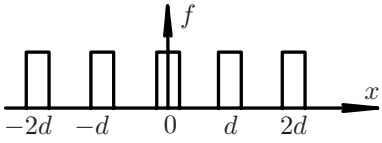


Рис. 2.8

Подставляя (2.36) в (2.30), находим

$$F(u) = \sum_{n} \int_{nd-b}^{nd+b} e^{-iux}\, dx,$$

или

$$F(u) = \frac{\sin bu}{u} \sum_n e^{iund}.$$

Окончательное выражение для спектра прямоугольной решётки имеет вид

$$F(u) = e^{i(N-1/2)u} \frac{\sin bu}{u} \left| \frac{\sin Ndu}{\sin du} \right| \tag{2.37}$$

и может быть найдено с помощью формулы (3.25), выведенной в § 3.5.
Задача. Определить положение и ширину главных максимумов функции (2.37).

Соотношение неопределённостей

Рассматриваемые оптические поля являются функциями *пространственных координат* (в общем случае координат на плоскости x, y). Преобразуя эти поля по Фурье, мы получаем функции двух переменных — пространственных частот u, v. Все свойства рассмотренного в гл. I преобразования Фурье функции одной переменной t переносятся на функции двух переменных x, y — ведь математику не очень заботит различие в физическом смысле переменных. Одним из фундаментальных свойств преобразования Фурье является соотношение неопределённостей. Это соотношение связывает пространственную протяжённость функции Δx, Δy с шириной её пространственного спектра Δu, Δv:

$$\Delta x \cdot \Delta u \gtrsim 2\pi, \qquad \Delta y \cdot \Delta v \gtrsim 2\pi. \tag{2.38}$$

Пара соотношений (2.38) является аналогом соотношения неопределённостей (1.44), связывающего длительность сигнала Δt с шириной его спектра $\Delta\omega$. Хорошей иллюстрацией (2.38) является пример 2. Протяжённость функции (2.28) есть $\Delta x = 2a$, $\Delta y = 2b$. Ширина её спектра, как это следует из (2.29), есть $\Delta u = 2\pi/a$, $\Delta v = 2\pi/b$. Эту ширину мы оцениваем по ширине главных лепестков функций $\frac{\sin au}{u}$ и $\frac{\sin bv}{v}$. Таким образом, в данном примере $\Delta x \cdot \Delta u \simeq 4\pi$, $\Delta y \cdot \Delta v \simeq 4\pi$.

Глава III

Дифракция

§ 3.1. Постановка задачи

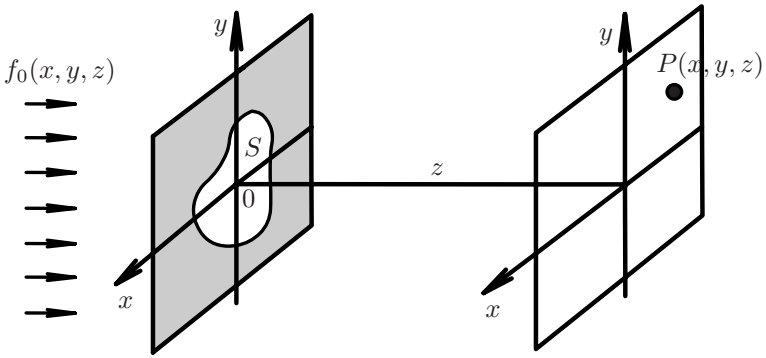


Рис. 3.1

В современной физике термин *дифракция* охватывает многообразные явления, связанные с распространением волн в неоднородной среде. В частности, неоднородной средой можно считать непрозрачный экран с отверстием (рис. 3.1). В такой дифракционной задаче (которую мы и будем рассматривать) однородность пространства нарушается в плоскости $z = 0$, где расположен экран. Пусть слева экран с отверстием освещается монохроматической волной, комплексная амплитуда которой $f_0(x,y,z)$ в области $z < 0$ определяет распределение амплитуд и фаз колебаний напряжённости поля в волне. Требуется определить, какова комплексная амплитуда волны в области $z > 0$ и, в частности, в некоторой точке $P(x,y)$, расположенной в плоскости $z = \text{const} > 0$. Для решения этой задачи необходимо найти ответ на два основных вопроса. *Первый вопрос* — определение граничных условий — ставится следующим образом: известно поле $f_0(x,y) = f_0(x,y,z)|_{z=0}$, созданное в плоскости $z = 0$ (*в отсутствие экрана!*) какими-либо «сторонними»

источниками. Это может быть поле плоской волны, нормально или косо падающей на плоскость $z = 0$, или поле точечного источника — сферическая волна. Совершенно неважно, с помощью каких «сторонних» источников и каким образом это поле создано — мы его считаем заданным. Затем в плоскость $z = 0$ помещается непрозрачный экран с отверстием. Требуется ответить на вопрос: какое поле $f(x,y)$ возникает в плоскости $z = 0_+$, непосредственно примыкающей к экрану справа от него. Это и есть то граничное поле, которое необходимо знать для решения дифракционной задачи.

Второй вопрос: каким образом, зная граничное поле $f(x,y)$, определить искомое поле $g(x,y)$ в некоторой плоскости $z = \text{const} > 0$? Математики такую задачу называют *краевой*: нужно найти решение уравнения Гельмгольца, удовлетворяющее заданным граничным условиям. Фактически это задача о распространении оптического сигнала в свободном пространстве: задано оптическое поле $f(x,y)$ в некоторой плоскости $z = 0_+$, надо знать, как это поле преобразуется в процессе распространения оптического сигнала и, в частности, какое поле $g(x,y)$ получится в некоторой плоскости $z = \text{const} > 0$ за экраном.

Решение дифракционной задачи мы начнём с ответа на второй вопрос.

§ 3.2. Распространение волн в свободном пространстве

Для начала будем предполагать, что заданное на границе $z = 0$ волновое поле можно представить в виде суперпозиции бегущих плоских волн (§ 2.4):

$$f(x,y) = \iint\limits_{u^2+v^2 \leqslant k^2} F(u,v) e^{i(ux+vy)}\, du dv. \tag{3.1}$$

Искомое поле $g(x,y)$ в плоскости $z = \text{const} > 0$ представляется в аналогичном виде:

$$g(x,y) = \iint\limits_{u^2+v^2 \leqslant k^2} G(u,v) e^{i(ux+vy)}\, du dv. \tag{3.2}$$

Напомним, что каждое слагаемое в (3.1)

$$df(x,y) = F(u,v) du dv \cdot e^{i(ux+vy)}$$

есть поле плоской волны с волновым вектором $\boldsymbol{k}(u,v)$, амплитудой $|F(u,v)|dudv$ и начальной фазой $\arg F(u,v)$. Мы уже знаем, что поле, создаваемое плоской волной в плоскости $z = \text{const} > 0$, связано с

полем этой волны в плоскости $z = 0$ равенством (2.16):

$$dg(x,y) = df(x,y) \cdot e^{i\sqrt{k^2-u^2-v^2}\,z}. \tag{3.3}$$

(Каждая плоская волна при распространении между двумя плоскостями приобретает набег фазы $\Delta\varphi = k_z z = \sqrt{k^2-u^2-v^2}\,z$.) Искомое поле в плоскости наблюдения $z = \text{const}$ является линейной суперпозицией полей (3.3), так что мы имеем

$$g(x,y) = \iint\limits_{u^2+v^2\leqslant k^2} F(u,v)e^{i\sqrt{k^2-u^2-v^2}\,z} \cdot e^{i(ux+vy)}\,dudv. \tag{3.4}$$

Из сравнения (3.4) и (3.2) следует, что спектр плоских волн искомого поля $g(x,y)$ есть

$$G(u,v) = F(u,v) \cdot e^{i\sqrt{k^2-u^2-v^2}\,z}. \tag{3.5}$$

Ещё раз подчеркнём смысл соотношения (3.5): каждая плоская волна, входящая в состав граничного поля $f(x,y)$, достигает плоскости наблюдения с неизменной амплитудой и волновым вектором (направлением распространения); приобретается лишь набег фазы, различный для разных плоских волн (для разных чисел u и v), и, следовательно, *изменяются фазовые соотношения между слагаемыми плоскими волнами* — именно это и приводит к отличию искомого поля $g(x,y)$ от граничного значения $f(x,y)$.

Теперь рассмотрим более общий случай: граничное поле $f(x,y)$ нельзя представить в виде (3.1). Такая ситуация возникает, когда $f(x,y)$ резко изменяется на расстояниях, сравнимых с длиной волны (§ 2.4). В этом случае имеет место более общее представление функции $f(x,y)$ — в виде интеграла Фурье (2.26), т. е. в виде суперпозиции не только бегущих плоских волн ($u^2 + v^2 \leqslant k^2$), но и функций $e^{i(ux+vy)}$, для которых $u^2 + v^2 > k^2$. Мы покажем, не изучая подробно физического смысла этих функций, что их вклад в искомое поле в плоскости $z = \text{const} > 0$ несуществен, если $z \gg \lambda$.

Представим искомую функцию $g(x,y,z)$ в виде интеграла Фурье:

$$g(x,y,z) = \iint G(u,v,z)e^{i(ux+vy)}\,dudv. \tag{3.6}$$

Написанное выражение следует понимать таким образом: мы фиксировали плоскость $z = \text{const}$ и, рассматривая поле g как функцию двух переменных (x,y), представили её в виде интеграла Фурье.

При этом отмечено, что спектр $G(u,v,z)$ является не только функцией пространственных частот, но и параметра z. Комплексная амплитуда $g(x,y,z)$, определяемая с помощью (3.6), должна удовлетворять уравнению Гельмгольца; достаточно потребовать, чтобы этому уравнению удовлетворяла подынтегральная функция в (3.6). Подставляя её в уравнение Гельмгольца (2.9), получаем для спектра $G(u,v,z)$ обыкновенное дифференциальное уравнение (по переменной z):

$$\frac{d^2}{dz^2}G + \left(k^2 - u^2 - v^2\right) G = 0.$$

Легко убедиться, что общее решение этого уравнения имеет вид

$$G(u,v,z) = C_1 e^{i\sqrt{k^2-u^2-v^2}\,z} + C_2 e^{-i\sqrt{k^2-u^2-v^2}\,z}, \tag{3.7}$$

где C_1 и C_2 — произвольные постоянные. Напомним, что (3.7) справедливо как при $u^2 + v^2 \leqslant k^2$, так в при $u^2 + v^2 > k^2$. Константу C_2 сразу можно положить равной нулю: действительно, для достаточно больших частот u, v ($u^2 + v^2 > k^2$) показатель экспоненты во втором слагаемом в (3.7) оказывается положительным, т. е. второе слагаемое неограниченно нарастает при $z \to \infty$; ясно, что такое «усиление» волны при распространении в свободном пространстве $z > 0$, в котором нет источников волн, невозможно. Что касается значений $u^2+v^2 \leqslant k^2$, то при этом второе слагаемое соответствует плоской волне, бегущей к плоскости $z = 0$ (§ 2.3). Таких «отражённых» волн не существует в полупространстве $z > 0$, заполненном однородным диэлектриком (или просто свободном), все бегущие волны должны иметь положительное k_z, следовательно, и в этом случае следует считать $C_2 = 0$. Константа C_1 может быть определена, если известен спектр граничного поля (при $z = 0$): $C_1 = F(u,v) = G(u,v,z)|_{z=0}$, Следовательно:

$$G(u,v,z) = F(u,v)e^{i\sqrt{k^2-u^2-v^2}\,z}. \tag{3.8}$$

Мы снова пришли к соотношению (3.5) и убедились, таким образом, что оно справедливо как при $u^2 + v^2 \leqslant k^2$, так и при $u^2 + v^2 > k^2$. Именно второе неравенство теперь необходимо обсудить. При условии $u^2 + v^2 > k^2$ экспоненциальный множитель можно записать в виде $e^{i\sqrt{k^2-u^2-v^2}\,z} = e^{-\mu z}$, где $\mu = |\sqrt{k^2 - u^2 - v^2}|$ — положительное число, следовательно,

$$G(u,v,z) = F(u,v)e^{-\mu z}, \tag{3.9}$$

т. е. спектральные компоненты с частотами, удовлетворяющими условию $u^2 + v^2 > k^2$, экспоненциально затухают с ростом z. Волны, соответствующие таким спектральным компонентам, называются *неоднородными* (или *затухающими*) *волнами*.

Оценим скорость затухания для конкретного примера. Пусть $u^2+v^2=1{,}1k^2$, тогда $\mu z \approx 2\frac{z}{\lambda}$. Если плоскость наблюдения отнести от плоскости $z=0$ на расстояние $z=2\lambda$, то имеем: $G(u,v,z) \approx 0{,}01F(u,v)$, т. е. амплитуда соответствующей спектральной составляющей примерно в 100 раз спадает по сравнению с её значением в плоскости $z=0$. Таким образом, достаточно отнести плоскость наблюдения хотя бы на несколько длин волн (а в оптике обычно интересуются расстояниями $z \gg \lambda$), как вкладом гармоник с пространственной частотой, удовлетворяющей условию $u^2+v^2>k^2$, можно пренебречь и считать, что при этом

$$G(u,v,z) \approx 0. \tag{3.10}$$

Отметим, что предельный случай $u^2+v^2=k^2$ соответствует плоской волне, распространяющейся перпендикулярно оси z.

Формулы (3.9) и (3.10) отражают чрезвычайно интересный факт. Пусть поле в плоскости $z=0$ имеет пространственные неоднородности, меньшие длины световой волны. Это означает, что в спектре поля имеются пространственные гармоники с частотами, удовлетворяющими условию $u^2+v^2>k^2$ (*соотношение неопределённостей* (2.38)!). Как мы выяснили, таким гармоникам соответствуют неоднородные волны, вкладом которых в суммарное поле в плоскости наблюдения $z=\text{const} \gg \lambda$ можно пренебречь. Исчезновение «высокочастотных» пространственных гармоник означает, что при распространении оптического сигнала *сглаживаются и исчезают пространственные неоднородности, меньшие длины волны.*

Резюмируя сказанное, можно сделать следующие выводы. Все изменения в пространственном распределении амплитуд и фаз в волне при её распространении сводятся к двум эффектам:
1) изменяются фазовые соотношения между бегущими плоскими волнами (согласно равенству (3.5));
2) сглаживаются мелкие пространственные неоднородности (меньшие длины волны) в связи с исчезновением неоднородных волн.

Отметим, что даже если в исходном распределении (в плоскости $z=0$) нет высокочастотных спектральных составляющих $u^2+v^2>k^2$, то и тогда эффект изменения фазовых соотношений между бегущими плоскими волнами приводит к тому, что пространственное распределение в плоскости наблюдения $z=\text{const}$ может кардинальным образом отличаться от исходного.

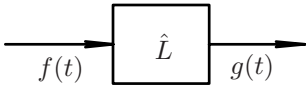


Рис. 3.2

Соотношение (3.5) следует понимать как связь между спектрами входного и выходного сигналов линейного пространственного филь-

тра, и, следовательно, величина

$$H(u,v) = e^{i\sqrt{k^2-u^2-v^2}\,z}$$

представляет собой *частотную характеристику свободного пространства*; оптической схеме, изображённой на рис. 3.1, можно поставить в соответствие эквивалентную блок-схему (рис. 3.2). Как следует из проведённого выше анализа, свободное пространство действительно обладает ярко выраженными фильтрующими свойствами: пространственные гармоники с частотами $u^2+v^2 > k^2$ практически не пропускаются нашим фильтром, так как $|H(u,v)| \approx 0$ при $u^2+v^2 > k^2$ (если $z \gg \lambda$). В то же время все пространственные гармоники в полосе $u^2 + v^2 \leqslant k^2$ пропускаются без искажения по амплитуде, хотя и с различными фазовыми набегами. Свойства фильтра — свободного пространства — можно характеризовать амплитудной характеристикой $|H(u,v)|$ и фазовой характеристикой $\arg H(u,v)$, которые изображены на рис. 3.3. Пренебрежение вкладом неоднородных волн означает, что реальная амплитудная характеристика (рис. 3.3а) заменяется прямоугольником ширины $2k$.

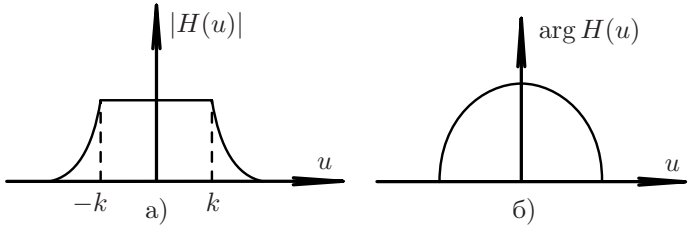


Рис. 3.3

§ 3.3. Граничные условия Кирхгофа

Пока мы ничего не говорили о том, как найти граничное поле и, следовательно, его спектр $G(u,v,0) = F(u,v)$, если оптическое поле, созданное в плоскости $z = 0$ в *отсутствие экрана* «сторонними» источниками, задано и равно $f_0(x,y)$.

Надо сказать, что точное определение граничного поля является чрезвычайно сложной проблемой, решенной до сих пор лишь для немногих дифракционных задач. Действительно, ведь поле, созданное «сторонними источниками», должно на поверхности экрана (если это экран проводящий) индуцировать токи, так что для суммарного поля (стороннего плюс поля токов на экране) должны выполняться известные в электродинамике граничные условия. Мы, следуя Кирхгофу,

определим *приближённые граничные условия* (поле $f(x,y)$) в плоскости $z = 0_+$) следующим равенством:

$$f(x,y) = \begin{cases} f_0(x,y), & \text{если точка } (x,y) \text{ лежит внутри отверстия } S, \\ 0, & \text{если точка } (x,y) \text{ лежит вне отверстия } S. \end{cases} \tag{3.11}$$

Смысл граничных условий Кирхгофа состоит в следующем: в той части плоскости $z = 0_+$, которая занята отверстием (т.е. не «затенена» непрозрачным экраном), поле предполагается равным полю сторонних источников. В той части плоскости $z = 0$, которая «затенена» непрозрачным экраном, граничное поле полагается равным нулю. Следует ясно понимать, что граничные условия Кирхгофа являются приближёнными — наличие экрана в действительности приводит к тому, что поле на отверстии отлично от падающего поля $f_0(x,y)$ и особенно сильно оно искажено вблизи краев отверстия. Но, как показывает сравнение теоретических выводов, основанных на приближённых граничных условиях Кирхгофа, с экспериментом, согласие оказывается поразительным *при условии, что линейные размеры отверстия D велики по сравнению с длиной волны $D > \lambda$.* (Сильное неравенство в действительности не является необходимым: достаточно потребовать, чтобы размеры отверстия превосходили длину волны хотя бы в несколько раз.)

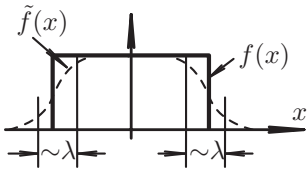


Рис. 3.4

Анализ поведения неоднородных волн позволяет качественно понять, почему замена точных граничных условий (согласно которым поле должно по определённому закону $\tilde{f}(x,y)$ спадать от значения $f_0(x,y)$ в области отверстия (вдали от края) до нуля где-то в области тени (рис. 3.4)) приближёнными условиями (в которых переход от значения $f_0(x,y)$ к нулю происходит скачком) не приводит к заметным ошибкам. Дело в том, что замена точного решения $\tilde{f}(x,y)$ приближением $f(x,y)$ приводит к резкому изменению спектра лишь на больших пространственных частотах $u^2 + v^2 \gtrsim k^2$, так происходит всегда, когда плавные изменения заменяются резкой ступенькой; но мы выяснили, что это не может заметно сказываться на искомом поле, поскольку высоким частотам $u^2 + v^2 \gtrsim k^2$ соответствуют неоднородные волны, амплитуда которых быстро уменьшается с ростом z и вкладом которых можно пренебречь при $z \gg \lambda$. Если $D > \lambda$, то всё определяется бегущими плоскими волнами, составляющими граничное поле; как бы мы ни меняли спектр неоднородных волн, ничего в конечном резуль-

тате не изменится. Поскольку размер отверстия определяет область пространственных частот Δu граничного поля ($\Delta u \cdot D \simeq 2\pi$), то условие $D > \lambda$ можно сформулировать на частотном языке: $\Delta u < k$.

Если ввести «функцию пропускания экрана», расположенного в плоскости $z = 0$, по формуле

$$p_s(x,y) = \begin{cases} 1 & \text{при } (x,y) \in S, \\ 0 & \text{при } (x,y) \notin S, \end{cases}$$

то граничные условия Кирхгофа можно записать в виде

$$f(x,y) = f_0(x,y) \cdot p_s(x,y). \tag{3.12}$$

Преобразуя (3.12) по Фурье, можно найти связь между пространственными спектрами волны, падающей на экран, и волны, прошедшей через отверстие. Используя (1.25), получаем

$$F(u,v) = F_0(u,v) \otimes\otimes \mathcal{P}_S(u,v), \tag{3.13}$$

((3.13) — операция свёртки функций двух переменных $F_0(u,v)$ и $\mathcal{P}_S(u,v)$), где $\mathcal{P}_S(u,v)$ — преобразование Фурье единично-нулевой функции $p_s(x,y)$:

$$\mathcal{P}_S(u,v) = \iint p_s(x,y) e^{-i(ux+vy)}\, dxdy.$$

Одномерные аналоги формул (3.12) и (3.13) имеют вид

$$f(x) = f_0(x) \cdot P_{_D}(x), \tag{3.14}$$

$$F(u) = F_0(u) \otimes \frac{\sin \frac{D}{2}u}{u}. \tag{3.15}$$

(В (3.15) использовано выражение (1.34) для фурье-образа единично-нулевой функции.)

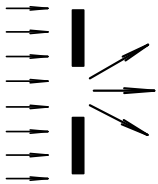


Рис. 3.5

Рассмотрим простой пример, поясняющий, каким образом граничные условия Кирхгофа приводят к явлению дифракции.

Согласно законам геометрической оптики, узкий параллельный пучок света можно получить с помощью щели в непрозрачном экране, освещаемом плоской волной (рис. 3.5). Плоская волна, освещающая экран, имеет пространственный спектр $F_0(u) = \delta(u)$ (см. пример 1

§ 2.6); пространственный спектр волны за щелью определим с помощью (3.15):

$$F(u) = \delta(u) \otimes \frac{\sin \frac{D}{2}u}{u} = \int \frac{\sin \frac{D}{2}v}{v}\delta(u-v)\ ,dv = \frac{\sin \frac{D}{2}u}{u}.$$

Согласно последнему равенству, волна за отверстием уже не является плоской, она представляется суперпозицией плоских волн с различными направлениями. Разброс пространственных частот Δu, оцениваемый по ширине главного максимума функции $\frac{\sin \frac{D}{2}u}{u}$, есть $\Delta u = \frac{2\pi}{D}$. Соответствующий разброс направлений волновых векторов плоских волн за отверстием есть $\Delta\alpha \simeq \frac{\Delta u}{k} \simeq \frac{\lambda}{D}$ (поскольку $u = k\sin\alpha \approx k\alpha$). Таким образом, чем более узкий пучок мы хотим получить (меньше D), тем шире спектр плоских волн за отверстием и, значит, больше дифракционная расходимость этого пучка — таков вывод из соотношений (3.12) и (3.13).

Теперь можно описать всю схему решения дифракционной задачи в такой последовательности.

1. По заданному полю «сторонних» источников $f_0(x,y)$ в плоскости $z = 0$ определяем поле $f(x,y)$ согласно граничным условиям Кирхгофа (3.11) и находим его спектр $F(u,v)$ — преобразование Фурье функции $f(x,y)$.

2. Зная спектр $F(u,v)$ на границе $z = 0_+$, находим спектр $G(u,v)$ в плоскости наблюдения $z = \text{const}$ с помощью равенства (3.5), справедливого для любых (u,v).

3. Искомое дифракционное поле в плоскости наблюдения есть обратное преобразование Фурье спектра $G(u,v)$, найденного по формуле (3.5):

$$g(x,y) = \frac{1}{4\pi^2}\iint F(u,v)e^{i\sqrt{k^2-u^2-v^2}\,z}e^{i(ux+vy)}\,dudv. \tag{3.16}$$

Замечание: одномерный аналог формулы (3.16) имеет вид

$$g(x) = \frac{1}{2\pi}\int F(u)e^{i\sqrt{k^2-u^2}\,z}e^{iux}\,du, \tag{3.17}$$

которому соответствует следующая связь между спектрами:

$$G(u) = F(u)\cdot e^{i\sqrt{k^2-u^2}\,z}, \tag{3.18}$$

где

$$H(u) = e^{i\sqrt{k^2-u^2}\,z} \tag{3.19}$$

— частотная характеристика свободного пространства в одномерном случае.

Может возникнуть естественный вопрос: стоит ли так много заниматься решением одной единственной задачи — дифракцией волны на отверстии в непрозрачном экране? Чтобы ответить на него, мы прервем последовательное изложение и обратимся к вопросу, который неожиданным образом имеет отношение к делу.

§ 3.4. Временна́я модуляция в радио и пространственная модуляция в оптике

Рассмотрим в самом общем виде принципы передачи сообщений в радиотехнике. Основным элементом любой радиостанции является генератор-источник высокочастотного ($10^5 - 10^8$ Гц) гармонического колебания $A_0 e^{i\omega_0 t}$, которое называют *несущим* и которое само по себе не содержит никакой информации. Для передачи сообщения (речи, музыки, телеизображения, вообще, любого сигнала) необходимо нарушение синусоидальности колебаний. *Отклонения от синусоидальности*, выражающие содержание радиопередачи, *называют в радиотехнике модуляцией*.

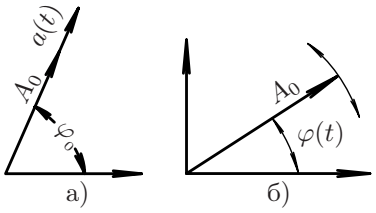


Рис. 3.6

Различают *амплитудную модуляцию*, которая на диаграмме (рис. 3.6а) изображается в виде вектора постоянного направления и переменной длины $A(t) = A_0 + a(t)$, и *частотную* (или *фазовую*) *модуляцию*, изображаемую вектором неизменной длины A_0, угол наклона которого $\varphi(t)$ со временем изменяется («качания» вектора, рис. 3.6б). Примеры амплитудно- и частотно-модулированных колебаний приведены на рис. 3.7а и б. Амплитудно-модулированное колебание имеет вид $f(t) = [A_0 + a(t)]\, e^{i\omega_0 t}$, а частотно-модулированное колебание есть $f(t) = A_0 e^{-i[\omega_0 t - \varphi(t)]}$, где $\varphi(t) = \varphi_0 + \Delta\varphi(t)$.

В общем случае, если одновременно осуществляются оба вида модуляции, модулированное колебание имеет вид

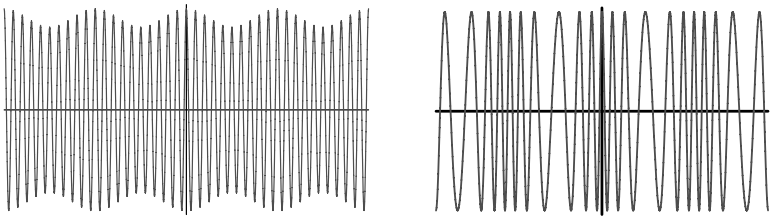
$$f(t) = [A_0 + a(t)]\, e^{-i[\omega_0 t - \varphi(t)]},$$

т. е. может быть записано в виде произведения *модулирующей* функции

$$m(t) = \left[1 + \frac{a(t)}{A_0}\right] e^{i\varphi(t)}$$

на несущее (*модулируемое*) колебание $f_0(t) = A_0 e^{-i\omega_0 t}$:

$$f(t) = m(t) \cdot f_0(t).$$



а) б)

Рис. 3.7

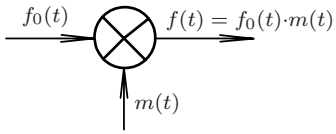


Рис. 3.8

Операция модуляции — умножение модулирующего сигнала (в котором и содержится передаваемое сообщение) на несущее колебание — изображается с помощью блок-схемы рис. 3.8. Модулирующее колебание можно создать, например, с помощью микрофона, преобразующего звуковые колебания (речь диктора или музыка) в электрические, которые затем поступают на устройство, называемое *модулятором*. Это устройство и осуществляет операцию перемножения сигналов $f_0(t)$ и $m(t)$.

Теперь обратимся к оптике, к вопросам передачи информации световой волной. При этом мы подразумеваем, что информация задаётся не в виде временно́го хода процесса (временна́я модуляция световых колебаний — это совершенно другая проблема, которой мы здесь не касаемся), а в виде *пространственного* распределения амплитуд и фаз колебаний в световой волне в некоторой плоскости $z = \text{const}$, т. е. сообщение содержится в некоторой комплексной функции $m(x,y) = a(x,y)e^{i\varphi(x,y)}$.

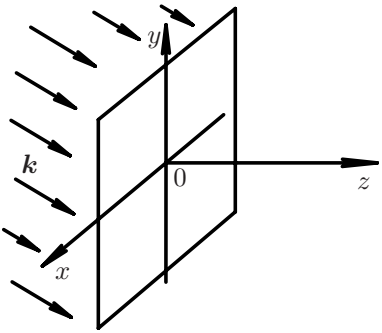


Рис. 3.9

Каким образом можно получить световое поле, которое содержало бы заданное сообщение?

Представим себе достаточно тонкую пластинку — транспарант, поглощательная способность которой меняется от точки к точке по закону $a(x,y)$ ((x,y) — координаты в плоскости транспаранта), а показатель преломления (или толщина) — по закону $n(x,y)$ (или $\Delta(x,y)$). Пусть такая пластинка освещается плоской волной с комплексной амплитудой $f_0(x,y) = A_0 e^{i(u_0 x + v_0 y)}$, ($A_0 = a_0 e^{i\varphi_0}$), где (u_0,v_0) — проекции волнового вектора на оси x и y: $u_0 = k\sin\alpha$, $v_0 = k\cos\alpha$ (рис. 3.9). Эта плоская волна — аналог несущего (модулируемого) колебания в радиотехнике. Что произойдёт с волной в плоскости выхода из пластинки $z = 0_+$? Переменная поглощательная способность транспаранта приводит к тому, что амплитуда волны на выходе будет не константой a_0, а некоторой функцией координат $a(x,y)$. Переменный показатель преломления или толщина приводят к различному в разных точках набегу фазы, так что и фронт прошедшей волны не будет плоским: он будет «деформирован» по некоторому закону $\varphi(x,y)$. Таким образом, комплексная амплитуда волны на выходе (в плоскости $z = 0_+$) будет иметь вид

$$f(x,y) = a(x,y)e^{i(u_0 x + v_0 y + \varphi(x,y))},$$

или

$$f(x,y) = t(x,y) \cdot f_0(x,y), \tag{3.20}$$

где $t(x,y)$ — «*комплексная пропускаемость*» транспаранта — *полный аналог модулирующего колебания* в радиотехнике. Вызванные поглощением в пластинке (абсорбцией) изменения амплитуды $a(x,y)$ аналогичны амплитудной модуляции $a(t)$, а изменения фазы $\varphi(x,y)$, связанные с вариациями показателя преломления (или толщины), аналогичны фазовой модуляции $\varphi(t)$. Ясно, что если поглощательная способность пластинки постоянна, то имеем чисто фазовую модуляцию (фазовая структура), а при постоянном n (или толщине) — чисто амплитудную модуляцию. Мы пока оставляем в стороне вопрос о том, как практически изготовить пластинку с заданной комплексной пропускаемостью. Конечно, пространственная модуляция в оптике не обязательно осуществляется с помощью тонких амплитудно-фазовых транспарантов; когда произвольный предмет освещается световой волной, то отражённая от предмета волна оказывается «промодулированной». Она содержит информацию как о поглощательной способности разных участков предмета, так и о его рельефе. Модулятором в данном случае является предмет, рассеивающий падающую волну.

Модулированная волна $f(x,y)$ является, как следует из (3.20), произведением несущей волны $f_0(x,y)$ и модулирующей функции $t(x,y)$. Так же, как и в радиотехнике, операцию модуляции можно изобра-

зить с помощью блок-схемы (рис. 3.10). Конечно, модулируемая волна не обязательно должна быть плоской (так же, как несущее колебание не обязательно гармоническим). Равенство (3.20) остаётся в силе для произвольной волны с комплексной амплитудой $f_0(x,y)$.

Введение комплексной пропускаемости позволяет решать довольно широкий круг разнообразных дифракционных задач, поскольку с помощью равенства (3.20) определяются граничные условия (поле в плоскости $z = 0_+$) для произвольных тонких экранов, не обязательно обладающих функцией пропускания $p_s(x,y)$; непрозрачный экран с отверстием (дифракцию на котором мы изучали выше) является лишь одним частным случаем.

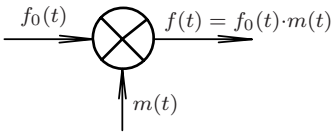


Рис. 3.10

От равенства (3.20), связывающего поле на выходе пространственного модулятора с полем на входе, можно перейти к связи между пространственными спектрами. Преобразуя обе части равенства (3.20) по Фурье, получаем

$$F(u,v) = F_0(u,v) \otimes\otimes T(u,v), \tag{3.21}$$

где $F(u,v)$, $F_0(u,v)$ и $T(u,v)$ — фурье-образы функций $f(x,y)$, $f_0(x,y)$ и $t(x,y)$ (это — обратная теорема свёртки, которую легко докажет читатель).

Приведём примеры амплитудных и фазовых транспарантов, часто применяемых в оптике.

1. *Дифракционная решётка* имеет функцию пропускания

$$t(x) = \sum_{n=-N}^{N} p_b(x - nd),$$

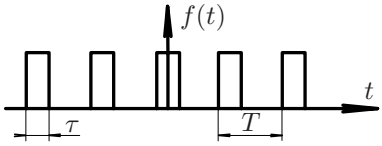


Рис. 3.11

изображённую на рис. 3.11 (d — период решётки, $2b$ — ширина щели, $2N + 1$ — число штрихов).

2. *Амплитудная синусоидальная решётка* с периодом $L = \frac{2\pi}{\Omega}$ и глубиной модуляции α: $t(x) = 1 + \alpha \cos \Omega x$.

3. *Фазовая синусоидальная решётка*: $t(x) = e^{i\alpha \cos \Omega x}$. (Это пример чисто фазовой структуры: амплитуда волны, прошедшей через решётку, не меняется, а фаза изменяется по синусоидальному закону.)

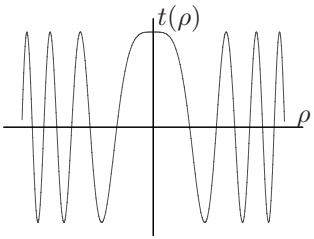


Рис. 3.12

4. *Зонная решётка Френеля*:

$$t(\rho) = 1 + \cos k\sqrt{z_0^2 + \rho^2},$$

где $\rho^2 = x^2 + y^2$, $z_0 = \text{const}$.

5. *Функция пропускания линзы.*

Равенство (3.20) можно использовать для описания работы *тонкой* линзы, если $f_0(x,y)$ и $f(x,y)$ — поля в плоскостях, примыкающих к линзе слева и справа (рис. 3.13). Действительно, мы полагаем, что луч света, входящий в линзу в точке с координатами (x,y), выходит из линзы в точке с теми же координатами (на том же расстоянии от оси), другими словами, пройдя через линзу, луч света не успевает заметно отклониться от оси — именно поэтому поле на выходе из линзы в точке (x,y) связано с полем на входе *в той же точке* (x,y). Функция $t(x,y)$ зависит от набега фазы при распространении волны от плоскости $z = 0_-$ до плоскости $z = 0_+$; поскольку поглощением света при прохождении линзы мы пренебрегаем, линза является чисто *фазовым модулятором*:

$$t(x,y) = e^{i\varphi(x,y)}.$$

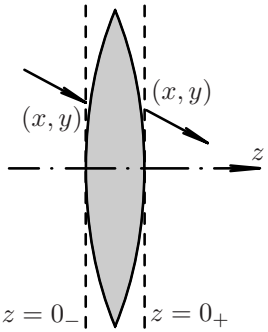


Рис. 3.13

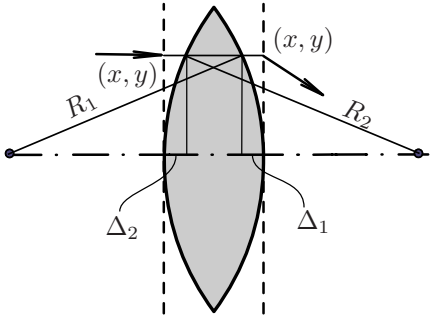


Рис. 3.14

Определим набег фазы для луча, входящего в линзу в точке с координатами (x,y). Из рис. 3.14 следует, что оптическая разность хода между лучом, проходящим через линзу в точке (x,y), и лучом, проходящим через центр линзы, есть

$$\Delta(x,y) = (1-n)(\Delta_1 + \Delta_2),$$

где

$$\Delta_1 = R_1 - \sqrt{R_1^2 - (x^2+y^2)},$$

$$\Delta_2 = R_2 - \sqrt{R_2^2 - (x^2+y^2)}.$$

Ограничимся лучами, близкими к оптической оси, заменив сферическую форму поверхностей линзы параболической:

$$\sqrt{R_1^2 - (x^2+y^2)} \approx R_1 - \frac{x^2+y^2}{2R_1}; \qquad \sqrt{R_2^2 - (x^2+y^2)} \approx R_2 - \frac{x^2+y^2}{2R_2}.$$

Приходим к равенству

$$\Delta(x,y) = -(n-1)\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\frac{x^2+y^2}{2} = -\frac{1}{2f}(x^2+y^2),$$

где $f = \left[(n-1)\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\right]^{-1}$ — фокусное расстояние. Найденная нами разность хода $\Delta(x,y)$ определяет относительный набег фазы волны, проходящей через линзу в произвольной точке (x,y): $\varphi(x,y) = k\Delta(x,y)$. Таким образом, функция пропускания линзы имеет вид

$$t(x,y) = e^{-\frac{k}{2f}(x^2+y^2)}. \tag{3.22}$$

Любая реальная линза имеет конечные размеры. Можно учесть этот факт, накрыв бесконечную линзу с функцией пропускания (3.22) непрозрачным экраном с отверстием (диафрагмой) диаметра $2a$, центр которого совпадает с центром линзы (рис. 3.15). Очевидно, функция пропускания задиафрагмированной линзы имеет вид

$$t(x,y) = p_a(x,y)e^{-\frac{k}{2f}(x^2+y^2)}, \tag{3.23}$$

где $p_a(x,y)$ — единично-нулевая функция:

$$p_a(x,y) = \begin{cases} 1 & \text{при } x^2+y^2 \leqslant a^2, \\ 0 & \text{при } x^2+y^2 > a^2. \end{cases}$$

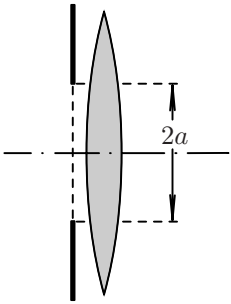


Рис. 3.15

§ 3.5. Некоторые важные задачи сложения гармонических колебаний

Рассмотренные ниже задачи сложения гармонических колебаний необходимы как для дальнейшего углублённого изучения общих вопросов дифракции, так и для решения конкретных дифракционных задач.

1. *Сумма N колебаний, фазы которых составляют арифметическую прогрессию*:

$$S = \sum_{n=1}^{N} a_0 e^{i(n-1)\delta} = a_0 + a_0 e^{i\delta} + a_0 e^{i2\delta} + \ldots \tag{3.24}$$

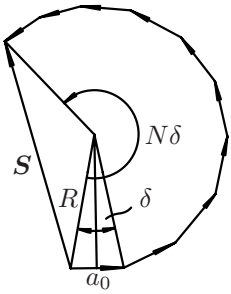


Рис. 3.16

Построим векторную диаграмму (рис. 3.16). Первое слагаемое в (3.24) изображается вектором длины a_0, направленным вдоль действительной оси; второе слагаемое $a_0 e^{i\delta}$ — вектор той же длины, составляющий угол δ с действительной осью; третий вектор составляет с действительной осью угол 2δ (и угол δ со вторым вектором) и т. д. Суммарное колебание — вектор $\boldsymbol{S}$, проведённый из начала 1-го вектора в конец последнего N-го вектора. Складываемые векторы образуют часть равностороннего многоугольника, вписанного в окружность, радиус которой равен $R = \frac{a_0}{2\sin\delta/2}$.

Из векторной диаграммы легко найти модуль вектора $\boldsymbol{S}$:

$$|\boldsymbol{S}| = 2R\left|\sin\frac{2\pi - N\delta}{2}\right| = 2R\left|\sin\frac{N\delta}{2}\right| = a_0\left|\frac{\sin\frac{N\delta}{2}}{\sin\frac{\delta}{2}}\right|. \tag{3.25}$$

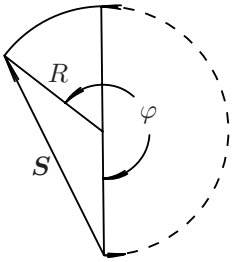


Рис. 3.17

2. *Спираль Френеля* — векторная диаграмма для вычисления интеграла

$$S = \int_0^{x_0} e^{i\beta x}\, dx.$$

Разобьем весь отрезок интегрирования [0,x_0] на N малых интервалов Δx ($x_0 = N\Delta x$). Начало n-го интервала находится в точке $x_n = n\Delta x$. Искомый интеграл приближённо равен сумме N векторов $\Delta x e^{i\beta n\Delta x}$, фазы которых образуют арифметическую прогрессию. Радиус окружности R, часть которой (при $\Delta x \to 0$) составляет сумма $\sum e^{i\beta n\Delta x}\cdot\Delta x$, находится так: при $\beta n\Delta x = \pi$ n-й вектор составляет угол π с действительной осью и, следовательно, находится на верхнем конце вертикального диаметра (рис. 3.17); суммарная длина цепочки векторов $n\Delta x$ равна длине полуокружности πR: $n\Delta x = \pi R$, откуда имеем $R = 1/\beta$. Так как дуга

окружности, стягиваемая суммарным вектором $\boldsymbol{S}$, имеет длину x_0, то угол φ равен: $\varphi = x_0/R$; окончательно получаем

$$S = \frac{2}{\beta}\left|\sin\frac{x_0\beta}{2}\right| e^{i\frac{x_0\beta}{2}}. \tag{3.26}$$

Как видно из векторной диаграммы (и из формулы (3.26)), величина интеграла является периодической функцией верхнего предела, изменяясь от максимального значения (равного $2/\beta$, когда $x_0 = \frac{\pi}{\beta}(2n \pm 1)$) до нуля при $x_0 = \frac{2\pi n}{\beta}$. Ясно, что при стремлении верхнего предела к бесконечности величина исследуемого интеграла не стремится к определённому пределу: конец суммарного вектора с ростом x_0 будет бесконечно двигаться по окружности радиуса $1/\beta$.

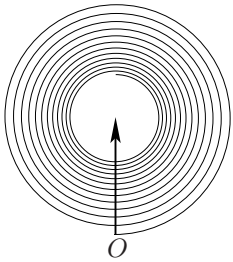


Рис. 3.18

Представим себе теперь, что длина каждого последующего элементарного вектора в цепочке векторов $\Delta x\, e^{i\beta n\Delta x}$ меньше длины предыдущего вектора; пусть, например, β является комплексным числом $\beta = \beta' + i\beta''$, где $\beta'' > 0$. Тогда длина n-го вектора есть $\Delta x\, e^{-\beta'' n\Delta x}$, т. е. с ростом n длины элементарных векторов экспоненциально уменьшаются. Ясно, что если имеется сколь угодно малое затухание амплитуд складываемых колебаний (т. е. имеется сколь угодно малое $\beta'' > 0$), то цепочка векторов $e^{i\beta n\Delta x}\Delta x$ при $\Delta x \to 0$ образует уже не дугу окружности, а скручивающуюся спираль, фокус которой (точка, в которую эта спираль скручивается) находится в центре первоначальной окружности (рис. 3.10). Значение интеграла $\int\limits_0^\infty e^{i\beta x}\,dx$ определяется тогда вектором, проведённым из точки 0 в фокус спирали, и, следовательно,

$$\int\limits_0^\infty e^{i\beta x}\,dx = \frac{i}{\beta} \tag{3.27}$$

(множитель $i = e^{i\frac{\pi}{2}}$ показывает, что суммарный вектор составляет угол $\pi/2$ с действительной осью).

Каждый раз, когда какая-либо физическая задача приводит нас к интегралу вида (3.27), физически правильное его значение получается, если предположить наличие бесконечно малого затухания $\beta'' > 0$ и затем в полученном результате перейти к пределу при $\beta'' \to 0$, в результате мы получаем равенство (3.27). Изображённая на рис. 3.18 векторная диаграмма называется *спиралью Френеля*.

3. *Спираль Корню* — векторная диаграмма для вычисления интеграла

$$S = \int\limits_0^{x_0} e^{i\beta x^2}\, dx. \tag{3.28}$$

Разбив отрезок интегрирования на малые интервалы Δx (так что $x_n = n\Delta x$), заменим интеграл (3.28) суммой

$$S \approx \sum_{n=0}^{N} e^{i\beta n^2 (\Delta x)^2}\, \Delta x, \qquad \left(N = \frac{x_0}{\Delta x}\right).$$

В отличие от предыдущего примера, здесь речь идёт о сложении колебаний, фазы которых изменяются по квадратичному закону (пропорционально n^2). Если в предыдущем примере угол между двумя последовательными векторами в цепочке остаётся постоянной величиной ($\Delta\varphi = \beta(n+1)\Delta x - \beta n\Delta x = \beta\Delta x$), то в данном примере этот угол с ростом n растёт линейно:

$$\delta\varphi_n = \beta[\delta x(n+1)]^2 - \beta[\delta x \cdot n]^2 = \beta(2n+1)(\delta x)^2.$$

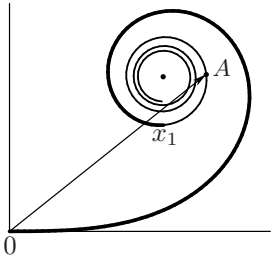


Рис. 3.19

Таким образом, закручивание цепочки векторов происходит здесь неравномерно: сначала (пока $\varphi_n = \beta(\Delta x)^2 n^2 \ll \pi$) кривая очень полога и векторы $S_n = \Delta x e^{i\beta(\Delta x)^2 n^2}$ выстраиваются (почти) в прямую линию — (элементарные колебания складываются почти синфазно, рис. 3.19). Затем при увеличении n угол между двумя последовательными векторами становится весьма заметным — кривая начинает быстро скручиваться. Длина дуги первого витка (в пределе при $\Delta x \to 0$), отмеченного на рис. 3.19 жирной линией, определяется условием $\beta x^2 = 2\pi$, откуда

$$x_1 = \sqrt{\frac{2\pi}{\beta}}. \tag{3.29}$$

Продолжая дальше прибавлять элементарные векторы, будем проходить всё более мелкие витки скручивающейся спирали. Конечно, пока речь идёт о дискретной сумме колебаний, получается не дуга, а скручивающаяся ломаная, которая переходит в плавную кривую при $\Delta x \to 0$, при этом сумма переходит в интеграл (3.28).

Очевидно, значение интеграла определяется вектором, проведённым из точки 0 в точку A, положение которой определяется условием: длина дуги, которую стягивает суммарный вектор, равна x_0.

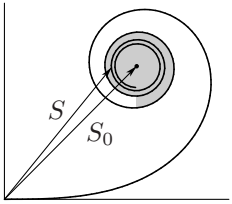


Рис. 3.20

С ростом x_0 конец суммарного вектора перемещается по виткам скручивающейся спирали и при $x_0 \to \infty$ попадает в точку, называемую *фокусом спирали*.

Векторная диаграмма позволяет понять, как изменяется значение интеграла (3.28) при изменении верхнего предела. Пока конец суммарного вектора перемещается по первому витку, т. е. x_0 изменяется в пределах: $0 < x_0 < x_1 = \sqrt{2\pi/\beta}$, изменение x_0 приводит к заметному изменению интеграла (длины результирующего вектора на диаграмме). Если же $x_0 > \sqrt{2\pi/\beta}$, т. е. конец суммарного вектора перешёл на второй, третий и т. д. витки спирали (заштрихованная область на рис. 3.20), то дальнейшее увеличение верхнего предела уже не может привести к заметному изменению интеграла (3.28), так как конец результирующего вектора уже не выходит из небольшой заштрихованной области, окружающей фокус спирали. Результирующий вектор S мало отличается при этом от вектора S_0, проведённого в фокус спирали, и мы имеем оценку:

$$\int_0^{x_0} e^{i\beta x^2}\,dx \simeq \int_0^{\infty} e^{i\beta x^2}\,dx \qquad \text{при} \qquad x_0 > \sqrt{\frac{2\pi}{\beta}}.$$

Интеграл $S_0 = \int\limits_0^{\infty} e^{i\beta x^2}\,dx$ легко вычисляется следующим образом. Вычислим сначала

$$S_0^2 = \iint\limits_0^{\infty} e^{i\beta(x^2+y^2)}\,dxdy.$$

Переходя к полярным координатам

$$\begin{cases} x = r\cos\theta, \\ y = r\sin\theta, \end{cases} \qquad dxdy = rdrd\theta,$$

имеем

$$S_0^2 = \int_0^{\infty}\int_0^{\pi/2} e^{i\beta r^2} r\,drd\theta = \frac{\pi}{4}\int_0^{\infty} e^{i\beta\xi}\,d\xi \quad (\xi = r^2).$$

Используя (3.27), получаем $S_0^2 = \pi i/\beta$, и, следовательно,

$$S_0 = \int_0^\infty e^{i\beta x^2}\,dx = \frac{1}{2}\sqrt{\frac{\pi}{\beta}}e^{i\frac{\pi}{4}}. \tag{3.30}$$

Таким образом, вектор, проведённый из начала координат в фокус спирали, имеет длину $\frac{1}{2}\sqrt{\pi/\beta}$ и составляет с действительной осью угол $\pi/4$.

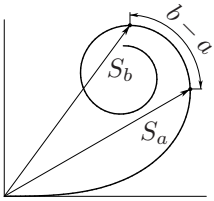


Рис. 3.21

Диаграмма, изображённая на рис. 3.21, позволяет графически определить как значение интеграла (3.28), так и величину

$$S_{ab} = \int_a^b e^{i\beta x^2}\,dx. \tag{3.31}$$

Действительно, вектор, определяющий S_{ab}, есть разность двух векторов: $S_b = \int_0^b e^{i\beta x^2}\,dx$ и $S_a = \int_0^a e^{i\beta x^2}\,dx$ (рис. 3.21). (Очевидно, вектор S_{ab} стягивает дугу, длина которой равна $|b-a|$).

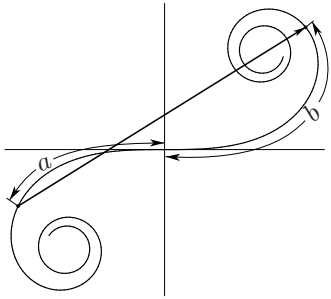


Рис. 3.22

Проведённые построения легко обобщаются на случай произвольных (в том числе отрицательных) пределов интегрирования. Полная векторная диаграмма (для произвольных a и b) изображена на рис. 3.22 и называется *спиралью Корню*. Вектор S_{ab}, изображённый на рис. 3.22, соответствует случаю $a < 0$, $b > 0$. Очевидно, величина интеграла $\int_{-\infty}^{\infty} e^{i\beta x^2}\,dx$ определяется вектором, соединяющим фокусы спирали

$$\int_{-\infty}^{\infty} e^{i\beta x^2}\,dx = \sqrt{\frac{\pi}{\beta}}e^{i\frac{\pi}{4}}. \tag{3.32}$$

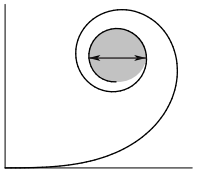


Рис. 3.23

Если $a < -x_1$ и $b > x_1$, то концы вектора S_{ab} лежат в небольших окрестностях, окружающих фокусы, и значение интеграла мало меняется с изменени-

ем пределов интегрирования, поэтому имеет место следующая оценка:

$$\int_a^b e^{i\beta x^2}\,dx \simeq \sqrt{\frac{\pi}{\beta}}e^{i\frac{\pi}{4}} \qquad \text{при} \qquad a < -\sqrt{\frac{2\pi}{\beta}},\quad b > \sqrt{\frac{2\pi}{\beta}}.$$

Наконец, следует отметить ещё одну характерную особенность спирали Корню. Если оба предела интегрирования a и b лежат в окрестности одного и того же фокуса $a > x_1$, $b > x_1$, то модуль интеграла (3.31) не превышает диаметра заштрихованного круга. Этот случай изображён на рис. 3.23.

4. *Представление сферической волны в виде суперпозиции плоских волн* (*формула Вейля*).

Пусть в начале координат плоскости $z = 0$ находится точечный источник света, излучающий сферическую волну. Поле этой волны в точке (x,y,z), расположенной в плоскости наблюдения $z = \text{const}$, есть

$$g(x,y,z) = \frac{e^{ikR}}{R} = \frac{e^{ik\sqrt{x^2+y^2+z^2}}}{\sqrt{x^2+y^2+z^2}}. \tag{3.33}$$

Наша задача состоит в том, чтобы представить это поле в виде суперпозиции плоских волн (как бегущих, так и неоднородных), другими словами, необходимо найти преобразование Фурье функции (3.33). Решим эту задачу следующим образом: сначала определим спектр $F(u,v)$ поля:

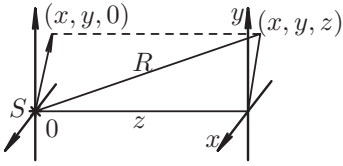


Рис. 3.24

$$f(x,y) = g(x,y,0) = \frac{e^{ikr}}{r} = \frac{e^{ik\sqrt{x^2+y^2}}}{\sqrt{x^2+y^2}},$$

создаваемого точечным источником S в плоскости $z = 0$. Затем, используя общую связь (3.5) между спектрами в двух плоскостях, отстоящих на расстоянии z, найдём искомый спектр $G(u,v)$ функции (3.33).

Спектр $F(u,v)$, определяемый равенством

$$F(u,v) = \iint \frac{e^{ik\sqrt{x^2+y^2}}}{\sqrt{x^2+y^2}} e^{-i(ux+vy)}\,dxdy, \tag{3.34}$$

вычисляется следующим образом. Перейдём в (3.34) к полярным координатам

$$\begin{cases} x = r\cos\varphi, \\ y = r\sin\varphi, \end{cases} \qquad \begin{cases} u = \rho\cos\theta, \\ v = \rho\sin\theta. \end{cases}$$

Подставляя в (3.34), получаем

$$F(\rho,\theta) = \iint \frac{e^{ikr}}{r} e^{-r\rho\cos(\varphi-\theta)}\, r dr d\varphi = \int\limits_0^{2\pi} d\varphi \int\limits_0^{\infty} e^{i[k-\rho\cos(\varphi-\theta)]r}\, dr.$$

Внутренний интеграл имеет вид $\int\limits_0^{\infty} e^{i\alpha r}\, dr$, т. е. может быть интерпретирован как сумма колебаний, фазы которых составляют арифметическую прогрессию (§ 3.5). Его величина равна

$$\frac{i}{\alpha} = \frac{i}{k-\rho\cos(\varphi-\theta)},$$

следовательно,

$$F(\rho,\theta) = i \int\limits_0^{2\pi} \frac{d\varphi}{k-\rho\cos(\varphi-\theta)} = \frac{2\pi i}{\sqrt{k^2-\rho^2}}.$$

Согласно общему соотношению (3.5), спектр $G(u,v)$ в плоскости наблюдения $z = \text{const}$ есть

$$G(u,v) = \frac{2\pi i}{\sqrt{k^2-u^2-v^2}} e^{i\sqrt{k^2-u^2-v^2}\, z}. \tag{3.35}$$

Получено важное равенство, дающее разложение сферической волны в интеграл Фурье:

$$\frac{e^{ikR}}{R} = \frac{i}{2\pi} \iint \frac{e^{i\sqrt{k^2-u^2-v^2}\, z}}{\sqrt{k^2-u^2-v^2}} e^{i(ux+vy)}\, du dv. \tag{3.36}$$

В дальнейшем мы воспользуемся соотношением, которое получается дифференцированием обеих частей (3.36) по z:

$$-\frac{1}{2\pi}\frac{d}{dz}\frac{e^{ikR}}{R} = \frac{1}{4\pi^2} \iint e^{i\sqrt{k^2-u^2-v^2}\, z} e^{i(ux+vy)}\, du dv. \tag{3.37}$$

§ 3.6. Принцип Гюйгенса–Френеля

Для понимания многих явлений оптики чрезвычайно плодотворными оказываются аналогии с линейными колебательными системами. В частности, задачу распространения оптического сигнала от плоскости

$z = 0$ до плоскости наблюдения $z = \text{const} > 0$ можно рассматривать как задачу линейной фильтрации: поле во «входной» плоскости $z = 0$ — это входной сигнал фильтра; плоскость наблюдения $z = \text{const}$ можно назвать «выходной» плоскостью, а поле в ней — выходным сигналом фильтра. Входной и выходной сигналы — это функции пространственных координат $f(x,y)$ и $g(x,y)$ — комплексные амплитуды в плоскости $z = 0$ и $z = \text{const}$, здесь, таким образом, речь идёт о пространственном фильтре.

В главе I мы говорили о возможности двух подходов к изучению линейных фильтров: *спектрального*, результатом которого является связь между спектрами входного и выходного сигналов:

$$G(\omega) = F(\omega) \cdot H(\omega), \tag{3.38}$$

и *временно́го*, результатом которого является связь между самими сигналами:

$$g(t) = \int f(\tau) h(t - \tau)\, d\tau, \tag{3.39}$$

причём было показано, что частотная характеристика $H(\omega)$ (которая определяет свойства фильтра при спектральном подходе) и импульсный отклик $h(t)$ (определяющий свойства фильтра при временно́м подходе) связаны преобразованием Фурье:

$$h(t) = \frac{1}{2\pi} \int H(\omega) e^{i\omega t}\, d\omega. \tag{3.40}$$

С точки зрения математики равенства (3.38) и (3.39) являются просто теоремой фурье-анализа, утверждающей, что если преобразования Фурье трёх функций связаны равенством (3.38), то сами функции связаны соотношением (3.39) (теорема свёртки). Эта теорема полностью переносится на случай функций двух переменных, так что из спектрального равенства

$$G(u,v) = F(u,v)H(u,v) \tag{3.41}$$

следует

$$g(x,y) = \iint f(\xi,\eta) h(x - \xi, y - \eta)\, d\xi d\eta, \tag{3.42}$$

где функции $h(x,y)$ и $H(u,v)$ связаны преобразованием Фурье:

$$h(x,y) = \frac{1}{4\pi^2} \iint H(u,v) e^{i(ux+vy)}\, du dv \tag{3.43}$$

(точно такая же связь между парами $f(x,y)$ и $F(u,v)$, $g(x,y)$ и $G(u,v)$).

Решение дифракционной задачи, выраженное равенством (3.8):

$$G(u,v) = F(u,v)e^{i\sqrt{k^2-u^2-v^2}\,z},$$

есть результат спектрального подхода к изучению свойств пространственного фильтра — свободного пространства, разделяющего «входную» ($z = 0$) и «выходную» ($z = \text{const}$) плоскости, причём множитель $e^{i\sqrt{k^2-u^2-v^2}\,z}$ имеет смысл частотной характеристики.

Ясно, что формулу (3.42) следует рассматривать как аналог временно́го подхода. Эта формула связывает комплексные амплитуды полей во входной плоскости $f(x,y)$ и в выходной плоскости $g(x,y)$, причём функцию $h(x,y)$ можно назвать импульсным откликом пространственного фильтра. Поскольку частотная характеристика $H(u,v) = e^{i\sqrt{k^2-u^2-v^2}\,z}$ нам известна, то с помощью (3.43) можно найти конкретный вид импульсного отклика:

$$h(x,y) = \frac{1}{4\pi^2}\iint e^{i\sqrt{k^2-u^2-v^2}\,z}e^{i(ux+vy)}\,dudv. \tag{3.44}$$

Сравнивая (3.44) и (3.37), получаем

$$h(x,y) = -\frac{1}{2\pi}\frac{d}{dz}\frac{e^{ikR}}{R}. \tag{3.45}$$

Подставляя найденное нами выражение для импульсного отклика в (3.42), можно искомое дифракционное поле записать в виде

$$g(x,y) = -\frac{1}{2\pi}\iint f(\xi,\eta)\frac{d}{dz}\frac{e^{ikR}}{R}\,d\xi d\eta \tag{3.46}$$

(здесь под R следует понимать расстояние между точкой наблюдения с координатами (x,y,z) и «переменной» точкой интегрирования $(\xi,\eta,0)$: $R = \sqrt{(x-\xi)^2+(y-\eta)^2+z^2}$).

Соотношение (3.46) называется *формулой Грина*. В условиях оптического эксперимента, в силу малости λ, почти всегда оказывается выполнено сильное неравенство $R \gg \lambda$ (при этом говорят, что точка наблюдения находится в «волновой зоне»). В волновой зоне выражение (3.45) для импульсного отклика упрощается; мы имеем

$$\frac{d}{dz}\frac{e^{ikR}}{R} = \frac{d}{dR}\left(\frac{e^{ikR}}{R}\right)\cos(\widehat{Rz}) = \\ = \left(ik - \frac{1}{R}\right)\frac{e^{ikR}}{R}\cos(\widehat{Rz}) \approx ik\frac{e^{ikR}}{R}\cos(\widehat{Rz}), \tag{3.47}$$

при $R \gg \lambda$. Следовательно,

$$h(x,y) = \frac{1}{i\lambda}\frac{e^{ikR}}{R}\cos(\widehat{Rz}). \tag{3.48}$$

Формула Грина (3.46) приобретает вид

$$g(x,y) = \frac{1}{i\lambda}\iint f(\xi,\eta)\frac{e^{ikR}}{R}\cos(\widehat{Rz})\,d\xi d\eta, \tag{3.49}$$

который допускает наглядную физическую интерпретацию, называемую принципом Гюйгенса–Френеля: дифрагированное поле $g(x,y)$ представляет собой суперпозицию сферических волн e^{ikR}/R, излучаемых точечными источниками, расположенными в плоскости $z = 0$ (плоскость (ξ, η)); каждый точечный источник (малая площадка $d\xi\, d\eta$, расположенная в точке с координатами (ξ, η)) излучает с той амплитудой и начальной фазой, которые «сообщаются» ему падающим полем $f(\xi,\eta)$. Амплитуда излучения зависит также от угла θ между осью z и направлением на точку наблюдения. Этот угол определяет видимую из точки наблюдения площадь элементарной излучающей площадки $d\xi d\eta$.

Принцип Гюйгенса–Френеля (3.49) был сформулирован задолго до того, как было получено его строгое математическое обоснование, на основании интуитивных соображений о существовании неких «вторичных» источников сферических волн, — пример, характерный для оптики, где часто догадка и интуиция опережали строгое математическое обоснование. Дело, по-видимому, в том, что интуиции помогает в этом случае природа, подарившая человеку зрение — возможность непосредственной регистрации (без каких-либо специальных устройств) световой волны.

Дальнейшее изучение дифракционных явлений будет основано или на спектральном равенстве (3.5) или на принципе Гюйгенса–Френеля (3.49).

Тот или иной характер дифракционных явлений определяется величиной *трёх основных параметров*: длиной волны λ (или волновым числом $k = 2\pi/\lambda$), расстоянием от препятствия до плоскости наблюдения z и характерным линейным размером препятствия D (например, размером отверстия в непрозрачном экране). Введём, следуя Г.С. Горелику [12], *волновой параметр* — безразмерную комбинацию величин λ, z и D:

$$P = \frac{\sqrt{\lambda z}}{D}.$$

Как показывает дальнейший анализ, дифракционные явления имеют существенно различный характер в трёх основных областях изменения волнового параметра P:

1) область геометрической оптики: $P \ll 1$;
2) область дифракции Френеля: $P \sim 1$;
3) область дифракции Фраунгофера: $P \gg 1$.

§ 3.7. Область геометрической оптики

Рассмотрим простую задачу: отверстие в непрозрачном экране освещается плоской, нормально падающей волной (рис. 3.25). При этом в плоскости наблюдения П, расположенной не слишком далеко от экрана с отверстием, можно наблюдать светлое пятно, тождественно повторяющее форму отверстия, другими словами, пятно получается геометрическим проецированием отверстия на плоскость наблюдения. Поставим вопрос: при каких условиях поле в плоскости $z = \text{const}$ оказывается тождественным полю в плоскости $z = 0_+$, т. е. при каких условиях «работает» геометрическая оптика? Обратимся к спектральному равенству (3.5) и разложим радикал в фазовом множителе в степенной ряд:

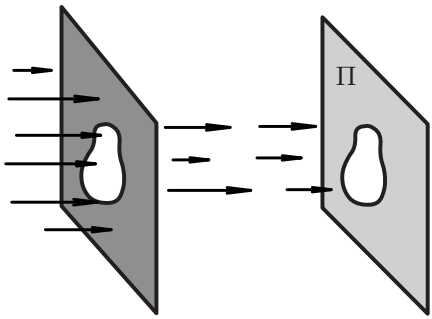


Рис. 3.25

$$\sqrt{k^2 - u^2 - v^2}\, z = kz \left[1 - \frac{u^2 + v^2}{2k^2} + \frac{(u^2 + v^2)^2}{8k^4} - \dots \right]. \qquad (3.50)$$

Очевидно, при $k \to \infty$ (или при $\lambda \to 0$) функция $H(u,v) = e^{i\sqrt{k^2-u^2-v^2}\, z}$ стремится к e^{ikz} для всех частот u, v, другими словами, набег фазы при распространении волны до плоскости $z = \text{const}$ одинаков для всех плоских волн, составляющих суммарную волну, и равен kz. Следовательно, фазовые соотношения между плоскими волнами в плоскости

наблюдения $z = \text{const}$ такие же, как и в плоскости $z = 0$. Очевидно, что интерференция плоских волн даст в плоскости наблюдения тот же результат, что и на границе $z = 0$. Математически этот факт выражается равенством: $G(u,v) = e^{ikz}F(u,v)$ и, следовательно, $g(x,y) = e^{ikz}f(x,y)$ при $k \to \infty$, т. е. поля отличаются только постоянным (не зависящим от x, y) множителем. Итак, *переход от волновой оптики к геометрической соответствует предельному переходу* $k \to \infty$ (*или* $\lambda \to 0$).

В реальных оптических экспериментах длина световой волны хотя и очень мала ($\sim 10^{-5}$ см), но всё же отлична от нуля. Поставим вопрос: при каких экспериментальных условиях распространение света может быть описано законами геометрической оптики? Снова обратимся к (3.50). Очевидно, набег фазы для всех плоских волн, образующих поле в плоскости $z = 0_+$, одинаков и равен kz, если можно отбросить все члены разложения (3.50), кроме первого. В частности, достаточно потребовать: $\frac{u^2+v^2}{2k}z \ll \pi$. Последнее неравенство должно выполняться для всех частот (u,v), при которых спектр $F(u,v)$ отличен от нуля. Эту область частот оценим из соотношения неопределённостей: $(u^2+v^2)D^2 \simeq 4\pi^2$, где D — протяжённость функции $f(x,y)$ (определяемая, например, размером отверстия в непрозрачном экране). Теперь написанное выше неравенство можно переписать в виде

$$P^2 = \frac{\lambda z}{D^2} \ll 1. \tag{3.51}$$

Итак, область геометрической оптики — это область малых значений волнового параметра.

§ 3.8. Дифракция Френеля

Снова будем исходить из спектрального равенства:

$$G(u,v) = F(u,v) \cdot H(u,v),$$

где $H(u,v) = e^{i\sqrt{k^2-u^2-v^2}\,z}$ — частотная характеристика. Оставим в разложении (3.50) для фазы частотной характеристики два слагаемых:

$$\sqrt{k^2-u^2-v^2}\,z \approx kz - \frac{u^2+v^2}{2k}z.$$

Для того чтобы написанное приближение было справедливо, достаточно потребовать малости первого отброшенного члена:

$$\frac{(u^2+v^2)^2}{8k^3}z \ll \pi.$$

Используя соотношение неопределённостей $(u^2+v^2)D^2 \simeq 4\pi^2$, перепишем последнее неравенство в виде

$$\frac{\lambda^3 z}{D^4} \ll 1. \tag{3.52}$$

Таким образом, мы видим, что (3.52) является достаточным условием применимости следующего приближения для частотной характеристики:

$$H(u,v) \approx e^{ikz} \cdot e^{-i\frac{u^2+v^2}{2k}z}. \tag{3.53}$$

Найдём импульсный отклик $h(x,y)$, соответствующий этому приближению. Подставляя (3.53) в (3.43), получаем

$$h(x,y) = \frac{e^{ikz}}{4\pi^2} \iint e^{-i\frac{u^2+v^2}{2k}z} \cdot e^{i(ux+vy)}\, du dv.$$

Последнее выражение легко преобразовать (дополнив фазу до полного квадрата) к виду

$$h(x,y) = \frac{e^{ikz}}{4\pi^2} e^{i\frac{k}{2z}(x^2+y^2)} \left[\int e^{-i\frac{z}{2k}u'^2}\, du'\right]^2.$$

Используя (3.30), приходим к окончательной формуле для импульсного отклика, справедливой при достаточном условии (3.52):

$$h(x,y) = \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}(x^2+y^2)}. \tag{3.54}$$

Выражение для поля $g(x,y)$ в этом приближении получим, подставляя (3.54) в (3.42):

$$g(x,y) = \frac{e^{ikz}}{i\lambda z} \iint f(\xi,\eta) e^{i\frac{k}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]}\, d\xi d\eta. \tag{3.55}$$

Формула (3.55), как правило, несравненно проще для использования, чем общее выражение (3.49), но условие её применимости нас не слишком удовлетворяет, так как мы можем быть уверены в справедливости (3.52) лишь при достаточно малых z и (или) достаточно больших D. Попробуем расширить область применимости формулы (3.55); для этого подойдем к решению задачи с другой стороны. Будем исходить из принципа Гюйгенса–Френеля (3.49). Использование этого более общего выражения для решения конкретных дифракционных задач затруднительно из-за сложного характера зависимости R от переменных интегрирования ξ и η:

$$R = \sqrt{(x-\xi)^2 + (y-\eta)^2 + z^2}.$$

Воспользуемся приближением

$$R \approx z + \frac{(x-\xi)^2 + (y-\eta)^2}{2z}, \tag{3.56}$$

полученным путём разложения радикала в ряд по степеням выражения $\frac{(x-\xi)^2+(y-\eta)^2}{z^2}$, причём оставлены лишь первые два члена разложения. Важно, чтобы отброшенные члены не вносили заметную поправку в фазу kR, в частности, достаточно потребовать малости первого из отброшенных членов разложения:

$$k\frac{\left[(x-\xi)^2 + (y-\eta)^2\right]^2}{6z^3} \ll \pi.$$

Последнее неравенство должно выполняться для всех ξ, η, при которых заметно отлично от нуля граничное поле $f(\xi, \eta)$. Пусть, например, D — линейный размер отверстия в непрозрачном экране, на котором происходит дифракция. Тогда размер интересующей нас области изменения переменных ξ, η есть $|\xi^2 + \eta^2| \lesssim D^2$. Условие справедливости приближения (3.56) можно теперь записать в виде

$$\frac{D^4}{\lambda z^3} \ll 1 \tag{3.57}$$

(аналогичное условие достаточно наложить на допустимый размер области наблюдения D_H). Подставляя (3.56) в (3.49) и полагая, что $\cos(\widehat{Rz}) \approx 1$ при D,$D_H \ll z$, снова приходим к выражению (3.55).

Выпишем неравенства (3.57) и (3.52):

$$\frac{D^4}{\lambda z^3} \ll 1, \qquad \frac{\lambda^3 z}{D^4} \ll 1. \tag{3.58}$$

Ещё раз подчеркнём, что при выполнении любого из них справедливо френелевское приближение, другими словами, выражение для поля (3.55) справедливо как при больших z и малых D (работает первое из неравенств (3.58)), так и при малых z и больших D (работает второе из неравенств (3.58)).

Мы часто в дальнейшем будем опускать постоянные множители и записывать (3.55) в символическом виде:

$$g(x,y) = f(x,y) \otimes\otimes e^{i\frac{k}{2z}(x^2+y^2)} \tag{3.59}$$

(сигнал на выходе линейного фильтра является свёрткой входного сигнала с импульсным откликом).

Эквивалентную связь между спектрами запишем в виде

$$G(u,v) = F(u,v)e^{-i\frac{z}{2k}(u^2+v^2)} \tag{3.60}$$

(опущен постоянный множитель e^{ikz}, не влияющий на форму спектров).

Легко видеть, что неравенства (3.58) не налагают слишком жёстких ограничений на значение волнового параметра P, который может быть как большим, так и малым (и, конечно, допустимы значения $P \sim 1$). Область применимости выражения (3.55) чрезвычайно широка: она, в частности, охватывает и условия $P \ll 1$ и $P \gg 1$. Мы выделяем области геометрической оптики и фраунгоферовой дифракции не потому, что перестаёт «работать» формула (3.55), а потому, что в этих областях можно получить более простые соотношения.

Пример 1. Дифракция Френеля на периодических структурах. Эффект Талбота.

Замечательный эффект, о котором пойдёт речь ниже, открыт Фоксом Талботом ещё в 1836 году: плоская монохроматическая волна, дифрагируя на периодической структуре (например, эквидистантно расположенных одинаковых объектах), воспроизводит на определённых расстояниях z_m её изображение. Изображение возникает «само собой», без использования каких-либо оптических систем, поэтому эффект называют эффектом *саморепродукции* (самовоспроизведения). Положение плоскостей самовоспроизведения определяется формулой

$$z_m = \frac{2d^2}{\lambda}m, \qquad m = 1, 2, \ldots, \tag{3.61}$$

где d — период структуры, λ — длина волны излучения. Ось z перпендикулярна плоскости структуры.

Итак, пусть транспарант, функция пропускания которого $t(x) = \sum f_0(x - nd)$ периодична по координате x с периодом d, освещается слева плоской нормально падающей волной. Тогда в плоскости $z = 0_+$, примыкающей к транспаранту справа, возникает пространственно-периодическое световое поле, с комплексной амплитудой

$$f(x) = \sum f_0(x - nd) \tag{3.62}$$

(функция $f_0(x)$ описывает комплексную пропускаемость отдельной ячейки периодической структуры).

Периодическая функция $f(x)$ может быть представлена рядом Фурье (1.26)

$$: f(x) = \sum c_n e^{inu_0 x}, \qquad u_0 = \frac{2\pi}{d}, \tag{3.63}$$

каждое слагаемое которого — поле плоской волны. Амплитуда и начальная фаза волны определяются соответствующим коэффициентом

$$c_n = \frac{1}{d} \int\limits_{-\frac{d}{2}}^{\frac{d}{2}} f_0(x) e^{-inu_0 x}\, dx$$

(см. формулу (1.27)), а направление — пространственной частотой $u_n = nu_0$. Оно составляет угол $\sin\alpha_n = \frac{nu_0}{k} = n\frac{\lambda}{d}$ с осью z.

Плоские волны разных пространственных частот u_n при распространении через промежуток свободного пространства z приобретают разные набеги фаз:

$$\varphi_n = k_z z = \sqrt{k^2 - u_n^2}\, z \approx kz - \frac{z}{2k} u_n^2 = kz - \frac{z}{2k} n^2 u_0^2$$

(использовано френелевское приближение). Разность фазовых набегов двух любых плоских волн с пространственными частотами $u_n = nu_0$ и $u_l = lu_0$ равна

$$\Delta\varphi_{nl} = \frac{z}{2k}(l^2 - n^2)u_0^2,$$

или, поскольку $u_0 = \frac{2\pi}{d}$ и $k = \frac{2\pi}{\lambda}$:

$$\Delta\varphi_{nl} = \pi(l^2 - n^2) \cdot \frac{\lambda z}{d^2}. \tag{3.64}$$

Нас интересуют расстояния, для которых величины $\Delta\varphi$ кратны 2π для любых плоских волн в суперпозиции (3.63), т. е. волн с пространственными частотами $u_n = nu_0$ и $u_l = lu_0$ для любых n и l. Равенство $\Delta\varphi_{nl} = 2\pi q$, где q — целое число, справедливо, если

$$z_m = \frac{2d^2}{\lambda} m \tag{3.65}$$

(это легко проверить, подставив выражение (3.65) для z_m в (3.64)). Разность фазовых набегов, равная целому числу 2π, означает, что результат интерференции плоских волн в плоскости $z = 0$ (где эти волны образуют периодическую структуру $f(x)$) и в плоскостях, отстоящих на расстояниях z_m, оказывается одинаковым. Следовательно, в этих

плоскостях воспроизводится исходная периодическая структура. Обратим внимание на отмеченную нами особенность законов распространения световых полей: периодичность светового поля в поперечном направлении (по координате x) приводит к периодическому характеру преобразования световых полей в процессе распространения в продольном направлении — вдоль оси z. Расстояние между плоскостями самовоспроизведения равно $\frac{2d^2}{\lambda}$.

Заметим также, что эффект самовоспроизведения имеет место, если световое поле представляет собой дискретную суперпозицию плоских волн вида

$$f(x) = \sum_n c_n e^{i\sqrt{n}\,u_0 x}. \tag{3.66}$$

В этом случае $\Delta\varphi_{nl} = \frac{z}{2k}(l-n)u_0^2$ также оказывается кратной 2π на тех же расстояниях z_m.

Пример 2. Дифракция плоской волны на щели ширины $2D$ ($|\xi| \leqslant D$, $-\infty \leqslant \eta \leqslant \infty$). Дифракционная формула Френеля (3.55) принимает вид

$$g(x,y) = \frac{e^{ikz}}{z} \int\limits_{-D}^{D} e^{i\frac{k}{2z}(x-\xi)^2}\,d\xi \cdot \int\limits_{-\infty}^{\infty} e^{i\frac{k}{2z}(y-\eta)^2}\,d\eta.$$

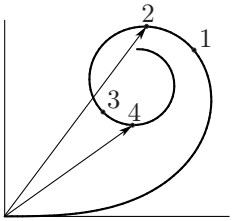


Рис. 3.26

Интеграл в бесконечных пределах вычисляется с помощью (3.30), и мы получаем

$$g(x,y) = g(x) = \frac{e^{ikz}}{\sqrt{i\lambda z}} \int\limits_{-D}^{D} e^{i\frac{k}{2z}(x-\xi)^2}\,d\xi.$$

Проследим за изменением интенсивности света на оси z ($x = 0$) при изменении ширины щели $2D$:

$$g(0) = \frac{e^{ikz}}{\sqrt{i\lambda z}} \int\limits_{-D}^{D} e^{i\frac{k}{2z}\xi^2}\,d\xi = \frac{2e^{ikz}}{\sqrt{i\lambda z}} \int\limits_{0}^{D} e^{i\frac{k}{2z}\xi^2}\,d\xi. \tag{3.67}$$

Интеграл в (3.67) оценивается с помощью спирали Корню (§ 3.5). С ростом D длина результирующего вектора на спирали растёт до тех пор, пока его конец не попадёт в точку 2 (рис. 3.26). Положение этой точки приближённо определяется условием $\frac{k}{2z}\xi^2 = \pi$ (последний элементарный вектор на диаграмме — в противофазе с первым)[1]. Длина

[1] Обычно этим условием заменяют более точное условие $\frac{k}{2z}\xi^2 = \frac{3\pi}{4}$, которое примерно соответствует положению точки 1 на спирали Корню.

дуги, стягиваемой результирующим вектором, равна интервалу интегрирования D, поэтому получаем приближённое условие максимума $\sqrt{\frac{2\pi z}{k}} = D$ или $D = \sqrt{\lambda z}$. При дальнейшем увеличении D интенсивность будет спадать до тех пор (точка 4), пока мы не придем к условию $\frac{k}{2z}\xi^2 = 2\pi$, откуда условие минимума $D = \sqrt{2\lambda z}$ (точнее, минимум соответствует точке 3, для которой $\frac{k}{2z}\xi^2 = \frac{7\pi}{4}$). Вообще, условие $D = \sqrt{n\lambda z}$ определяет максимумы при n нечётном и минимумы при n чётном.

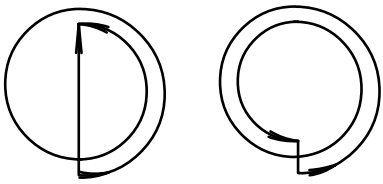
Пример 3. Дифракция плоской волны на круглом отверстии радиуса R. Будем интересоваться полем на оси отверстия, на расстоянии z от него. Мы имеем

$$g(0) = \frac{e^{ikz}}{i\lambda z} \iint\limits_{|\xi^2+\eta^2|<R^2} e^{i\frac{k}{2z}(\xi^2+\eta^2)}\, d\xi d\eta.$$

Переходя к полярным координатам и учитывая, что $d\xi d\eta = r dr d\varphi$, получаем

$$g(0) = \frac{2\pi e^{ikz}}{i\lambda z} \int\limits_0^{R^2} e^{i\frac{k}{2z}\rho}\, d\rho, \tag{3.68}$$

где $\rho = r^2$. Интеграл в (3.68) оценивается с помощью спирали Френеля (3.5). Ясно, что условие $\frac{k}{2z}R^2 = \pi n$ даёт максимум при n нечётном (последний элементарный вектор на диаграмме в противофазе с первым, рис. 3.27а) и минимум (почти нуль) при n чётном (последний вектор — в фазе с первым, рис. 3.27б).



а) б)

Рис. 3.27

§ 3.9. Дифракция Фраунгофера

Дифракция Фраунгофера — это предельный случай френелевской дифракции, соответствующей сильному неравенству $P \gg 1$, поэтому будем исходить из дифракционной формулы (3.55).

Фазовый множитель в (3.55) представим в виде

$$e^{i\frac{k}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]} = e^{i\frac{k}{2z}(x^2+y^2)} e^{i\frac{k}{2z}(\xi^2+\eta^2)} e^{-i\frac{k}{z}(x\xi+y\eta)}.$$

Область интегрирования по ξ и η ограничена размерами отверстия в непрозрачном экране $\xi^2+\eta^2 \leqslant D^2$, поэтому имеем

$$\frac{k}{2z}(\xi^2+\eta^2) \leqslant \frac{kD^2}{2z} = \pi\frac{D^2}{\lambda z} = \pi\frac{1}{P^2} \ll \pi$$

при $P \gg 1$. Таким образом, во всей области интегрирования $e^{i\frac{k}{2z}(\xi^2+\eta^2)} \approx 1$, и мы получаем

$$g(x,y) = \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}(x^2+y^2)} \iint f(\xi,\eta) e^{-i\frac{k}{z}(x\xi+y\eta)}\, d\xi d\eta. \tag{3.69}$$

Интеграл в (3.69) представляет собой преобразование Фурье функции $f(\xi,\eta)$ от аргументов $u = k\frac{x}{z}$, $v = k\frac{y}{z}$, следовательно,

$$g(x,y) = \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}(x^2+y^2)} F\left(\frac{kx}{z}, \frac{ky}{z}\right). \tag{3.70}$$

Выражение (3.70) даёт чрезвычайно простой рецепт определения поля в зоне Фраунгофера: необходимо найти преобразование Фурье $F(u,v)$ «входного» поля $f(x,y)$, подставив в качестве аргументов $u = k\frac{x}{z}$, $v = k\frac{y}{z}$. Множитель, стоящий перед функцией $F\left(k\frac{x}{z}, k\frac{y}{z}\right)$, не влияет на распределение интенсивности света по плоскости наблюдения:

$$I(x,y) = gg^* = \frac{1}{\lambda^2 z^2}\left|F\left(k\frac{x}{z}, k\frac{y}{z}\right)\right|^2. \tag{3.71}$$

Выражение (3.70) показывает, что поле в любой точке наблюдения (x,y) определяется значением спектра «входного» поля только на одной частоте $u = k\frac{x}{z}$, $v = k\frac{y}{z}$.

Каждая пространственная гармоника частоты пучком с направлением $\sin\alpha = u/k$. Нулевой пучки с частотами $u = 0$ и $z = z_0 = kD/u$. Так как ширина спектра разделения $z > D^2/\lambda$. Чем больше z, тем все пространственные гармоники (все пучки) с частотой $u = kx/z$.

Используя примеры преобразований Фурье, приведённые в § 2.6, читатель легко решит задачи дифракции Фраунгофера на щели, на прямоугольной решётке, на амплитудной синусоидальной решётке.

§ 3.10. Теорема Котельникова в оптике

Мы будем говорить о теореме Котельникова, имея в виду пространственный аналог соотношения (1.45), то есть о комплексных амплитудах световых полей — функций координат $f(x, y)$, заданных в некоторой фиксированной плоскости $z = 0$. Для сокращения выкладок будем рассматривать функции одной переменной $f(x)$.

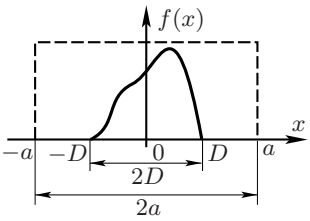


Рис. 3.28

Введём определение. *Финитная функция* — функция, отличная от нуля лишь в некоторой ограниченной области (рис. 3.28): $f(x) \equiv 0$ при $|x| \geqslant D$. В частности, согласно граничным условиям Кирхгофа (3.11), поле на выходе из отверстия размера D (в плоскости, примыкающей к отверстию) является финитной функцией. Кроме того, транспарант, функция пропускания которого содержит необходимую информацию, всегда имеет конечные размеры и освещается пучком света конечного поперечного сечения. И в этом случае поле на выходе транспаранта описывается финитной функцией.

Далее, при распространении света от границы $z = 0$ на расстояниях $z \gg \lambda$ затухание неоднородных волн (с пространственными частотами $|u| > k$) приводит к тому, что световое поле $g(x)$ на удалении от транспаранта описывается *функцией с финитным спектром* $G(u)$: $G(u) \equiv 0$ при $|u| > k$ (3.10). В реальных оптических системах конечные размеры объективов ещё сильнее ограничивают полосу пропускания *пространственных частот* $|u| \leqslant \Omega < k$, поэтому световое поле $g(x)$ в *выходной плоскости* системы описывается функцией с финитным спектром. Такое поле полностью определяется своими значениями в отсчётных точках $x_n = nx_0$, если расстояние между ними $x_0 = \frac{\pi}{u_0}$ не превосходит величины $\frac{\pi}{\Omega}$:

$$g(x) = \sum_n g\left(n\frac{\pi}{u_0}\right) \frac{\sin u_0(x - n\frac{\pi}{u_0})}{u_0(x - n\frac{\pi}{u_0})}, \tag{3.72}$$

где $u_0 \geqslant \Omega$, Ω — граничная частота спектра выходного сигнала.

Формула (3.72) является очевидным пространственным аналогом соотношения (1.45).

Примем теперь во внимание, что любую функцию $f(x)$ можно рассматривать как фурье-преобразование её собственного спектра $F(u)$, т. е. если $f(x) \leftrightarrow F(u)$, то $F(x) \leftrightarrow 2\pi f(-u)$. Другими словами, два последовательно применённых к функции $f(x)$ преобразования Фурье

приводят с точностью до инверсии к исходной функции (см. 1.29)). Поэтому естественным следствием (3.72) является так называемая *обратная теорема Котельникова*: спектр $F(u)$ финитной функции $f(x)$ ($f(x) \equiv 0$ при $|x| \geqslant D$, рис. 3.28) может быть представлен рядом

$$F(u) = \sum_n F\left(n\frac{\pi}{a}\right) \frac{\sin a(u - n\frac{\pi}{a})}{a(u - n\frac{\pi}{a})}, \tag{3.73}$$

т. е. полностью восстанавливается по своим значениям в отсчётных точках $u_n = nu_0 = n\frac{\pi}{a}$, если интервал между отсчётами $u_0 = \frac{\pi}{a}$ не превышает величины $\frac{\pi}{D}$ (т. е. $a \geqslant D$).

Применяя к обеим частям равенства (3.73) обратное преобразование Фурье, получаем следующее представление для собственно финитной функции $f(x)$:

$$f(x) = p_a(x) \sum_n F\left(n\frac{\pi}{a}\right) e^{in\frac{\pi}{a}x}. \tag{3.74}$$

Здесь $p_a(x)$ — «прямоугольная» функция

$$p_a(x) = \begin{cases} \frac{1}{2a} & \text{при} \quad |x| \leqslant a, \\ 0 & \text{при} \quad |x| > a, \end{cases}$$

показанная на рис. 3.28 пунктиром. Формула (3.74) с очевидностью следует из (3.73), поскольку

$$p_a(x) \leftrightarrow \frac{\sin au}{au},$$

и, следовательно, (см. (1.34) и (1.37))

$$p_a(x)e^{in\frac{\pi}{a}x} \leftrightarrow \frac{\sin a(u - n\frac{\pi}{a})}{a(u - n\frac{\pi}{a})}.$$

Ещё раз подчеркнём важную особенность представления (3.74). Имеется произвол в выборе интервала отсчётов $u_0 = \frac{\pi}{a}$: при любом $u_0 \leqslant \frac{\pi}{D}$ (т. е. при любом $a \geqslant D$) формула (3.74) справедлива. Максимально допустимый интервал отсчётов равен $(u_0)_{\max} = \frac{\pi}{D}$. Напомним, D определяет *минимальный интервал* значений x ($|x| \leqslant D$, вне которого поле $f(x)$ тождественно равно нулю (рис. 3.28). В этом случае формула (3.74) имеет вид

$$f(x) = p_D(x) \sum_n F\left(n\frac{\pi}{D}\right) e^{in\frac{\pi}{D}x}. \tag{3.75}$$

В частности, формула (3.75) определяет граничное поле при дифракции плоской волны на транспаранте конечного размера $|x| \leqslant D$, пропускаемость которого (модуляционная характеристика) описывается финитной функцией $f(x)$.

Наглядную «оптическую» интерпретацию формулы (3.75) поясняет рис. 3.29. Поле $f(x)$ представляется суммой пучков света

$$f_n(x) = p_D(x) F(u_n) e^{iu_n x} \qquad (3.76)$$

($u_n = n\pi/D$), каждый из которых имеет плоский фазовый фронт. Направление распространения пучка определяется его пространственной частотой:

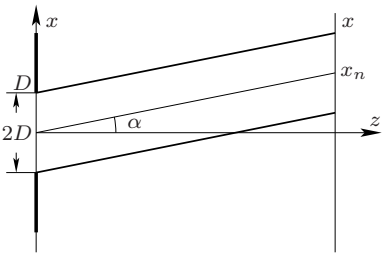


Рис. 3.29

$$u_n = k \sin \alpha_n = n \frac{\pi}{D}, \qquad \text{т. е.} \qquad \sin \alpha_n = n \frac{\lambda}{2D},$$

а поперечное сечение пучка (при малых α) равно $2D$. Амплитуда и начальная фаза колебаний определяются весовым множителем $F(u_n)$, т. е. значением спектра поля $f(x)$ в соответствующей отсчётной точке $u_n = n\frac{\pi}{D}$. Пространственный спектр пучка (3.76) есть не что иное, как n-е слагаемое в ряде Котельникова, т. е.

$$F_n(u) = F\left(n \frac{\pi}{D}\right) \cdot \frac{\sin D(u - n\frac{\pi}{D})}{D(u - n\frac{\pi}{D})}. \qquad (3.77)$$

Спектр $F_n(u)$ показан на рис. 3.30. Ширина спектра, определяемая по ширине главного максимума функции $F_n(u)$, равна $\Delta u \simeq \frac{2\pi}{D}$. Каждый пучок можно рассматривать, следовательно, как «пакет», состоящий из непрерывного набора плоских волн, пространственные частоты которых заполняют интервал

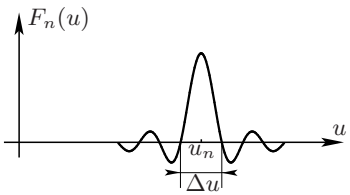


Рис. 3.30

$$u_n - \frac{\pi}{D} \leqslant u \leqslant u_n + \frac{\pi}{D}. \qquad (3.78)$$

Что происходит с каждым пучком (3.76), т. е. с каждым «пакетом», по мере распространения в области $z > 0$?

Будем исходить из спектрального равенства (3.18), связывающего пространственные спектры $F(u)$ и $G(u)$ в двух плоскостях, разделённых промежутком z; где $H(u)$ — частотная характеристика свободного пространства, которая во френелевском приближении имеет вид

$$H(u) = e^{i\sqrt{k^2-u^2}\,z} \approx e^{ikz}\, e^{-i\frac{z}{2k}u^2}.$$

Используя (3.77), находим спектр n-го пучка в плоскости $z = \text{const} > 0$:

$$G_n(u) = e^{ikz}\, F(u_n) \frac{\sin D(u-u_n)}{D(u-u_n)}\, e^{-i\frac{z}{2k}u^2}.$$

Оценим фазовый множитель $e^{-i\frac{z}{2k}u^2}$ с учётом того факта, что функция отсчётов $\frac{\sin D(u-u_n)}{D(u-u_n)}$ существенно отлична от нуля лишь в полосе частот (3.78): $v = |u-u_n| \lesssim \frac{\pi}{D}$. Мы имеем

$$\frac{z}{2k}u^2 = \frac{z}{2k}(u_n+v)^2 \approx \frac{z}{2k}u_n^2 + \frac{z}{k}u_n v.$$

Последнее равенство справедливо при условии

$$\frac{z}{2k}v^2 \lesssim \frac{z}{2k}\left(\frac{\pi}{D}\right)^2 \ll \pi,$$

т. е. при

$$\frac{D^2}{\lambda z} \gg 1. \tag{3.79}$$

Полагая условие применимости геометро-оптического приближения (3.79) выполненным, получаем

$$G_n(u) = e^{ikz} F(u_n)\, e^{-i\frac{z}{2k}u_n^2}\, \frac{\sin Dv}{Dv}\, e^{-\frac{z}{k}u_n v}.$$

Преобразуя обе части последнего равенства по Фурье, найдём поле, созданное в плоскости наблюдения n-м пучком

$$g_n(x) = e^{ikz} F(u_n)\, e^{iu_n x}\, e^{-i\frac{z}{2k}u_n^2}\, p_D(x-x_n), \tag{3.80}$$

где $x_n = \frac{z}{k}u_n$, $u_n = n\frac{\pi}{D}$.

Итак, каждый пучок, образующий в плоскости $z = 0$ поле (3.76), создаёт в плоскости наблюдения $z > 0$ поле (3.80). Результат имеет очевидную физическую интерпретацию (рис. 3.29): распространение пучка через промежуток свободного пространства сводится к двум

эффектам — квадратичному набегу фазы (в точности равному набегу фазы плоской волны с тем же направлением распространения) и поперечному смещению пучка на расстояние $x_n = n\frac{\lambda}{2D}z$.

Таким образом, в области, определяемой неравенством (3.79), искомое поле в плоскости наблюдения представляется суммой пучков (3.80):

$$g(x) = e^{ikz}\sum_n F(u_n)\, e^{iu_n x}\, e^{-i\frac{z}{2k}u_n^2}\, p_D(x - x_n), \tag{3.81}$$

или

$$g(x) = e^{ikz}\sum_n e^{-i\frac{z}{2k}u_n^2}\, f_n(x - x_n).$$

Важное утверждение, которое является прямым следствием теоремы Котельникова, состоит в следующем: выражение для поля, аналогичное (3.81), справедливо в действительности в любой плоскости наблюдения $z = \mathrm{const} > 0$, не обязательно удовлетворяющей условию геометрической оптики (3.79).

Действительно, теорема Котельникова допускает свободу выбора расстояния между отсчётными точками, которое лишь не должно превышать величины $\frac{\pi}{D}$. Пусть для данных z и D условие (3.79) не выполняется. Тогда поле, создаваемое каждым пучком, уже не преобразуется в соответствии с (3.81), иными словами, нельзя пренебречь дифракционной расходимостью пучка. Поступим следующим образом. Выберем константу $a > D$ так, чтобы при заданном расстоянии z выполнялось условие

$$\frac{a^2}{\lambda z} \gg 1. \tag{3.82}$$

Интервал отсчётов в теореме Котельникова выберем равным $\frac{\pi}{a}$. Теорема Котельникова запишется в виде (3.73) и, следовательно, граничное поле $f(x)$ — в виде (3.74). При условии (3.82) поле в плоскости $z = \mathrm{const} > 0$ примет вид, аналогичный (3.81), с той существенной разницей, что отсчётные точки $u_n = \frac{n\pi}{a}$ лежат более плотно, направления распространения слагаемых пучков определяются теперь формулой $\sin\alpha_n = n\frac{\lambda}{2a}$, а поперечное сечение пучков равно $2a$. Важно отметить, что сумма пучков $f_n(x)$ с поперечным сечением $2a > 2D$ даёт граничное поле $f(x)$, тождественно равное нулю вне области $|x| \geqslant D$ (а не $|x| \geqslant a$!), таков удивительный результат интерференции пучков. Используя это новое разложение (с более густой сеткой отсчётных точек), мы приходим к формуле

$$g(x) = e^{ikz}\sum_n F\left(n\frac{\pi}{a}\right)\, e^{in\frac{\pi}{a}x}\, e^{-i\frac{z}{2k}\left(n\frac{\pi}{a}\right)^2}\, p_D\left(x - n\frac{\lambda z}{2a}\right). \tag{3.83}$$

Фактически всё это означает, что при решении задачи дифракции на транспаранте размером $2D$ всегда можно выбрать дискретный набор пучков ограниченного сечения $2a > 2D$, которые являются собственными функциями задачи. Символически этот факт можно записать с помощью операторного равенства

$$L[f_n(x)] = e^{ikz} e^{-i\frac{z}{2k}u_n^2} f_n(x - x_n), \qquad u_n = n\frac{\pi}{a},\ x_n = n\frac{\lambda z}{2a}. \tag{3.84}$$

За возможность сохранить структуру формулы при произвольном z (не удовлетворяющем условию (3.79)) мы заплатили увеличением числа слагаемых. Легко показать (упражнение для читателя), что число слагаемых ряда (3.83), дающих вклад в суммарное поле в любой точке x плоскости наблюдения, равно $2P_0 = 2\frac{a^2}{\lambda z}$. Так, в точку $x = 0$ дают вклад P_0 слагаемых с $n > 0$ и P_0 слагаемых с $n < 0$. При смещении в точку наблюдения $x = x_0 = \frac{\lambda z}{2a} > 0$ число слагаемых с $n > 0$ увеличивается на единицу и становится равным $P_0 + 1$, а число слагаемых с $n < 0$ на единицу уменьшается и становится равным $P_0 - 1$. При смещении в пределах $\Delta x = \frac{\lambda z}{2a}$ состав слагаемых в сумме (3.83) не меняется. А все изменения при переходе через очередную точку $x_n = n\Delta x$ связаны просто с заменой одного из слагаемых в сумме.

Таким образом, теорема Котельникова даёт простой и удобный способ решения дифракционных задач, особенно эффективный в случае, если речь идёт о френелевской дифракции, где традиционные методы расчёта оказываются чрезвычайно трудоёмкими.

Глава IV

Дифракционная теория формирования изображения и разрешающая способность

§ 4.1. Элементарная оптическая система

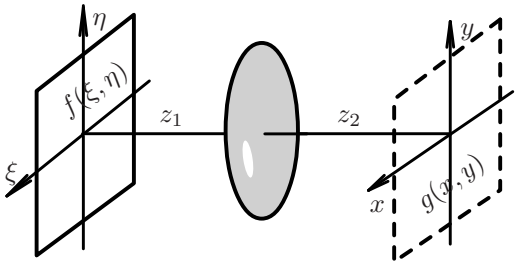


Рис. 4.1

Исчерпывающий анализ работы любого оптического устройства и изучение его предельных возможностей может быть основано только на волновом подходе. Будем рассматривать элементарную оптическую систему, изображённую на рис. 4.1, имея в виду, что любое самое сложное устройство составлено из таких элементарных систем. Будем полагать, что во «входной» плоскости (ξ,η), расположенной на расстоянии z_1 от линзы, задано поле $f(\xi,\eta)$ (например, в этой плоскости помещён предмет — транспарант с комплексной пропускаемостью $f(\xi,\eta)$, освещённый плоской волной). Задача состоит в определении поля $g(x,y)$ в «выходной» плоскости, расположенной на расстоянии z_2 справа от линзы. Замечательное свойство изображённой на рис. 4.1 системы состоит в том, что она представляет собой *линейный пространственный фильтр*; действительно, во-первых, свойством линейности обладают участки свободного пространства (от входной плоскости до линзы и от линзы до выходной плоскости); во-вторых, вспомним, как преобразуется волна, проходя через линзу (соотношение (3.20)): если поле на входе в линзу есть $f(x,y) = f_1(x,y) + f_2(x,y)$, то поле на выходе

из линзы есть $f(x,y)t(x,y) = f_1(x,y)t(x,y) + f_2(x,y)t(x,y)$, т. е. равно сумме полей, возникающих на выходе из линзы при отдельно взятых входных полях $f_1(x,y)$ и $f_2(x,y)$.

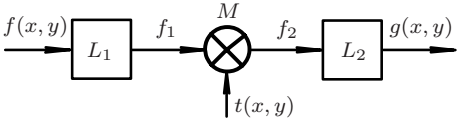


Рис. 4.2

Блок-схема, эквивалентная изучаемой системе, изображена на рис. 4.2. Линейные пространственные фильтры L_1 и L_2 описывают преобразование волны на участках свободного пространства (рис. 4.1). Эти фильтры разделены пространственным модулятором M — линзой с функцией пропускания $t(x,y)$. Ясно из изложенного выше, что вся блок-схема представляет собой некоторый линейный пространственный фильтр L: $f(x,y) \to \boxed{L} \to g(x,y)$. Как и ранее, можно использовать два подхода для изучения работы нашего фильтра: при спектральном подходе мы исследуем преобразование пространственного спектра; при «полевом» подходе изучается преобразование самого поля как функции координат, т. е. комплексной амплитуды. Для экономии места будем в дальнейшем пользоваться одномерными аналогами полученных в главе III соотношений. С помощью частотной характеристики $H(u) = e^{-i\frac{z}{2k}u^2}$ (одномерный аналог соотношения (3.60)) найдём связь между спектрами на входе и выходе фильтров L_1 и L_2 (френелевское приближение):

$$F_1(u) = F(u)e^{-i\frac{z_1}{2k}u^2}, \qquad G(u) = F_2(u)e^{-i\frac{z_2}{2k}u^2}.$$

Спектры F_1 и F_2 на входе и выходе из линзы свяжем с помощью (3.21): $F_2(u) = F_1(u) \otimes T(u)$, где $T(u)$ — фурье-образ функции пропускания линзы $t(x) = p_a(x)e^{-i\frac{k}{2f}x^2}$. Весь процесс преобразования можно описать одной формулой

$$G(u) = \left\{ \left[F(u) \cdot e^{-i\frac{z_1}{2k}u^2} \right] \otimes T(u) \right\} \cdot e^{-i\frac{z_2}{2k}u^2}, \tag{4.1}$$

дающей связь между спектрами входного $F(u)$ и выходного $G(u)$ сигналов.

При «полевом» описании нужно использовать одномерные аналоги соотношений (3.59) и (3.20), и мы получаем

$$g(x) = \left\{ \left[f(x) \otimes e^{i\frac{k}{2z_1}x^2} \right] \cdot t(x) \right\} \otimes e^{i\frac{k}{2z_2}x^2}. \tag{4.2}$$

Ниже будут подробно исследованы два наиболее интересных случая: $z_2 = f$ (поле в фокальной плоскости) и $\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}$ (поле в плоскости изображения).

§ 4.2. Поле в фокальной плоскости линзы

1. Линза бесконечных размеров $t(x) = e^{-i\frac{k}{2f}x^2}$

Для фурье-образа функции $t(x)$ имеем

$$T(u) = \int e^{-i\frac{k}{2f}x^2} \cdot e^{-iux}\,dx = e^{i\frac{f}{2k}u^2} \int e^{-i\frac{k}{2f}\left(x+\frac{f}{k}u\right)^2}\,dx.$$

Последний интеграл вычисляется с помощью (3.32). Поскольку его величина не зависит от u, то с точностью до константы будем писать: $T(u) = e^{i\frac{f}{2k}u^2}$. Выражение в фигурных скобках в (4.1) запишем в виде

$$F_2(u) = \left[F(u) \cdot e^{-i\frac{z_1}{2k}u^2}\right] \otimes e^{i\frac{f}{2k}u^2} =$$

$$= \int F(v)e^{-i\frac{z_1}{2k}v^2} \cdot e^{i\frac{f}{2k}(u-v)^2}\,dv = e^{i\frac{f}{2k}u^2} \int F(v)e^{i\left(\frac{f}{2k}-\frac{z_1}{2k}\right)v^2}e^{-i\frac{f}{k}uv}\,dv =$$

$$= e^{i\frac{f}{2k}u^2} \int F\left(\frac{kx}{f}\right) e^{i\frac{k}{2f}\left(1-\frac{z_1}{f}\right)x^2}e^{-iux}\,dx \quad (4.3)$$

(замена переменной $x = \frac{f}{k}v$). Используя (4.1), при $z_2 = f$ получаем

$$G_f(u) = \int F\left(\frac{kx}{f}\right) e^{i\frac{k}{2f}\left(1-\frac{z_1}{f}\right)x^2}e^{-iux}\,dx,$$

и, следовательно, поле в фокальной плоскости $g_f(x)$ есть

$$g_f(x) = e^{i\frac{k}{2f}\left(1-\frac{z_1}{f}\right)x^2} \cdot F\left(\frac{kx}{f}\right). \quad (4.4)$$

В чём смысл полученного соотношения? Можно представить себе, что от входной плоскости, где расположен предмет $f(\xi)$, распространяется совокупность плоских волн, составляющих поле предмета. Каждая плоская волна из этой совокупности, проходя через линзу, преобразуется в сферическую волну, сходящуюся в определённую точку в фокальной плоскости. В точку x фокусируется плоская волна с направлением $\alpha = x/f$ (рис. 4.3) и, следовательно, с пространственной частотой $u = kx/f$; амплитуда и фаза этой волны определяются функцией $F(u)$ — пространственным спектром предмета $f(\xi)$. Этот факт и отражается соотношением (4.4): поле, образующееся в фокальной плоскости линзы есть, с точностью до фазового множителя $e^{i\frac{k}{2f}\left(1-\frac{z_1}{f}\right)x^2}$, преобразование Фурье от поля предмета $f(\xi)$ (с аргументом $u = kx/f$). Фазовый множитель зависит от положения входной

плоскости z_1 (т. е. от положения предмета). При её перемещении изменяется фазовое распределение в фокальной плоскости, однако на наблюдаемом распределении интенсивности это не сказывается:

$$I(x) = gg^* = \left| F\left(\frac{kx}{f}\right)\right|^2 . \tag{4.5}$$

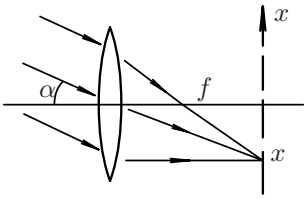


Рис. 4.3

Отметим, что в случае, когда предмет находится в передней фокальной плоскости $z_1 = f$, поле в задней фокальной плоскости является точным преобразованием Фурье поля предмета:

$$g_f(x) = F\left(\frac{kx}{f}\right). \tag{4.6}$$

Если предмет расположен вплотную к линзе ($z_1 = 0$) или поле $f(\xi)$ задано в плоскости, примыкающей к линзе, то

$$g_f(x) = e^{i\frac{k}{2f}x^2} \cdot F\left(\frac{kx}{f}\right). \tag{4.7}$$

Сопоставляя формулы (4.5) и (3.71), мы видим, что с точностью до масштаба дифракционная картина Фраунгофера совпадает с картиной поля в фокальной плоскости линзы; таким образом, для наблюдения дифракции Фраунгофера нет надобности относить плоскость наблюдения на большие расстояния ($z \gg D^2/\lambda$), достаточно наблюдать поле в фокальной плоскости линзы.

2. Линза конечных размеров $t(x) = p_a(x)e^{-i\frac{k}{2f}x^2}$

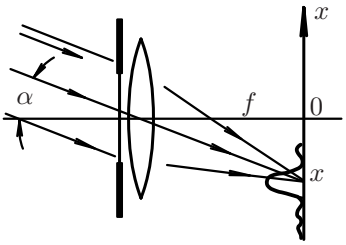


Рис. 4.4

Мы уже говорили о том, что линзу конечных размеров можно представить как бесконечную линзу, вплотную к которой помещена диафрагма — непрозрачный экран с отверстием, оставляющим открытой лишь её центральную часть (рис. 4.4).

Рассмотрим сначала плоскую волну e^{iu_0x}, падающую на линзу под углом α к оптической оси ($u_0 = k\alpha$). Волна, прошедшая через отверстие в непрозрачном экране, уже не является плоской — это целая непрерывная совокупность плоских волн со спектром

$\frac{\sin a(u-u_0)}{u-u_0}$ (см. § 3.3)). Каждая из волн этой совокупности фокусируется линзой в свою точку в фокальной плоскости. Воспользовавшись формулой (4.5), находим

$$I_f(x) = \left| \frac{\sin a\left(\frac{kx}{f} - u_0\right)}{\frac{kx}{f} - u_0} \right|^2 .$$

Итак, каждая первоначально плоская волна фокусируется не строго в точку $x_0 = f\alpha$, а в некоторую небольшую окрестность этой точки — пятнышко размера $\Delta x \approx \frac{\lambda f}{a}$. В этом и проявляется эффект дифракции на зрачке линзы (влияние её конечного размера).

Рассмотрим теперь общий случай, когда предмет-транспарант, расположенный на расстоянии z_1 от задиафрагмированной линзы, имеет комплексную пропускаемость $f(x)$, и, следовательно, поле вблизи диафрагмы имеет спектр $F_1(u) = F(u)e^{-i\frac{z_1}{2k}u^2}$. Спектр поля непосредственно за диафрагмой $\tilde{F}(u)$ найдём с помощью (3.21):

$$\tilde{F}(u) = F_1(u) \otimes \frac{\sin au}{u}.$$

Именно этот спектр необходимо подставить в (4.7) для определения поля в фокальной плоскости:

$$g_f(x) = e^{i\frac{k}{2f}x^2} \cdot \tilde{F}\left(\frac{kx}{f}\right). \tag{4.8}$$

Довольно трудно интерпретировать результат, исходя непосредственно из формулы (4.8).

Проанализируем сначала выражение для $\tilde{F}(u)$. Мы имеем

$$\tilde{F}(u) = \left[F(u)e^{-i\frac{z_1}{2k}u^2}\right] \otimes \frac{\sin au}{u} = \int F(u-v)e^{-i\frac{z_1}{2k}(u-v)^2}\frac{\sin av}{v}\,dv = \\ = e^{-i\frac{z_1}{2k}u^2}\int F(u-v)\frac{\sin av}{v}e^{-i\frac{z_1}{2k}v^2}e^{i\frac{z_1}{k}uv}\,dv. \tag{4.9}$$

Заметим, что подынтегральная функция заметно отлична от нуля лишь для значений $|v| \lesssim \pi/a$, а при таких значениях v $e^{-i\frac{z_1}{2k}v^2} \approx 1$, если $\frac{\lambda z_1}{a^2} \ll 1$. Полагая последнее неравенство выполненным (приближение геометрической оптики относительно размеров линзы), находим

$$\tilde{F}(u) = e^{-i\frac{z_1}{2k}u^2}\int F(u-v)\frac{\sin av}{v}e^{i\frac{z_1}{k}uv}\,dv.$$

Согласно определению,

$$F(u-v)=\int f(\xi)e^{-i(u-v)\xi}\,d\xi.$$

Следовательно,

$$\tilde{F}(u)=e^{-i\frac{z_1}{2k}u^2}\int f(\xi)e^{-iu\xi}\left[\int\frac{\sin av}{v}e^{i(\frac{z_1}{k}u+\xi)v}\,dv\right]d\xi.$$

Внутренний интеграл равен $p_a\left(\frac{z_1}{k}u+\xi\right)$ (см. пример 1 § 1.7), поэтому получаем

$$\tilde{F}(u)=e^{-i\frac{z_1}{2k}u^2}\int f(\xi)p_a\left(\frac{z_1}{k}u+\xi\right)e^{-iu\xi}\,d\xi. \tag{4.10}$$

Окончательное выражение для поля в фокальной плоскости найдём, подставляя (4.10) в (4.8) и полагая $u=\frac{kx}{f}$:

$$g_f(x)=e^{i\frac{k}{2f}(1-\frac{z_1}{f})x^2}\int f(\xi)p_a\left(\frac{z_1}{k}u+\xi\right)e^{-i\frac{kx}{f}\xi}\,d\xi. \tag{4.11}$$

Интеграл в последнем выражении представляет собой фурье-образ функции $f_0(\xi)=f(\xi)p_a\left(\frac{z_1}{f}x+\xi\right)$, поэтому (4.11) можно переписать в виде

$$g_f(x)=e^{i\frac{k}{2f}(1-\frac{z_1}{f})x^2}\cdot F_0\left(\frac{kx}{f}\right), \tag{4.12}$$

где $F_0(u)$ — фурье-образ функции

$$f_0(\xi)=f(\xi)p_a\left(\frac{z_1}{k}u+\xi\right)=f(\xi)p_a\left(\frac{z_1}{f}x+\xi\right).$$

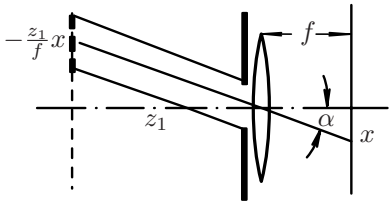


Рис. 4.5

Роль диафрагмы, как видно из сравнения (4.12) и (4.4), сводится к тому, что поле в любой точке x фокальной плоскости определяется не полем всего предмета $f(\xi)$, а лишь той частью предмета $f_0(\xi)$, которая получается проецированием диафрагмы размера $2a$ на входную плоскость в направлении $\alpha=x/f$. На рис. 4.5 предмет изображён в виде решётки, причём жирными штрихами выделена та часть решётки, которая определяет поле в точке x фокальной плоскости. Ограничение эффективного размера предмета, обусловленное конечностью размера линзы, называется *виньетированием.*

Пример. Поле в фокальной плоскости объектива микроскопа. Специфика этой задачи состоит в том, что, во-первых, предмет находится вблизи передней фокальной плоскости, а во-вторых, он имеет чрезвычайно малые размеры. Ясно, что проекция диафрагмы на входную плоскость «перехватывает» предмет лишь в том случае, если $\operatorname{tg}\alpha \approx \sin\alpha \leqslant \frac{a}{f}$ (рис. 4.6), и, следовательно, поле в задней фокальной плоскости отлично от нуля лишь в области $|x| \leqslant a$; отсюда определяем область пространственных частот, дающих вклад в поле $g_f(x)$: $|u| \leqslant \frac{ka}{f}$. Более высокие пространственные частоты (плоские волны с направлением $\sin\alpha > \frac{a}{f}$) не дают вклада в поле в фокальной плоскости. Мы имеем

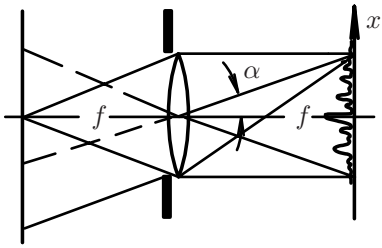


Рис. 4.6

$$g_f(x) \approx \begin{cases} F\left(\frac{kx}{f}\right) & \text{при} \quad |x| \leqslant a, \\ 0 & \text{при} \quad |x| > a. \end{cases}$$

Задача. При каких условиях эффект виньетирования не сказывается на поле в фокальной плоскости?

Задача. Транспарант с пропускаемостью $f(\xi,\eta)$ (рис. 4.1) освещается сферической волной, источник которой находится на расстоянии z_s от линзы. Найти положение плоскости z_2, поле в которой пропорционально преобразованию Фурье пропускаемости транспаранта.

Ответ: $\dfrac{1}{z_s} + \dfrac{1}{z_2} = \dfrac{1}{f}$.

§ 4.3. Функция рассеяния точки

Определить поле в плоскости изображения $\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}$ можно было бы с помощью общих соотношений (4.1) или (4.2) для произвольного входного поля $f(\xi,\eta)$. Поступим несколько по-иному. Пусть во входной плоскости в точке (ξ,η) имеется точечный источник света единичной амплитуды, излучающий с нулевой начальной фазой (рис. 4.7). Такой источник излучает сферическую волну e^{ikr}/r, которая, проходя через линзу, создаёт в плоскости

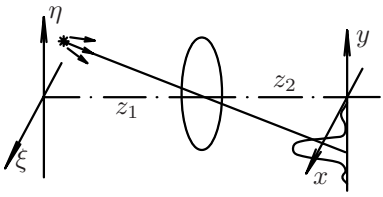


Рис. 4.7

изображения некоторое поле $h(x,y,\xi,\eta)$. Это поле является функцией координат (x,y) в выходной плоскости и положения точечного источника (ξ,η) во входной плоскости и называется *функцией рассеяния точки*. Зная функцию рассеяния точки, можно найти поле в плоскости изображения для произвольного поля на входе $f(\xi,\eta)$. Действительно, согласно принципу Гюйгенса–Френеля (3.49), каждую точку входной плоскости (бесконечно малый элемент $d\xi d\eta$) можно считать источником сферической волны:

$$f(\xi,\eta)\frac{e^{ikr}}{r}\,d\xi d\eta \tag{4.13}$$

(амплитуда и начальная фаза колебаний этой волны задаются значением $f(\xi,\eta)$ в данной точке (ξ,η)). Теперь вспомним свойство линейности: если сферическая волна единичной амплитуды (и нулевой начальной фазы) создаёт в плоскости изображения поле $h(x,y,\xi,\eta)$, то волна (4.13), очевидно, создаёт в плоскости изображения поле

$$f(\xi,\eta)h(x,y,\xi,\eta)\,d\xi d\eta. \tag{4.14}$$

Результирующее изображение, создаваемое всем предметом (т. е. всеми точечными источниками Гюйгенса–Френеля), является суммой полей (4.14):

$$g(x,y) = \iint f(\xi,\eta)h(x,y,\xi,\eta)\,d\xi d\eta. \tag{4.15}$$

Мы пришли к интегралу суперпозиции, определяющему поле $g(x,y)$ в плоскости изображения через заданное поле предмета $f(\xi,\eta)$ и функцию рассеяния точки $h(x,y,\xi,\eta)$.

Читатель непременно обратит внимание на бросающуюся в глаза аналогию, вытекающую из сравнения соотношений (1.21) и (4.15). Выходной сигнал временно́го фильтра (например, напряжение на конденсаторе в колебательном контуре) связан с входным сигналом (внешней ЭДС, действующей на контур) соотношением (1.21), в котором $h(t,\tau)$ — импульсная реакция линейного фильтра — выходной сигнал фильтра в момент времени t, возбуждённого δ-импульсом в момент времени τ. Функция рассеяния точки в оптической системе формирования изображения является полным аналогом импульсной реакции временного фильтра, при этом точечный источник света, расположенный во входной плоскости, играет роль δ-импульса входного воздействия. Функция рассеяния точки (как и импульсная реакция) полностью определяет свойства системы, так как, зная её, можно найти изображение, создаваемое произвольным предметом.

Найдём конкретный вид функции рассеяния точки (одномерный случай, рис. 4.8). Поле на входе в линзу $f_1(x)$ создаётся сферической волной, излучаемой точечным источником, расположенным в точке ξ входной плоскости. Воспользуемся параболическим приближением (2.22). Опустив множители, не зависящие от переменной x, имеем

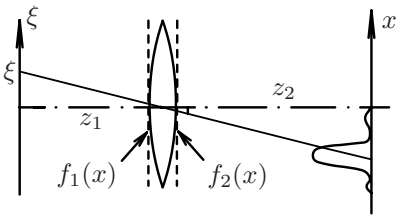


Рис. 4.8

$$f_1(x) = e^{i\frac{k}{2z_1}(x-\xi)^2}.$$

Линза, как мы знаем, является пространственным модулятором с функцией пропускания

$$t(x) = p_a(x)e^{-i\frac{k}{2f}x^2}$$

($2a$ — диаметр линзы — размер «входного зрачка» оптической системы), поэтому поле на выходе из линзы имеет вид

$$f_2(x) = f_1(x)\cdot t(x) = e^{i\frac{k}{2z_1}(x-\xi)^2}e^{-i\frac{k}{2f}x^2}p_a(x). \qquad (4.16)$$

Наконец, рассматривая свободное пространство, отделяющее плоскость изображения от плоскости выхода из линзы, как линейный фильтр с импульсной реакцией

$$h_2(x) = e^{i\frac{k}{2z_2}x^2}$$

(френелевское приближение), находим поле в плоскости изображения:

$$h(x,\xi) = f_2(x,\xi) \otimes e^{i\frac{k}{2z_2}x^2}.$$

Распишем последнее выражение, используя (4.16):

$$h(x,\xi) = \int e^{-i\frac{k}{2f}y^2}\cdot e^{i\frac{k}{2z_1}(y-\xi)^2}\cdot p_a(y)\cdot e^{i\frac{k}{2z_2}(x-y)^2}\,dy.$$

Не зависящие от переменной интегрирования множители вынесем из-под знака интеграла:

$$h(x,\xi) = e^{i\frac{k}{2z_2}x^2}e^{i\frac{k}{2z_1}\xi^2}\int e^{i\frac{k}{2}\left(\frac{1}{z_1}+\frac{1}{z_2}-\frac{1}{f}\right)y^2}p_a(y)e^{-ik\left(\frac{\xi}{z_1}+\frac{x}{z_2}\right)y}\,dy.$$

Мы рассматриваем плоскость изображения, для которой справедлива формула линзы и, следовательно:

$$e^{i\frac{k}{2}\left(\frac{1}{z_1}+\frac{1}{z_2}-\frac{1}{f}\right)y^2}=1,$$

поэтому

$$h(x,\xi)=e^{i\frac{k}{2z_2}x^2}e^{i\frac{k}{2z_1}\xi^2}\int p_a(y)e^{-ik\left(\frac{\xi}{z_1}+\frac{x}{z_2}\right)y}\,dy. \tag{4.17}$$

Интеграл в (4.17) имеет вид преобразования Фурье единично-нулевой функции $p_a(y)$ (функции зрачка). Используя (1.34), получаем окончательно:

$$h(x,\xi)=e^{i\frac{k}{2z_2}x^2}e^{i\frac{k}{2z_1}\xi^2}\frac{\sin\frac{ka}{z_2}\left(x+\frac{z_2}{z_1}\xi\right)}{\frac{k}{z_2}\left(x+\frac{z_2}{z_1}\xi\right)}. \tag{4.18}$$

Интенсивность в плоскости изображения есть квадрат модуля функции (4.18):

$$I(x)=|h(x,\xi)|^2=\left|\frac{\sin\frac{ka}{z_2}\left(x+\frac{z_2}{z_1}\xi\right)}{\frac{k}{z_2}\left(x+\frac{z_2}{z_1}\xi\right)}\right|^2. \tag{4.19}$$

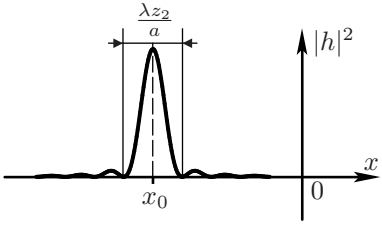


Рис. 4.9

Функция $|h(x,\xi)|^2$ изображена на рис. 4.9. Максимум интенсивности находится в точке $x_0=-\frac{z_2}{z_1}\xi$, соответствующей положению геометрического изображения точечного источника. Принципиальный результат, который не может быть получен в рамках геометрической оптики, состоит в том, что изображение точечного источника не является точкой; дифракция на входной диафрагме, ограничивающей размер линзы, приводит к тому, что изображение представляет собой яркое пятнышко, размер которого можно оценить по ширине главного максимума функции (4.19). Этот размер равен

$$\Delta x=\frac{\lambda z_2}{a}. \tag{4.20}$$

§ 4.4. Разрешающая способность (когерентные и некогерентные источники)

Теперь возникает вопрос: если во входной плоскости есть два близко расположенных точечных источника, сумеем ли мы «разрешить»

их, то есть можно ли по картине интенсивности, возникающей в плоскости изображения, отличить два точечных источника от одного?

Совершенно ясно, что если функции рассеяния двух источников не перекрываются (рис. 4.10), то такие источники всегда различимы: по картине рис. 4.10 можно однозначно ответить на вопрос, один или два источника находятся во входной плоскости; поэтому, учитывая (4.20), имеем неравенство $|x_1 - x_2| > \frac{\lambda z_2}{a}$ (или $|\xi_1 - \xi_2| > \frac{\lambda z_1}{a}$) — условие разрешения *для любых источников*. Гораздо сложнее обстоит дело в случае сильного перекрытия функций рассеяния. К изучению этого вопроса мы и перейдём.

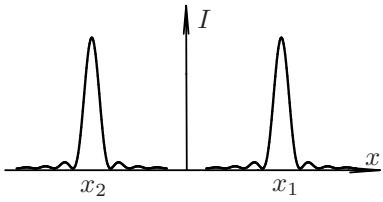


Рис. 4.10

Пусть поле, создаваемое источником S_1, есть h_1, а поле источника S_2 есть h_2. В силу линейности фильтра суммарное поле равно $g(x) = h_1(x,\xi_1) + h_2(x,\xi_2)$. Интенсивность в плоскости изображения определяется формулой

$$I(x) = |h_1(x,\xi_1) + h_2(x,\xi_2)|^2 .$$

Раскрывая последнее равенство, получаем

$$I(x) = |h_1(x,\xi_1)|^2 + |h_2(x,\xi_2)|^2 + 2\,\mathrm{Re}\, h_1(x,\xi_1)h_2^*(x,\xi_2). \qquad (4.21)$$

Первые два слагаемых — это интенсивности, создаваемые каждым из источников S_1 и S_2 в отдельности. Третье слагаемое — так называемый *интерференционный член*, он зависит от разности фаз волн, приходящих в точку x плоскости изображения от источников S_1 и S_2.

Рассмотрим случай, когда оба источника излучают с одинаковой (единичной) амплитудой, а начальные фазы излучаемых ими волн равны соответственно φ_1 и φ_2. Тогда

$$h_1(x,\xi_1) = e^{i\varphi_1} \cdot h(x,\xi_1), \qquad h_2(x,\xi_2) = e^{i\varphi_2} \cdot h(x,\xi_2), \qquad (4.22)$$

где $h(x,\xi)$ — функция рассеяния, определяемая равенством (4.18) (т. е. поле, соответствующее источнику, излучающему с нулевой начальной фазой). Подставляя (4.22) в (4.21), находим

$$I(x) = |h(x,\xi_1)|^2 + |h(x,\xi_2)|^2 + 2\,\mathrm{Re}\, e^{i\Delta\varphi} h(x,\xi_1)h^*(x,\xi_2). \qquad (4.23)$$

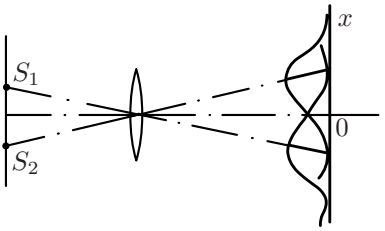


Рис. 4.11

Проанализируем (4.23), расположив источники симметрично относительно оптической оси ($\xi_1 = -\xi_2 = \xi$) таким образом, чтобы максимум функции рассеяния от одного источника совпал с первым минимумом функции рассеяния от другого (рис. 4.11). Очевидно, $|S_1S_2| = 2\xi = \frac{\lambda z_1}{2a}$. О наличии или отсутствии двух источников во входной плоскости можно судить по тому, есть ли заметный провал в суммарной интенсивности (4.23) в точке $x = 0$. Поскольку $h(0,\xi) = h(0,-\xi)$, то получаем

$$I(0) = 2|h(0,\xi)|^2(1 + \cos\Delta\varphi). \tag{4.24}$$

Наилучшим образом изображения источников различимы при $\Delta\varphi = \pi$: интенсивность в точке, разделяющей изображения ($x = 0$), при этом равна нулю. Источники практически неразличимы при $\Delta\varphi = 0$. Интенсивность в точке $x = 0$ при этом равна $4|h(0,\xi)|^2$. Подставляя в (4.24) значение $\xi = \frac{\lambda z_1}{4a}$, убеждаемся, что $I(0)$ примерно в 1,8 раза превышает значение интенсивности в точках, соответствующих положению геометрических изображений. Наконец, при $\Delta\varphi = \pi/2$ имеем $I(0) = 2|h(0,\xi)|^2$, т. е. складываются интенсивности, создаваемые каждым из источников. Используя (4.19), можно показать, что провал в суммарной интенсивности в точке $x = 0$ составляет примерно 20% от максимального значения (в точках $x_1 = -x_2 = \frac{\lambda z_2}{4a}$). Все три ситуации изображены на рис. 4.12а, б, в.

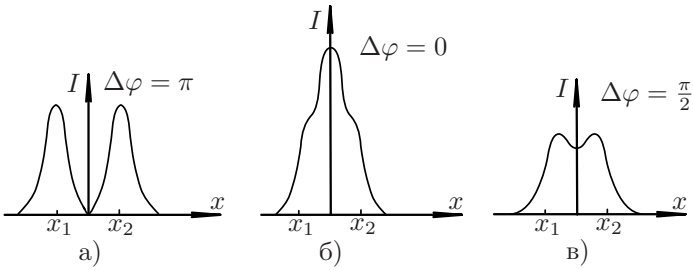


Рис. 4.12

Выше была изучена возможность разрешения двух источников, излучающих с постоянной, не зависящей от времени разностью фаз $\Delta\varphi$.

Такие источники называются *когерентными*. Вывод, который можно сделать из проведённого анализа, состоит в следующем: разрешение двух близко расположенных когерентных источников существенным образом зависит от разности фаз излучаемых ими волн. Нельзя однозначно ответить на вопрос, различимы ли когерентные источники, функции рассеяния которых сильно перекрываются, — всё определяется значением разности фаз колебаний поля в источниках.

Мы рассматривали источники гармонических волн (строго определённой частоты), такие источники всегда когерентны. Но понятие когерентности шире, чем понятие монохроматичности. Для когерентности существенно, чтобы разность фаз колебаний сохранялась (почти) неизменной лишь в течение времени, определяемом инерционностью приёмника (времени наблюдения). Когерентные колебания в точках S_1 и S_2 можно создать, освещая эти точки лучом лазера или даже при определённых условиях, используя обычный тепловой источник. Для углублённого изучения теории когерентности отсылаем читателя к книге [11].

Рассмотрим теперь другой предельный случай, когда разность фаз волн, излучаемых источниками S_1 и S_2, с течением времени быстро и хаотически изменяется так, что за время наблюдения $\Delta\varphi$ с равной вероятностью принимает любые значения в интервале $[-\pi,\pi]$ — такие источники называются *некогерентными*. Среднее за время наблюдения значение $\cos\Delta\varphi$ при этом равно нулю, и мы получаем

$$I(x) = |h(x,\xi_1)|^2 + |h(x,\xi_2)|^2,$$

т. е. *интенсивность, создаваемая в плоскости изображения некогерентными источниками, равна сумме интенсивностей, создаваемых отдельно взятыми источниками.*

В оптике часто пользуются критерием Релея разрешения некогерентных источников, который состоит в следующем: два точечных источника считаются различимыми (находящимися на пределе разрешения), если максимум интенсивности света, создаваемого в плоскости изображения одним источником, совпадает с первым минимумом интенсивности от другого источника (именно так мы располагали ранее когерентные источники). Картина, соответствующая пределу разрешения, изображена на рис. 4.12в. (Точно такая же картина возникает при когерентных источниках, излучающих с фазовым сдвигом $\pi/2$.) Напомним, что провал в центре интенсивности в точке $x = 0$, разделяющей изображения (по которому и судят о наличии двух источников), составляет $\sim$20%. Конечно, следует понимать условность критерия Релея. Используя современные методы обработки фотографий, можно

различить источники, находящиеся на существенно меньшем расстоянии.

§ 4.5. Оптическое изображение при когерентном и некогерентном освещении предмета

Снова рассмотрим два предельных случая.

Первый случай — *полностью когерентное освещение предмета*: разность фаз колебаний, создаваемых освещающей волной в любых двух точках предмета, сохраняется неизменной в течение времени наблюдения. При этом всё происходит так, как если бы волна была строго гармонической, и поле в плоскости изображения определяется интегралом суперпозиции (4.15), а его интенсивность есть

$$I(x) = gg^* = \left| \int f(\xi) h(x,\xi)\, d\xi \right|^2, \tag{4.25}$$

где $h(x,\xi)$ — функция рассеяния (4.18).

Второй случай — *полностью некогерентное освещение*: разность фаз колебаний, создаваемых освещающей волной в любых двух точках предмета, хаотически меняется за время наблюдения. Пусть предмет-транспарант с комплексной пропускаемостью $f(\xi)$ освещается такой «пространственно-некогерентной» волной единичной амплитуды. Любую точку предмета ξ_n можно считать «вторичным» источником сферической волны

$$f(\xi_n) \frac{e^{ikr_n}}{r_n} e^{i\varphi(\xi_n,t)},$$

начальная фаза которой $\varphi(\xi_n,t)$ быстро и хаотически меняется со временем (и независимо от изменения фазы в любой другой точке ξ_m). В плоскости изображения такой источник создаёт поле

$$f(\xi_n) h(x,\xi_n) e^{i\varphi(\xi_n,t)},$$

фаза которого также хаотически быстро изменяется за время наблюдения. Суммарное поле от всех точек есть

$$g = \sum_n f(\xi_n) h(x,\xi_n) e^{i\varphi(\xi_n,t)},$$

а его интенсивность

$$I(x) = \overline{gg^*} = \sum_m \sum_n f(\xi_n) f^*(\xi_m) h(x,\xi_n) h^*(x,\xi_m) \overline{e^{i\Delta\varphi_{nm}(t)}}, \tag{4.26}$$

где $\Delta\varphi_{nm} = \varphi(\xi_n,t) - \varphi(\xi_m,t)$. В результате усреднения за время наблюдения (большое по сравнению с характерным временем изменения разностей фаз $\Delta\varphi_{nm}(t)$) исчезнут все члены в сумме (4.26) с $n \neq m$ (так как $\overline{e^{i\Delta\varphi}} = \overline{\cos\Delta\varphi} + i\overline{\sin\Delta\varphi} = 0$) и, следовательно,

$$I(x) = \sum_n |f(\xi_n)|^2 |h(x,\xi_n)|^2. \tag{4.27}$$

Мы снова пришли к закону сложения интенсивностей для некогерентного освещения: результирующая интенсивность получается сложением интенсивностей $I_n = |f(\xi_n)|^2|h(x,\xi_n)|^2$, создаваемых отдельными точками предмета. Переходя к непрерывному распределению излучающих площадок $\Delta\xi$ (и учитывая, что интенсивность излучения каждой малой площадки пропорциональна её площади $\Delta\xi$), имеем вместо (4.27):

$$I(x) = \int |f(\xi)|^2 |h(x,\xi)|^2 \, d\xi. \tag{4.28}$$

Формула (4.28), так же как и (4.15), является интегралом суперпозиции, определяющим выходной сигнал линейного фильтра. Важное отличие состоит в том, что входным и выходным сигналами фильтра являются не поля (комплексные амплитуды $f(x)$ и $g(x)$), а интенсивности $I_f = \overline{ff^*}$ и $I_g = \overline{gg^*}$ во входной и выходной плоскостях. Роль импульсной реакции играет квадрат модуля функции рассеяния, т. е. распределение интенсивности, которое создаётся в плоскости изображения точечным источником, находящимся в плоскости предмета.

Итак, при *когерентном освещении* мы имеем пространственный фильтр, *линейный по полю* (формула (4.15)), а при *некогерентном освещении — линейный по интенсивности* (формула (4.28)).

§ 4.6. Анализ оптического изображения (спектральный подход)

Проще всего проанализировать работу любого линейного фильтра, используя спектральный подход, поскольку спектры входного и выходного сигналов связаны чрезвычайно простым соотношением (1.14). Этот подход, однако, возможен только в том случае, когда линейный фильтр является *стационарным*, т. е. когда изменение момента появления δ-импульса на входе фильтра приводит лишь к сдвигу во времени импульсной реакции (§ 1.6), без изменения её функционального вида, т. е. когда $h(t,\tau) = h(t-\tau)$. Аналогичное свойство пространственного фильтра называется *пространственной инвариантностью* или *изопланарностью*. Это свойство означает, что при переносе точечного источника по входной плоскости (изменении его координат ξ,

η) поле в выходной плоскости — функция рассеяния $h(x,y,\xi,\eta)$ — переносится как целое, не изменяя своего функционального вида, т. е. $h(x,y,\xi,\eta) = h(x-\xi,y-\eta)$. Из выражения (3.45) следует, что свободное пространство является изопланарным фильтром (именно поэтому мы могли использовать спектральный подход при решении дифракционной задачи).

Является ли изопланарной оптическая система формирования изображения? Здесь следует рассмотреть два случая: когерентное и некогерентное освещение предмета.

1. *Когерентное освещение.* Поле в плоскости изображения определяется интегралом суперпозиции (4.15) с импульсной реакцией (4.18). Введя замену переменной $\frac{z_1}{z_2}x = -x_1$ (что означает изменение масштаба и положительного направления оси x в «выходной» плоскости), имеем вместо (4.18)

$$h(x_1,\xi) = \left[e^{i\frac{kz_2}{2z_1^2}x_1^2} e^{i\frac{k}{2z_1}\xi^2}\right] \frac{\sin\frac{ka}{z_1}(x_1-\xi)}{\frac{k}{z_1}(x_1-\xi)}. \tag{4.29}$$

Выражение (4.29) не удовлетворяет условию изопланарности при произвольных значениях x_1 и ξ — мешает фазовый множитель

$$e^{i\frac{kz_2}{2z_1^2}x_1^2} e^{i\frac{k}{2z_1}\xi^2}.$$

Предоставляем читателю убедиться, что влиянием этого множителя можно пренебречь при выполнении следующих неравенств:

$$\frac{k}{2z_1}b^2 \ll \pi; \qquad \frac{z_2}{ka^2} \ll \pi, \tag{4.30}$$

где b — размер предмета (или допустимая область перемещения точечного источника по входной плоскости), a — размер линзы. Первое из неравенств (4.30) — довольно серьёзное ограничение на размер предмета (или на допустимую область перемещения точечного источника по входной плоскости). Реально его выполнить довольно сложно, действительно, при $z_1 = 25$ см и $\lambda = 5{\cdot}10^{-5}$ см $b \ll 0{,}03$ см. Второе условие — гораздо более мягкое ограничение на минимальный размер линзы: при $z_2 = 25$ см и $\lambda = 5{\cdot}10^{-5}$ см $a \gg 0{,}03$ см (обычно линзы имеют существенно больший размер). Итак, при очень жёстком ограничении (4.30а) и очень мягком условии (4.30б) функция рассеяния $h(x,\xi)$ удовлетворяет условию изопланарности и имеет вид

$$h(x_1,\xi) = h(x_1-\xi) = \frac{\sin\frac{ka}{z_1}(x_1-\xi)}{\frac{k}{z_1}(x_1-\xi)}. \tag{4.31}$$

При этом интеграл суперпозиции (4.15), определяющий поле в плоскости изображения, является свёрткой двух функций: входного сигнала $f(x)$ и импульсной реакции

$$h(x) = \frac{\sin \frac{ka}{z_1} x}{\frac{k}{z_1} x},$$

$$g(x_1) = f(x_1) \otimes \frac{\sin \frac{ka}{z_1} x_1}{\frac{k}{z_1} x_1}. \tag{4.32}$$

Поскольку функции $f(x)$, $h(x)$ и $g(x)$ связаны операцией свёртки, то для их преобразований Фурье $F(u)$, $H(u)$ и $G(u)$ имеем $G(u) = F(u) \cdot H(u)$, где

$$H(u) = P_a\left(\frac{z_1}{k} u\right) = \begin{cases} 1 & \text{при} \quad |u| \leqslant \frac{ka}{z_1}, \\ 0 & \text{при} \quad |u| > \frac{ka}{z_1} \end{cases} \tag{4.33}$$

— частотная характеристика оптической системы, формирующей изображение. Она изображена на рис. 4.13. Мы видим, что все пространственные частоты внутри полосы $|u| \leqslant \frac{ka}{z_1}$ «пропускаются» фильтром без искажений, другими словами, все плоские волны, составляющие спектр предмета, направления распространения которых удовлетворяют условию $\sin\alpha \leqslant \frac{a}{z_1}$, достигают плоскости изображения с теми же амплитудами и фазами, с какими они входят в состав предмета. Плоские волны, лежащие вне конуса $\sin\alpha \leqslant \frac{a}{z_1}$, не участвуют в формировании изображения: фильтр с характеристикой (4.33) «обрезает высокие пространственные частоты» $|u| > \frac{ka}{z_1}$. Чем больше размер линзы, тем шире полоса пропускания (и тем уже функция рассеяния (4.18)). Ясно, что изображение является идеальным: $G(u) = F(u)$ (и $g(x) = f(x)$), если спектр предмета равен нулю вне полосы пропускания $|u| \leqslant \frac{ka}{z_1}$. Как следует из (4.33), линза бесконечных размеров даёт идеальное изображение любого предмета: частотная характеристика $H(u) \equiv 1$ при $a \to \infty$ для всех частот «u» (при этом функция рассеяния $h(x) = \delta(x)$). Разумеется, следует помнить, что наш анализ основан на френелевском приближении (§ 3.8), справедливом во всяком случае при $|u| < k$. Очевидно, неоднородные волны $|u| > k$ «отфильтровываются» свободным пространством, отделяющим плоскость предмета от линзы (§ 3.2), и, конечно, не участвуют в формировании

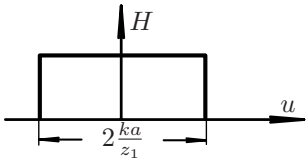


Рис. 4.13

изображения, поэтому максимальная полоса пропускания, достижимая при бесконечной линзе, есть $|u| \leqslant k$. Рисунок 4.14 иллюстрирует характер изменения спектра изображения при изменении размеров линзы. Максимальной пространственной частоте $u_{\max} \simeq \frac{ka}{z_1}$ спектра изображения соответствует, согласно соотношению неопределённости $u_{\max} \cdot x_{1\min} \simeq 2\pi$, минимальный размер деталей в изображении $x_{1\min} \simeq \frac{2\pi z_1}{ka} = \frac{\lambda z_1}{a}$, равный предельно разрешимому размеру деталей предмета (именно такая оценка разрешающей способности линзы получена в § 4.4).

Поскольку для идеальной системы ($a = \infty$) $u_{\max} = k$, то из соотношения неопределённости получаем $x_{\min} \simeq \lambda$: минимальный предельно разрешимый размер деталей предмета, который может быть получен в идеальной оптической системе, равен (по порядку величины) длине волны.

2. *Некогерентное освещение.* Будем исходить из соотношения (4.28), смысл которого состоит в том, что интенсивность в плоскости изображения представляется суперпозицией интенсивностей, создаваемых всеми излучающими точками предмета: каждая точка создаёт в плоскости изображения интенсивность $|h(x,\xi)|^2$ — функция, определяемая формулой (4.19); это и есть импульсная реакция некогерентной системы, которая после изменения масштаба и направления оси x в плоскости изображения имеет вид

$$\tilde{h}(x_1,\xi) = |h(x_1,\xi)|^2 = \left| \frac{\sin \frac{ka}{z_1}(x_1 - \xi)}{\frac{k}{z_1}(x_1 - \xi)} \right|^2 . \tag{4.34}$$

Функция $\tilde{h}(x_1,\xi)$ зависит только от разности $x_1 - \xi$, и, таким образом, при некогерентном освещении нет необходимости в требованиях (4.30) — система является изопланарной при любом размере предмета «b» и любом размере линзы «a» (разумеется, в рамках френелевского приближения): $\tilde{h}(x_1,\xi) = \tilde{h}(x_1, -\xi)$. Следовательно, интеграл в (4.28) представляет собой свёртку двух функций: интенсивности во входной плоскости $I_f(x)$ (входной сигнал некогерентной системы) и импульсной реакции

$$\tilde{h}(x) = \left| \frac{\sin \frac{ka}{z_1} x}{\frac{k}{z_1} x} \right|^2 .$$

Переходя к преобразованиям Фурье $\tilde{G}(u)$, $\tilde{F}(u)$ и $\tilde{H}(u)$ функций $I_g(x)$, $I_f(x)$ и $\tilde{h}(x)$, получаем

$$\tilde{G}(u) = \tilde{F}(u) \cdot \tilde{H}(u),$$

где

$$\tilde{H}(u) = \int |h(x)|^2 e^{-iux}\, dx \tag{4.35}$$

— частотная характеристика некогерентной системы (называемая оптической передаточной функцией (ОПФ)).

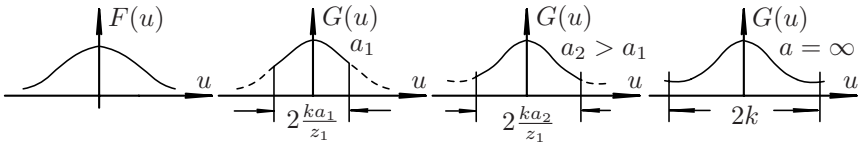


Рис. 4.14

Найдём связь между ОПФ и частотной характеристикой когерентной системы. Формулу (4.35) можно рассматривать как преобразование Фурье произведения двух функций $h(x)$ и $h^*(x)$. Согласно теореме о фурье-образе произведения двух функций, эта величина равна свёртке фурье-образов функций $h(x)$ и $h^*(x)$. Мы имеем

$$H(u) = \int h(x) e^{-iux}\, dx,$$

отсюда

$$H^*(u) = \int h^*(x) e^{+iux}\, dx,$$

и, следовательно:

$$H^*(-u) = \int h^*(x) e^{-iux}\, dx$$

т. е. $H^*(-u)$ есть фурье-образ функции $h^*(x)$, и окончательно получаем

$$\tilde{H}(u) = H(u) \otimes H^*(-u). \tag{4.36}$$

Формула (4.36) даёт связь между ОПФ и частотной характеристикой когерентной системы. Как следует из (4.33), $H^*(-u) = H(u)$, и, таким образом

$$\tilde{H}(u) = H(u) \otimes H(u). \tag{4.37}$$

Задача. Используя (4.37), показать, что если $H(u)$ имеет вид, изображённый на рис. 4.15а, то $\tilde{H}(u)$ — функция, изображённая на рис. 4.15б.

Можно ли, исходя из сравнения частотных характеристик $H(u)$ и $\tilde{H}(u)$, сделать вывод о том, какой характер освещения предмета (когерентный или некогерентный) даёт изображение лучшего качества? При обсуждении этой проблемы возникает вопрос о выборе критерия

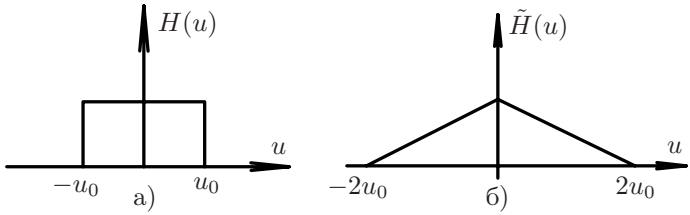


Рис. 4.15

качества. Мы уже сравнивали оба способа освещения при изучении разрешающей способности и убедились, что однозначного ответа на вопрос, какое освещение лучше, не существует. Разрешающая способность — не единственный критерий качества изображения. Изображение можно сравнивать и по другим параметрам. Обсудим один из них: контраст спектральных составляющих интенсивности изображения, рассмотрев в качестве предмета амплитудную синусоидальную решётку: $f(x) = 1 + \alpha \cos \Omega x$ при $\alpha < 1$. Мы имеем

$$F(u) = \delta(u) + \frac{\alpha}{2}[\delta(u - \Omega) + \delta(u + \Omega)],$$

$$I_f \approx 1 + 2\alpha \cos \Omega x,$$

$$\tilde{F}(u) = \delta(u) + \alpha[\delta(u - \Omega) + \delta(u + \Omega)].$$

Рассмотрим когерентное освещение. При $\Omega < u_0$ (где $u_0 = \frac{ka}{z_1}$) спектральные компоненты $\delta(u \pm \Omega)$ оказываются в пределах полосы пропускания частотной характеристики $H(u)$ и, следовательно, $G(u) = F(u)$, $g(x) = f(x)$ и $I_g(x) = I_f(x)$. Итак, контраст составляющих на частотах $\pm\Omega$ относительно постоянного фона (δ-функции в нуле) равен контрасту в предмете. При $\Omega > u_0$ компоненты $\delta(u \pm \Omega)$ лежат вне полосы пропускания фильтра и, следовательно, $G(u) = \delta(u)$, $g(x) = 1$, $I_g(x) \equiv 1$ — изображение отсутствует.

Контраст составляющих на частотах $\pm\Omega$ равен нулю. Очевидно, при некогерентном освещении и при условии $\Omega < u_0$ контраст спектральных компонент на частотах $\pm\Omega$ оказывается ниже, чем при когерентном освещении (так как функция $\tilde{H}(u)$ линейно спадает с ростом u, в то время как $H(u) = \text{const}$ внутри полосы $|u| < u_0$). Если же $u_0 < \Omega < 2u_0$, то контраст компонент $\delta(u \pm \Omega)$, хотя и уменьшен по сравнению с контрастом в предмете, но всё же отличен от нуля, так как при этом компоненты $\delta(u + \Omega)$ попадают в пределы полосы частот $|u| < 2u_0$, пропускаемой некогерентной системой. Итак, вывод:

при $\Omega < u_0$ когерентное освещение оказывается лучше некогерентного, а при $\Omega > u_0$ (но меньше $2u_0$) «лучше» некогерентное освещение решётки. (При $\Omega > 2u_0$, очевидно, изображение решётки вообще нельзя получить ни при каком способе освещения). Таким образом, можно снова заключить, что качество изображения, получаемого в данной оптической системе, существенным образом зависит не только от способа освещения (когерентно или некогерентно), но и от характеристик самого предмета.

В заключение этого параграфа — несколько слов о сравнении когерентных и некогерентных систем.

Первое существенное различие состоит в том, что входные и выходные сигналы когерентных систем — это, вообще говоря, комплексные функции координат $f(x,y)$ и $g(x,y)$, в то время как в некогерентных системах это действительные положительные функции $I_f(x,y)$ и $I_g(x,y)$. Отсюда, в частности, следует, что два предмета $a(x,y)e^{i\varphi_1(x,y)}$ и $a(x,y)e^{i\varphi_2(x,y)}$ (с одинаковыми амплитудными и различными фазовыми распределениями) дают при когерентном освещении два различных изображения, тогда как при некогерентном освещении изображения оказываются одинаковыми (так как входной сигнал некогерентной системы $I_f(x) = ff^* = a^2(x)$ при замене одного предмета другим не меняется).

Второе различие: импульсный отклик когерентной системы — комплексная функция координат $h(x)$, при этом импульсный отклик некогерентной системы — действительная положительная функция $|h(x)|^2$. Сказанное справедливо не только для функции (4.18). В следующей главе мы убедимся, что можно создать оптические системы с самыми различными импульсными откликами, т. е. с самыми различными законами преобразования входного сигнала в выходной.

Таким образом, можно сказать, что и класс функций, и класс преобразований, осуществляемых над этими функциями, в когерентных системах более широк.

Последняя глава будет посвящена анализу когерентных оптических систем.

Глава V

Пространственная фильтрация и голография

§ 5.1. Общие принципы пространственной фильтрации

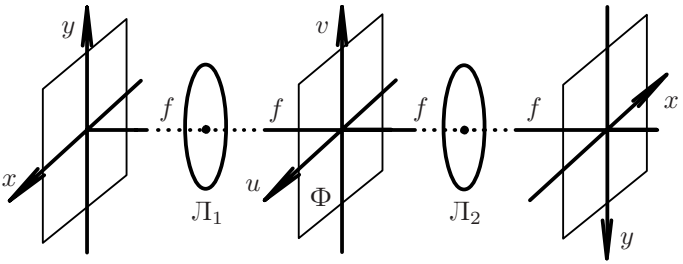


Рис. 5.1. Схема Катрона

Основные идеи пространственной фильтрации проще всего понять, изучив работу оптической системы, изображённой на рис. 5.1 (схема Катрона). Эта система состоит из двух линз $Л_1$ и $Л_2$ с общей фокальной плоскостью Ф (задняя фокальная плоскость линзы $Л_1$ совпадает с передней фокальной плоскостью линзы $Л_2$). Плоскость Ф мы будем называть *частотной* или *фурье-плоскостью*. Передняя фокальная плоскость линзы $Л_1$ является «входной» плоскостью нашей системы (и поле в ней $f(x,y)$ — входной сигнал фильтра), а задняя фокальная плоскость линзы $Л_2$ — «выходная» плоскость и поле в ней $g(x,y)$ — выходной сигнал фильтра.

Образование выходного сигнала (изображения) можно рассматривать в два этапа: во-первых, линза $Л_1$ преобразует входное поле $f(x,y)$ в поле $g_Ф(x,y)$ в частотной плоскости Ф, и эти поля связаны преобразованием Фурье (§ 4.2):

$$g_Ф(x,y) = F\left(\frac{kx}{f}, \frac{ky}{f}\right). \tag{5.1}$$

Поле $g_\Phi(x,y)$ является, по терминологии Аббе, «первичным» изображением. На втором этапе происходит преобразование первичного изображения $g_\Phi(x,y)$ в искомое поле $g(x,y)$ («вторичное» изображение) с помощью линзы $Л_2$; поля $g(x,y)$ и $g_\Phi(x,y)$ также связаны преобразованием Фурье:

$$g(x,y) = \iint g_\Phi(\xi,\eta) e^{-i\left(\frac{kx}{f}\xi + \frac{ky}{f}\eta\right)}\, d\xi d\eta. \tag{5.2}$$

Итак, наша система осуществляет два последовательных преобразования Фурье, в результате чего мы получаем $g(x,y) = f(-x,-y)$, т. е. перевёрнутое изображение предмета. Действительно, после использования (5.1) и замены переменной $\frac{k\xi}{f} = u$, $\frac{k\eta}{f} = v$ формула (5.2) записывается в виде

$$g(x,y) = \iint F(u,v) e^{-i(ux+vy)}\, du dv = f(-x,-y).$$

Отметим, что точно таким же образом можно было бы рассматривать формирование изображения в однолинзовой системе (рис. 4.1): первый этап — преобразование волны от входной плоскости к задней фокальной плоскости линзы, где возникает спектр предмета (первичное изображение), и второй этап — распространение волны в свободном пространстве, отделяющем фокальную плоскость от выходной, где возникает «вторичное» изображение. Такой подход к формированию оптического изображения называется *принципом двойной дифракции.*

Вернёмся к схеме рис. 5.1. Изменив направления осей координат в выходной плоскости на противоположные (как показано на рис. 5.1), получим в этих новых осях: $g(x,y) = f(x,y)$, т. е. выходной сигнал тождествен входному, и, следовательно, наша система имеет импульсную реакцию $h(x,y) = \delta(x,y)$ и частотную характеристику $H(u,v) \equiv 1$ (напомним, что мы пренебрегаем дифракцией, обусловленной конечностью размеров линз). Расположим теперь в частотной плоскости Ф плоский транспарант-маску с комплексной пропускаемостью $T(\frac{kx}{f}, \frac{ky}{f})$. Наш транспарант осуществляет пространственную модуляцию падающей на него волны (§ 3.4): поле $g_\Phi(x,y)$ на выходе транспаранта (на выходе частотной плоскости) есть произведение падающего на плоскость Ф поля $g_\Phi(x,y) = F\left(\frac{kx}{f}, \frac{ky}{f}\right)$ на комплексную пропускаемость транспаранта $T\left(\frac{kx}{f}, \frac{ky}{f}\right)$:

$$g_{\Phi_+}(x,y) = F\left(\frac{kx}{f}, \frac{ky}{f}\right) \cdot T\left(\frac{kx}{f}, \frac{ky}{f}\right).$$

Для определения выходного сигнала $g(x,y)$ необходимо теперь вместо функции g_{Φ} подставить в (5.2) функцию g_{Φ_+}. Мы получим (в измененных направлениях осей координат в выходной плоскости)

$$g(x,y) = \iint F\left(\frac{k\xi}{f}, \frac{k\eta}{f}\right) T\left(\frac{k\xi}{f}, \frac{k\eta}{f}\right) e^{i\left(\frac{kx}{f}\xi + \frac{ky}{f}\eta\right)}\, d\xi d\eta. \qquad (5.3)$$

Как следует из изложенного выше, поле в плоскости Ф и введённая нами пропускаемость транспаранта, устанавливаемого в этой плоскости, являются функциями переменных $u = \frac{k\xi}{f}$, $v = \frac{k\eta}{f}$, имеющих размерность пространственных частот. Можно сразу проградуировать оси координат в плоскости Ф в этих переменных $u = \frac{k\xi}{f}$, $v = \frac{k\eta}{f}$, поставив в соответствие каждой точке (ξ,η) пару чисел (u,v) — составляющих вектора $\boldsymbol{k}$ той плоской волны, которая фокусируется линзой $Л_1$ в эту точку. В переменных u, v формула (5.3) записывается в виде

$$g(x,y) = \iint F(u,v)T(u,v)e^{i(ux+vy)}\, dudv, \qquad (5.4)$$

и, как ясно из сравнения (5.4) с общим соотношением, связывающим функцию $g(x,y)$ с её преобразованием Фурье $G(u,v)$,

$$g(x,y) = \iint G(u,v)e^{i(ux+vy)}\, dudv,$$

поле на выходе частотной плоскости $g_{\Phi}(u,v) = F(u,v) \cdot T(u,v)$ (на выходе транспаранта) является спектром выходного сигнала:

$$G(u,v) = F(u,v)T(u,v). \qquad (5.5)$$

Соотношение (5.5) связывает спектры входного и выходного сигналов линейного фильтра, и, следовательно, функция $T(u,v)$ является его частотной характеристикой:

$$H(u,v) = T(u,v). \qquad (5.6)$$

Суть дела в действительности чрезвычайно проста: каждая плоская волна, входящая в состав волны, посылаемой предметом, фокусируется линзой $Л_1$ в некоторую точку в фокальной плоскости Ф. Помещая в эту плоскость транспарант, мы получаем возможность *избирательного воздействия на отдельные плоские волны*, входящие в спектр предмета, поскольку значение пропускаемости транспаранта в некоторой точке определяет изменение амплитуды и фазы лишь одной спектральной компоненты: именно той плоской волны, которая фокусируется в эту точку.

Итак, функция пропускания транспаранта, установленного в частотной плоскости, однозначно определяет частотную характеристику оптической системы. Помещая в частотную плоскость различные маски, мы получаем фильтры с различными частотными характеристиками (с различными законами преобразования входного сигнала в выходной) и имеем, таким образом, эффективный способ влияния на характеристики изображения.

При выводе (5.6) мы не учитывали влияние конечного размера линз, формирующих изображение. В § 3.4 было показано, что линзу конечного размера можно рассматривать как бесконечную линзу, накрытую диафрагмой; при этом частотная характеристика $H(u,v)$ системы есть единично-нулевая функция с шириной, определяемой размерами диафрагмы. Точно такую же частотную характеристику мы получим, рассматривая схему рис. 5.1 с бесконечными линзами, если в частотную плоскость поместить диафрагму того же размера $2a$. Функция пропускания такой диафрагмы (одномерный случай) есть

$$T_0\left(\frac{kx}{f}\right) = P_a = \begin{cases} 1 & \text{при} \quad |x| \leqslant a, \\ 0 & \text{при} \quad |x| > a, \end{cases}$$

и согласно (5.6) имеем

$$H_0(u) = \begin{cases} 1 & \text{при} \quad |u| \leqslant \frac{ka}{f}, \\ 0 & \text{при} \quad |u| > \frac{ka}{f}. \end{cases}$$

Ясно, что если теперь дополнительно к диафрагме поместить в частотную плоскость маску с произвольной функцией пропускания $T\left(\frac{kx}{f}\right)$, то полная комплексная пропускаемость двух транспарантов есть $T_0\left(\frac{kx}{f}\right) \cdot T\left(\frac{kx}{f}\right)$, и частотная характеристика нашей системы определяется равенством

$$H(u) = H_0(u) \cdot T(u), \tag{5.7}$$

которое заменяет (5.6), если необходимо учесть дифракционные эффекты, обусловленные конечностью размеров линз.

Изложенный выше принцип пространственной фильтрации (с помощью масок, устанавливаемых в частотной плоскости) позволяет решать не только сугубо оптические задачи (улучшения качества изображений и характеристик оптических систем), но также широкий круг задач, связанных с обработкой информации. Оптические системы обработки информации имеют очевидные преимущества перед существующими электронными системами:

во-первых, высокая скорость обработки, которая лимитируется лишь скоростью ввода и съёма данных (сам процесс преобразования входного сигнала в выходной происходит со скоростью распространения оптического сигнала);

во-вторых, информация, подлежащая обработке, записывается в виде комплексной пропускаемости плоского транспаранта, расположенного во входной плоскости, т. е. оптическая система обладает двумя степенями свободы. Обработка происходит *одновременно* по всей области задания входной функции $f(x,y)$. Электронные системы обладают одной степенью свободы — временем, и обработка входного сигнала $f(t)$ происходит *последовательно во времени*. На плоском транспаранте можно записать большое число функций одной переменной $f_n(x)$ (n — номер «канала», определяемый координатой y_n: $f_n(x) = f(x,y_n)$) и, таким образом, осуществлять одновременную обработку в большом числе каналов;

в-третьих, огромная информационная ёмкость запоминающих устройств и высокая надёжность кодирования и хранения информации (§ 5.5).

Вот вкратце причины, определяющие перспективность оптических методов обработки информации.

Рассмотренные ниже примеры, не претендуя на техническое решение проблем, являются хорошей иллюстрацией возможностей методов пространственной фильтрации.

§ 5.2. Методы улучшения качества изображения

1. Улучшение контраста

Пусть предмет-транспарант, помещённый во входной плоскости, имеет пропускаемость $f(x) = 1 + f_0(x)$, причём полезная информация заключена в вариациях пропускаемости (функции $f_0(x)$), а постоянная составляющая (единица) создаёт сильный постоянный фон ($|f_0(x)| \ll 1$), так что контраст изображения оказывается слабым. Поле в частотной плоскости Ф (фурье-образ функции $f(x)$) есть $g_\Phi(u) = \delta(u) + F_0(u)$ (постоянный фон концентрируется линзой $Л_1$ в точку — начало координат частотной плоскости — и даёт δ-функцию в поле g_Φ, $F_0(u)$ — спектр функции $f_0(x)$). Поместим в частотную плоскость транспарант с пропускаемостью (см. рис. 5.2)

$$T(u) = \begin{cases} 0 & \text{при} \quad |u| \leqslant \varepsilon, \\ 1 & \text{при} \quad |u| > \varepsilon. \end{cases} \tag{5.8}$$

Он представляет собой маленький непрозрачный экран, помещённый в начало координат частотной плоскости и полностью поглощает падающий на него свет. Поэтому поле на выходе частотной плоскости (являющееся спектром изображения) отличается от поля g_Φ только отсутствием δ-функции в начале координат (которая «поглощается» непрозрачным экраном):

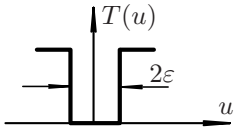


Рис. 5.2

$$G(u) = F(u) \cdot T(u) = g_\Phi(u) - \delta(u) = F_0(u).$$

В плоскости изображения получаем «отфильтрованный» полезный сигнал $g(x) = f_0(x)$ (мы предполагаем, что полезная составляющая не имеет в спектре δ-функцию на нулевой частоте и, следовательно, транспарант с пропускаемостью (5.8) не влияет на функцию $F_0(u)$).

Задача. Оценить размер непрозрачного экрана, который следует установить в частотной плоскости, если размер предмета-транспаранта равен b.

2. Исключение периодического шума

Пусть поле во входной плоскости есть $f(x) = f_0(x) + A\sin\omega_0 x$, где $f_0(x)$ — полезная составляющая, а $A\sin\omega_0 x$ — периодический шум, который требуется устранить.

Задача. Определить пропускаемость транспаранта $T(u)$, который следует установить в частотной плоскости.

3. Улучшение разрешающей способности

Согласно (5.7), можно было бы попытаться исключить влияние дифракции, поместив в частотную плоскость транспарант с пропускаемостью $T(u) = \frac{1}{H_0(u)}$ (при этом $H(u) \equiv 1$ и $h(x) = \delta(x)$, т. е. система становится идеальной). Конечно, точно такой транспарант изготовить невозможно, так как $H_0(u) = 0$ вне полосы $|u| < \frac{ka}{f}$, тем не менее принцип пространственной фильтрации оказывается полезным при решении двух рассмотренных ниже задач.

Из критерия Релея следует, что разрешающая способность оптической системы улучшается при увеличении размеров линзы, так как при этом сужается главный максимум функции рассеяния (формула (4.20)). Возникает вопрос: можно ли, не изменяя размеров линзы и используя изложенные выше принципы пространственной фильтрации, сделать функцию рассеяния более узкой и, следовательно, улучшить разрешающую способность в смысле критерия Релея?

Выберем масштаб в плоскости Ф таким образом, чтобы граничная частота функции $H_0(u)$ равнялась единице: $H_0(u) = 0$ при $|u| > 1$;

тогда функция рассеяния $h_0(x) = \frac{\sin x}{x}$ имеет ширину главного максимума $\Delta x_0 = 2\pi$.

1. Перекроем центр частотной плоскости непрозрачным экраном размера $2l$ $(l < 1)$:

$$T(u) = \begin{cases} 0 & \text{при} \quad |u| \leqslant l, \\ 1 & \text{при} \quad |u| > l. \end{cases}$$

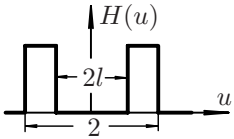


Рис. 5.3

Теперь частотная характеристика системы в силу (5.7) имеет вид

$$H(u) = H_0(u)T(u) = \begin{cases} 1 & \text{при} \quad l \leqslant |u| \leqslant 1, \\ 0 & \text{вне этих значений,} \end{cases}$$

а соответствующая функция рассеяния

$$h(x) = \frac{\sin x}{x} - \frac{\sin lx}{x}$$

имеет ширину главного максимума $\delta x = \frac{2\pi}{1+l}$. При изменении l от нуля до единицы главный максимум сужается от Δx_0 до $\Delta x_0/2$.

2. *Экран Уилкинса.* Заменим непрозрачный экран в предыдущем примере прозрачной пластинкой, вносящей фазовую задержку в π:

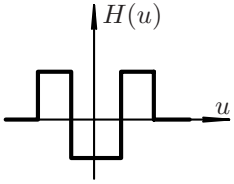


Рис. 5.4

$$T(u) = \begin{cases} -1 & \text{при} \quad |u| \leqslant l, \\ 1 & \text{при} \quad |u| > l. \end{cases}$$

Тогда

$$H(u) = \begin{cases} -1 & \text{при} \quad |u| \leqslant l, \\ 1 & \text{при} \quad l < |u| \leqslant 1. \end{cases} \tag{5.9}$$

Функция рассеяния (фурье-образ (5.9)) имеет вид

$$h(x) = \frac{\sin x}{x} - 2\frac{\sin lx}{x}. \tag{5.10}$$

Всегда можно выбрать $l < 1$ так, чтобы функция (5.10) обратилась в нуль в любой наперед заданной точке x_0, поэтому ширина дифракционного максимума $\Delta x = 2x_0$ может быть сделана как угодно малой. Однако при этом с уменьшением Δx световой поток всё более равномерно распределяется по всем дифракционным максимумам функции $|h(x)|^2$, нулевой максимум перестаёт играть роль «главного» (в котором сосредоточена большая часть энергии), а критерий Релея перестаёт быть критерием, реально определяющим способность разрешения.

§ 5.3. Методы наблюдения фазовых структур

Многие объекты обладают высокой степенью прозрачности и почти не поглощают падающий свет. При просвечивании таких объектов плоской волной поле на выходе из объекта имеет одинаковую амплитуду, а изменения показателя преломления или толщины объекта сказываются лишь на вариациях фазы прошедшей волны: $f(x) = ae^{i\varphi(x)}$. Если оптическая система идеальна, то в плоскости изображения получаем $g(x) = f(x) = ae^{i\varphi(x)}$, и любой приёмник (фиксирующий интенсивность $I = gg^*$) даёт: $I = a^2 = \text{const}$, т. е. фазовый объект невидим.

Рассмотрим два метода преобразования фазовой модуляции поля в плоскости предмета в амплитудную (и, следовательно, в модуляцию интенсивности) в плоскости изображения, полагая, что вариации фазы при прохождения предмета малы ($\varphi(x) \ll 1$) и, следовательно, $f(x) \approx 1 + i\varphi(x)$.

1. Метод тёмного поля

Первое слагаемое в функции $f(x)$ (единица) даёт плоскую волну, распространяющуюся от предмета вдоль оптической оси. Эта волна фокусируется линзой $Л_1$ в точку — начало координат плоскости Ф, т. е. даёт δ-функцию в поле g_f. Метод тёмного поля состоит в устранении этой волны, причём остальные спектральные компоненты функции $f(x)$ — плоские волны, распространяющиеся по другим направлениям и фокусирующиеся в других точках частотной плоскости, должны быть пропущены без искажений. Достигнуть этого можно, установив в частотной плоскости (в начале координат) маленький непрозрачный экран, имеющий функцию пропускания (5.8). Считая систему идеальной ($H_0 \equiv 1$), получим поле на выходе частотной плоскости (являющееся спектром изображения):

$$G(u) = F(u) \cdot T(u) \approx [\delta(u) + i\Phi(u)] \cdot T(u) = i\Phi(u),$$

где $\Phi(u)$ — преобразование Фурье функции $\varphi(x)$; следовательно, поле в плоскости изображения $g(x) = i\varphi(x)$, а наблюдаемое распределение интенсивности $gg^* = \varphi^2(x)$, т. е. фазовые вариации объекта превратились в вариации интенсивности в плоскости изображения. Отметим, что аналогичный метод преобразования частотно-модулированного колебания в колебание, модулированное по амплитуде (устранение δ-функции в спектре), называется в радиотехнике *приёмом без несущей.*

2. Метод фазового контраста

Установим в частотной плоскости маску с пропускаемостью

$$T(u) = \begin{cases} i = e^{i\frac{\pi}{2}} & \text{при} \quad |u| \leqslant \varepsilon, \\ 1 & \text{при} \quad |u| > \varepsilon. \end{cases}$$

(Маска представляет собой маленькую прозрачную пластинку, вносящую фазовую задержку в $\pi/2$.) То есть вместо устранения несущего колебания — плоской волны, распространяющейся вдоль оптической оси, вводится фазовый сдвиг, равный $\pi/2$. Так как $F(u) = \delta(u) + i\Phi(u)$, то на выходе фурье-плоскости получаем

$$G(u) = F(u) \cdot T(u) = i\delta(u) + i\Phi(u).$$

Результирующее изображение $g(x) = i[1 + \varphi(x)]$, а его интенсивность

$$gg^* = 1 + 2\varphi(x). \tag{5.11}$$

Как следует из (5.11), вариации интенсивности в плоскости изображения линейно связаны с вариациями фазы предмета. Это — несомненное преимущество метода фазового контраста по сравнению с методом тёмного поля (где интенсивность пропорциональна $\varphi^2(x)$). Если ввести фазовую задержку в $\frac{3}{2}\pi$, то получим отрицательный фазовый контраст: $I = 1 - 2\varphi(x)$. Контраст получаемого изображения оказывается низким ($|\varphi(x)| \ll 1$), однако положение можно улучшить, если пластинка наряду с фазовой задержкой вносит и сильное ослабление:

$$T(u) = \begin{cases} R{\cdot}i & \text{при} \quad |u| \leqslant \varepsilon, \\ 1 & \text{при} \quad |u| > \varepsilon, \end{cases}$$

где $R \ll 1$. Тогда

$$G(u) = iR\delta(u) + i\Phi(u), \qquad g(x) = i[R + \varphi(x)],$$

и при $R \ll 1$ фон оказывается сильно ослабленным по сравнению с полезными вариациями сигнала: $gg^* = R^2 + 2R\varphi(x)$.

Аналогичный методу фазового контраста радиотехнический метод преобразования фазовой модуляции в амплитудную называется *приёмом с изменением фазы несущей.*

Задача. Ещё один метод наблюдения фазовых объектов — метод дефокусировки. Рассмотрите в качестве предмета синусоидальную фазовую решётку: $f(x) = e^{i\alpha\cos\Omega x}$ ($\alpha \ll 1$). Определите, на сколько нужно сдвинуть плоскость изображения в схеме рис. 5.1, чтобы вариации интенсивности в этой плоскости имели максимальный контраст.

§ 5.4. Мультипликация (размножение) изображений

Пусть поле во входной плоскости оптической системы описывается *финитной функцией* $f(x)$, т. е. $f(x) \equiv 0$ при $|x| \geqslant x_0$ (рис. 5.5).

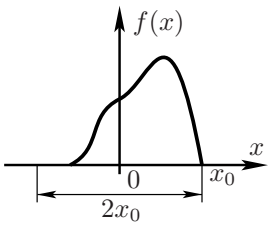


Рис. 5.5

Расположим в фурье-плоскости оптической системы маску-транспарант с функцией пропускания:

$$H(u) = \sum_n \delta(u - nu_0). \qquad (5.12)$$

Это — идеализированная «δ-решётка» с бесконечно узкими щелями.

Напомним, пространственная частота «u» связана с координатой x в фурье-плоскости равенством $u = \frac{kx}{f}$, поэтому период решётки $d_0 = \frac{fu_0}{k}$. Маска определяет частотную характеристику оптической системы, поэтому спектр изображения (пóля $g(x)$ в *выходной плоскости* оптической системы) есть

$$G(u) = F(u)\cdot H(u) = F(u)\sum_n \delta(u-nu_0) = \sum_n F(nu_0)\delta(u-nu_0). \quad (5.13)$$

Из сплошного спектра $F(u)$ входного сигнала маска пропускает лишь эквидистантный набор равноотстоящих по частоте гармоник $u_n = nu$. «Гребёнка» таких гармоник, согласно (1.31), представляет собой спектр периодического поля. Таким образом, в выходной плоскости получаем мультиплицированное (с периодом $d = \frac{2\pi}{u_0} = \frac{\lambda f}{d_0}$) изображение (рис. 5.6):

$$g(x) = \sum_n f(x - nd). \qquad (5.14)$$

Отдельные слагаемые мультиплицированного изображения не перекрываются, если $d > 2x_0$.

Использование вместо установленной в фурье-плоскости δ-решётки реальной решётки, имеющей щели конечной ширины b, приводит к тому, что число элементов размноженного изображения оказывается конечным. В этом случае функция пропускания решётки

$$H(u) = \sum_n P_\Delta(u - nu_0),$$

где

$$P_\Delta(u) = \begin{cases} 1 & \text{при} \quad |u| \leqslant \frac{\Delta}{2}, \\ 0 & \text{при} \quad |u| > \frac{\Delta}{2}. \end{cases}$$

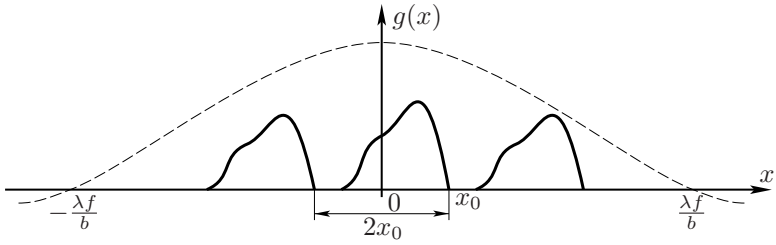


Рис. 5.6

Ширина щелей b связана с соответствующим частотным интервалом Δ: $\Delta = \frac{kb}{f}$. С учётом конечной ширины щелей спектр изображения есть

$$G(u) = F(u) \sum_n P_\Delta(u - nu_0) \simeq \sum_n F(nu_0) P_\Delta(u - nu_0). \qquad (5.15)$$

(В пределах ширины n-й щели спектр входного поля можно считать константой: $F(u) \simeq F(nu_0)$ при $|u - nu_0| < \frac{\Delta}{2}$.) Поле в выходной плоскости (обратное преобразование Фурье (5.15)) имеет вид

$$g(x) = \frac{\sin \frac{\Delta}{2} x}{x} \sum_n F(nu_0) e^{inu_0 x} = \left(\frac{\sin \frac{kb}{2f} x}{x} \right) \sum_n f(x - nd).$$

Огибающая — множитель $\left(\frac{\sin \frac{kb}{2f} x}{x}\right)$, показанный на рисунке пунктиром, описывает картину дифракции Фраунгофера на отдельной щели решётки. Он определяет размер области Δx в выходной плоскости, в которой наблюдаются повторяющиеся изображения входного поля $f(x)$: $|\Delta x| \lesssim \frac{\lambda f}{b}$ (полуширина главного максимума огибающей). Поэтому реально число повторяющихся изображений равно приблизительно $N \simeq \frac{2d}{b}$ (как число главных максимумов в картине дифракции Фраунгофера на решётке с периодом d и шириной щелей b).

В рассмотренных выше задачах использовались транспаранты с очень простой функцией пропускания, принимающей лишь два значения (0 и 1 или 0 и -1 и т. д.); это так называемые *бинарные фильтры*. Во многих задачах оптической фильтрации и оптической обработки информации возникает необходимость в транспарантах, пропускаемость которых варьируется сложным образом. Амплитудные вариации пропускаемости довольно просто создать обычными фотографическими методами: освещая разные участки фотопластинки светом

различной интенсивности, мы получим (после проявления) транспарант, почернение которого (и, следовательно, пропускаемость по амплитуде) пропорционально интенсивности света во время экспонирования. Однако получить транспарант со сложным законом вариаций фазы довольно трудно; необходимо, чтобы или толщина, или показатель преломления пластинки были некоторой заданной функцией координат, причём нужную точность обеспечить нелегко из-за малости длины световой волны. Трудности удалось преодолеть благодаря голографии — способу регистрации и последующего восстановления световой волны, при котором обеспечивается сохранение как амплитудной, так и фазовой информации (Д. Габор, 1948 г.).

§ 5.5. Голография

Пусть некоторый предмет (например, «скульптура», изображённая на рис. 5.7) освещается когерентным светом лазера, и пусть волна, отражённая предметом, создаёт в плоскости $z = 0$ поле с комплексной амплитудой $f_s(x,y) = a(x,y)e^{i\varphi(x,y)}$. Представим себе, что нам удалось создать в плоскости $z = 0$ то же распределение амплитуд и фаз колебаний (т. е. то же комплексное поле $f_s(x,y)$) *в отсутствие предмета*. Ясно, что в силу единственности решения уравнения Гельмгольца $\nabla^2 g + k^2 g = 0$, удовлетворяющего на плоскости $z = 0$ условию $g(x,y,z)|_{z=0} = f_s(x,y)$, в области $z > 0$ мы получим поле, тождественное полю, создаваемому реально присутствующим предметом: у наблюдателя, находящегося справа от плоскости $z = 0$, создаётся полное впечатление реально существующего предмета, находящегося именно на том месте (в области $z < 0$), в котором отражённая им волна создаёт в плоскости $z = 0$ поле $f_s(x,y)$.

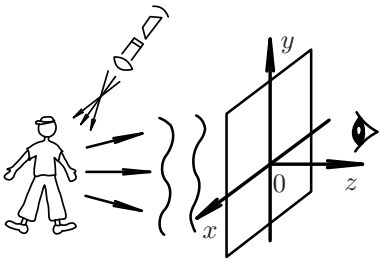


Рис. 5.7

Вопрос состоит в том, как реализовать на плоскости $z = 0$ комплексное поле $f_s(x,y)$ в отсутствие предмета. Попробуем сделать это с помощью фотопластинки, установленной в плоскости $z = 0$ в схеме эксперимента, изображённого на рис. 5.7. Фотопластинка (чувствительная только к интенсивности падающего света) зарегистрирует величину $I(x,y) = |f_s(x,y)|^2 = a^2(x,y)$. Это означает, что пропускаемость пластинки после проявления пропорциональна величине $a^2(x,y)$. Мы видим, что информация о фазе падающей волны теряется при квадратичном детектировании. Создастся

впечатление, используя для регистрации сре́ды, чувствительные только к интенсивности, вообще нельзя сохранить информацию о фазовом рельефе волны. Оказывается, это не так. Одна из возможностей — интерференционный метод записи: нужно к волне, идущей от предмета (предметная или сигнальная волна), добавить волну с известным распределением амплитуд и фаз (опорная волна). Записанная при фоторегистрации интерференционная картина сохраняет, как мы увидим, информацию как об амплитуде, так и о фазе предметной волны. Такой метод регистрации называется *голографией*, а фотопластинка, на которой регистрируется интерференционная картина, — *голограммой*.

Запись голограммы

Изменим в соответствии со сказанным выше схему эксперимента, изображённого на рис. 5.7.

На фотопластинку, установленную в плоскости $z = 0$, теперь падают две волны (рис. 5.8): предметная волна, создающая на пластинке поле $f_s(x,y)$, и опорная волна, поле которой $f_0(x,y)$. Суммарное поле есть $f_s(x,y) + f_0(x,y)$, а его интенсивность — интерференционная картина, регистрируемая на фотопластинке-голограмме, есть

$$I(x,y) = |f_s(x,y) + f_0(x,y)|^2. \tag{5.16}$$

Опуская некоторые детали, будем полагать, что после проявления пропускаемость голограммы по амплитуде пропорциональна интенсивности света при записи $t(x,y) \sim I(x,y)$, и мы получаем, используя (5.16):

$$t(x,y) = |f_s(x,y)|^2 + |f_0(x,y)|^2 + f_s(x,y)f_0^*(x,y) + f_s^*(x,y)f_0(x,y). \tag{5.17}$$

Реконструкция изображения

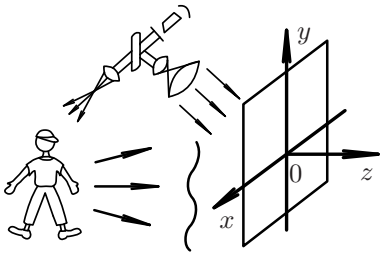


Рис. 5.8

Очевидно, что функция $t(x,y)$, определяемая формулой (5.17), действительная и положительная. Это означает, что проявленная голограмма является *чисто амплитудным пространственным модулятором*: при просвечивании она изменяет лишь амплитуду падающей на неё волны, не внося изменений в форму волнового фронта. Заметим, однако, что третье слагаемое в (5.17) содержит полную информацию как об амплитуде, так и о фазе предметной волны. Эту информацию можно восстановить и, следовательно,

восстановить (реконструировать) изображение предмета следующим образом. Установим проявленную пластинку-голограмму в плоскость $z = 0$ (где она экспонировалась при записи) и осветим голограмму волной, создающей на голограмме поле $f_p(x,y)$. Поле за голограммой (в плоскости $z = 0_+$) есть: $f_+(x,y) = f_p(x,y)t(x,y)$ или после использования (5.17):

$$f_+(x,y) = f_p(x,y)\Big[|f_0(x,y)|^2 + |f_s(x,y)|^2 + \\ + f_s^*(x,y)f_0(x,y) + f_s(x,y)f_0^*(x,y)\Big]. \quad (5.18)$$

Итак, поле в плоскости $z = 0_+$ представляется в виде суммы четырёх слагаемых, следовательно, и волну за голограммой (в области $z > 0$) можно представить в виде суммы четырёх волн. Исследуем физический смысл общего соотношения (5.18) (и каждого слагаемого), рассмотрев простой пример.

Голограмма точечного источника

Пусть предметом является точечный источник S (рис. 5.9), а в качестве опорной волны возьмём плоскую волну, нормально падающую на плоскость голограммы (плоскость $z = 0$). Мы имеем

$$f_0(x,y) = a_0 = \text{const};$$

$$f_s(x,y) = be^{ikr} = be^{ik\sqrt{x^2+y^2+z_0^2}}$$

(точечный источник создаёт на голограмме волну со сферическим фронтом; её амплитуду можно считать константой, если $x,y \ll z_0$).

Используя (5.17), получим

$$t(x,y) = \left|a_0 + be^{ikr}\right|^2 = a_0^2 + b^2 + a_0be^{ikr} + a_0be^{-ikr}. \quad (5.19)$$

Последнее равенство можно переписать в виде

$$t(x,y) = a_0^2 + b^2 + 2a_0b\cos kr = 2a_0^2(1 + \cos kr)$$

(будем считать для простоты амплитуды опорной и предметной волн равными $b = a_0$). Поскольку

$$r = \sqrt{z_0^2 + \rho^2} \approx z_0 + \frac{\rho^2}{2z_0} \qquad (\rho^2 = x^2 + y^2, \quad \text{см. рис. 5.9}),$$

то

$$t(\rho) = 2a_0^2\left[1 + \cos\left(kz_0 + \frac{k}{2z_0}\rho^2\right)\right]. \quad (5.20)$$

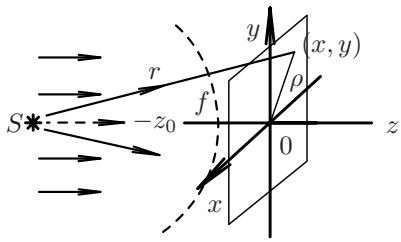


Рис. 5.9

Пластинка с функцией пропускания (5.20) называется *зонной решёткой Габора*.

Для восстановления изображения осветим голограмму с функцией пропускания (5.19) плоской, нормально падающей волной единичной амплитуды: $f_p = 1$, рис. 5.10. Поле за голограммой (в плоскости $z = 0_+$) имеет вид

$$f_+ = 2a_0^2 + a_0^2 e^{ikr} + a_0^2 e^{-ikr}.$$

Итак, волна за голограммой является суммой трёх волн: каждое слагаемое в поле на границе $z = 0_+$ ответственно за появление «своей» волны в области $z > 0$. Первое слагаемое $2a_0^2$ — постоянная составляющая — представляет собой поле плоской волны с волновым вектором, перпендикулярным плоскости $z = 0$, и, следовательно, за голограммой мы имеем плоскую волну (распространяющуюся вдоль оси z). Второе слагаемое $a_0^2 e^{ikr}$ тождественно полю $f_s(x,y)$, создаваемому на плоскости $z = 0$ источником S при записи голограммы (расходящаяся сферическая волна). Наблюдателю, находящемуся справа от плоскости $z = 0$, соответствующая волна будет казаться исходящей из точки S' (рис. 5.10) (хотя в процессе реконструкции самого источника нет). Это мнимое изображение предмета.

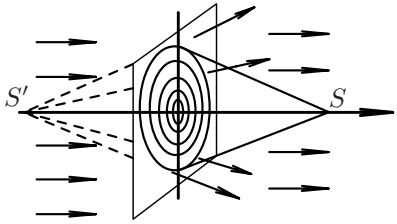


Рис. 5.10

Рассмотрим теперь третье слагаемое $a_0^2 e^{-ikr}$. Это слагаемое комплексно сопряжено со вторым. Значит, фазовый фронт третьей волны «вывернут наизнанку» относительно фронта второй волны. Действительно, поскольку полное поле $f_+(x,y)$ на выходе голограммы должно иметь плоский фазовый фронт, совпадающий с плоскостью голограммы (так как $f_+(x,y)$ — действительная положительная функция), то сферическое фазовое искривление, создаваемое вторым слагаемым, должно быть скомпенсировано третьим слагаемым. Итак, третье слагаемое ответственно за появление сферической волны, *сходящейся* к точке S — действительному изображению предмета. Яркую точку S можно наблюдать на экране, установленном в плоскости $z = z_0 = \text{const}$ справа от голограммы.

Итак, зонная решётка, освещаемая плоской волной, создаёт в области $z > 0$ три волны: плоскую, сходящуюся сферическую и расходящуюся сферическую; таким образом, она работает одновременно и как плоскопараллельная пластинка, и как собирающая линза с фокусным расстоянием z_0, и как рассеивающая линза (с фокусом $-z_0$). Мы рассмотрели в качестве примера процесс получения голограммы и последующего восстановления изображения очень простого предмета — точечного источника света. Но, согласно принципу Гюйгенса–Френеля, любое поле $f(x,y)$ может быть представлено в виде суперпозиции сферических волн, излучаемых точечными источниками, поэтому в общем соотношении (5.18) физический смысл каждого из слагаемых остаётся прежним: третье слагаемое в (5.18) создаёт мнимое изображение предмета (наблюдатель, глядя на голограмму, как в окно, будет видеть за «окном» предмет). Четвёртое слагаемое в (5.18) приводит к появлению действительного изображения, которое можно наблюдать на экране, расположив его в нужном месте справа от голограммы. Здесь следует ещё раз подчеркнуть, что связь между полем, в котором экспонируется фотопластинка-голограмма при записи, и полем, возникающим за голограммой при восстановлении, нелинейна; поэтому может показаться, что если поле предмета представляется в виде линейной суперпозиции сферических волн, то это не означает, что при восстановлении мы получим линейную суперпозицию сферических волн, создающую изображение. В действительности, нас интересует не полное поле $f_+(x,y)$ за голограммой, а лишь одна его часть (третье слагаемое в (5.18)), а это слагаемое связано с полем $f_s(x,y)$ линейно. Именно поэтому рассмотренный частный пример обобщается на случай произвольного предмета; при этом сложный интерференционный узор, записанный на голограмме, можно представить суперпозицией зонных решёток Френеля, каждая из которых является результатом интерференции опорной волны и сферической волны, создаваемой какой-либо точкой предмета.

Чем отличается голографическое изображение от фотографического?

При фотографировании каждая точка предмета посылает расходящуюся сферическую волну, которая с помощью объектива фотоаппарата фокусируется в небольшое пятнышко на фотопластинке. Так как яркость разных точек объекта различна, то различна и интенсивность света, падающего на разные участки фотопластинки. Почернение разных участков проявленной фотопластинки (которое пропорционально интенсивности) повторяет распределение яркости на предмете — мы

получаем негативное изображение. Отметим три особенности фотографического изображения.

Во-первых, на фотопластинке фиксируется лишь распределение интенсивности, и, следовательно, фотоизображение содержит не полную информацию о предмете (нет распределения фаз и, в частности, нет рельефа предмета). Во-вторых, на каждом небольшом участке фотопластинки получается изображение лишь небольшого («сопряженного») участка предмета. В-третьих, трёхмерные объекты регистрируются в виде плоских двумерных изображений.

Голограмма, как мы видели, сохраняет полную информацию об объекте; возникающий на ней интерференционный узор зависит как от амплитуды, так и от фазы предметной волны.

Наблюдатель, который при восстановлении смотрит сквозь голограмму (как в окно), видит как бы реальный предмет; так, изменяя положение, наблюдатель может увидеть те детали объекта, которые не были видны ранее — голографическое изображение трёхмерно. Важнейшее свойство голограммы состоит в том, что любой её малый участок содержит информацию о всем объекте: ведь поле в каждой точке голограммы является суперпозицией полей, посылаемых всеми точками предмета (и опорной волны). Другими словами, интерференционная картина на каждом небольшом участке голограммы содержит информацию об амплитуде и фазе излучения всех точек предмета, поэтому изображение может быть восстановлено с помощью небольшого осколка голограммы, полученной при записи (рис. 5.11). Это значительно увеличивает надёжность хранения записанной на голограмме информации.

Голограмма Лейта–Упатниекса

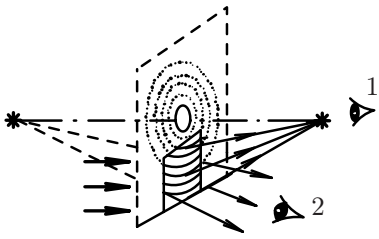


Рис. 5.11

Беда в том, что при наблюдении голографических изображений по схеме рис. 5.10, действительное и мнимое изображения создают взаимные помехи: изображения S' и S находятся на одной прямой с точкой, в которой расположен глаз наблюдателя; «шумовой» фон создаётся также постоянной составляющей — плоской волной, бегущей вдоль оси z. Положение можно исправить, используя для восстановления периферийный участок голограммы (как показано на рис. 5.11), при этом для наблюдения действительного и мнимого изображений наблюдатель дол-

жен находиться в разных положениях (1 и 2 на рис. 5.11), и взаимных помех нет.

Более традиционным является способ, предложенный Лейтом и Упатниексом. В этом способе опорной волной служит плоская волна, *наклонно падающая на голограмму*: $f_0(x,y) = a_0 e^{-iv_0 y}$. Функция пропускания голограммы имеет в этом случае вид

$$t(x,y) = \left|a_0 e^{-iv_0 y} + f_s(x,y)\right|^2 = \\ = a_0^2 + |f_s(x,y)|^2 + a_0 f_s(x,y) e^{iv_0 y} + a_0 f_S^*(x,y) e^{-iv_0 y}, \quad (5.21)$$

или через амплитуду $a(x,y)$ и фазу $\varphi(x,y)$ предметной волны:

$$t(x,y) = a_0^2 + a^2(x,y) + 2a_0 a(x,y)\cos[v_0 y - \varphi(x,y)]. \quad (5.22)$$

Как следует из (5.22), информация об амплитуде и фазе предметной волны заключена в модуляции «несущего колебания» $a_0 \cos v_0 y$ («частота» которого определяется «пространственной частотой» опорного пучка $v_0 = k \sin\alpha$). Функция $a(x,y)$ даёт амплитудную модуляцию (т. е. меняет контраст интерференционных полос), а функция $\varphi(x,y)$ осуществляет фазовую модуляцию несущего колебания, т. е. определяет форму и вариации частоты полос.

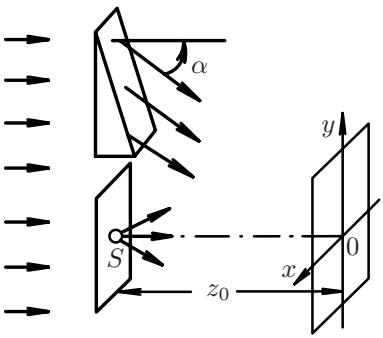


Рис. 5.12

Проследим, каким образом введение несущего колебания (наклонного опорного пучка) обеспечивает возможность независимого наблюдения действительного и мнимого изображений (на примере голограммы точечного источника). На рис. 5.12 источником является маленькое отверстие S в непрозрачном экране. Заменим для простоты расчётов сферический волновой фронт параболическим, тогда с точностью до постоянного множителя имеем

$$f_s(x,y) = b e^{i\frac{k}{2z_0}(x^2+y^2)},$$

и мы получаем, используя (5.21):

$$t(x,y) = \left|a_0 e^{-iv_0 y} + b e^{i\frac{k}{2z_0}(x^2+y^2)}\right|^2.$$

Для восстановления изображения используем плоскую, нормально падающую волну: $f_p(x,y) = 1$. Тогда поле за голограммой (в плоскости $z = 0_+$) есть

$$f_+(x,y) = \left| a_0 e^{-iv_0 y} + be^{i\frac{k}{2z_0}(x^2+y^2)} \right|^2 .$$

Легко видеть, что слагаемые, ответственные за появление мнимого и действительного изображений, можно представить в виде

$$\begin{aligned} f_3(x,y) &= f_s(x,y) f_0^*(x,y) = a_0 b e^{-i\frac{z_0}{2k}v_0^2} \cdot e^{i\frac{k}{2z_0}\left[(y+z_0\frac{v_0}{k})^2+x^2\right]}, \\ f_4(x,y) &= f_s^*(x,y) f_0(x,y) = a_0 b e^{i\frac{z_0}{2k}v_0^2} \cdot e^{-i\frac{k}{2z_0}\left[(y+z_0\frac{v_0}{k})^2+x^2\right]}. \end{aligned} \qquad (5.23)$$

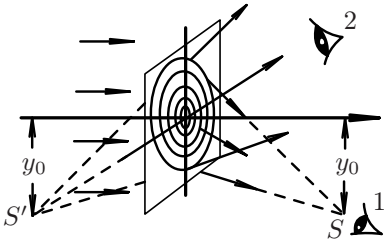


Рис. 5.13

Первое из равенств (5.23) представляет собой поле сферической волны (в параболическом приближении), источник которой находится в точке с координатами $x = 0$, $y_0 = -z_0\frac{v_0}{k} = -z_0 \sin\alpha$, разумеется, на расстоянии z_0 от голограммы, слева от неё — это мнимое изображение предмета. Соответственно действительное изображение находится в точке с теми же координатами $x = 0$, $y_0 = -z_0 \sin\alpha$ на расстоянии z_0 справа от голограммы (рис. 5.13). Мы видим, что введение наклонного опорного пучка приводит к сдвигу действительного и мнимого изображений таким образом, что их можно наблюдать в разных положениях (1 и 2 на рис. 5.13), в которых они не создают взаимных помех.

При восстановлении голограммы произвольного предмета слагаемые, дающие мнимое и действительное изображения, имеют вид

$$f_3 = f(x,y)e^{iv_0 y}, \qquad f_4 = f^*(x,y)e^{-iv_0 y}.$$

Изображения могут наблюдаться независимо лишь в том случае, если несущая частота достаточно велика (т. е. достаточно велик угол наклона опорного пучка). Действительно, только в этом случае пространственные спектры слагаемых f_3 и f_4 не перекрываются.

Задача. Показать, исходя из критерия неперекрывания пространственных спектров f_3 и f_4 (а также слагаемого $f_2 = |f_s(x,y)|^2$, что условие, при котором можно без помех наблюдать действительное и мнимое изображения, есть $\sin\alpha_{\min} \simeq \frac{3\lambda}{\delta}$, где δ — минимальный размер деталей предмета.

Разрешающая способность голограммы

До сих пор мы никаким образом не учитывали влияние конечного размера голограммы на восстановленные изображения. Именно поэтому голографические изображения точечного источника являются точками. Что изменится, если принять во внимание конечные размеры голограммы? Вместо (5.17) для функции пропускания голограммы следует написать: $t_1 = p_a(x) \cdot t(x)$, где $t(x)$ определяется соотношением (5.17), и соответственно для поля на выходе голограммы (в плоскости $z = 0_+$) при освещении её плоской нормально падающей волной имеем $f_+(x) = a_0 p_a(x) t(x)$.

Рассмотрим, например, слагаемое, ответственное за появление действительного изображения $f_4 = a_0 p_a(x) e^{-ikr}$ (с мнимым изображением всё аналогично). Функция f_4 представляет собой поле со сферическим фазовым фронтом, причём протяжённость фронта ограничена размерами голограммы: $f_4 = 0$ при $|x| > a$. Можно провести аналогию с линзой конечного размера. Функция пропускания такой линзы есть $t(x) = p_a(x) e^{-i\frac{k}{2f}x^2}$ (см. § 3.4). Используя параболическое приближение, функцию f_4 можем записать в виде $f_4 = p_a(x) e^{-i\frac{k}{2z_0}x^2}$ (с точностью до постоянного множителя). Таким образом, поле f_4 аналогично полю, возникающему за линзой конечного размера при освещении её плоской волной. Мы знаем, что в фокусе линзы при этом получается яркое пятнышко (функция рассеяния), размер которого определяется формулой $\delta x \simeq \frac{\lambda f}{a}$, и, следовательно, поле f_4 даёт в качестве действительного изображения не точку, а пятно с поперечным размером $\delta x \simeq \frac{\lambda z_0}{a}$ (это пятно возникает на расстоянии z_0 справа от голограммы). Ясно, что если для восстановления используется малый осколок голограммы, то это приводит к ухудшению разрешающей способности (к увеличению размеров изображения точечного источника), а в общем случае — к размытию мелких деталей в изображении предмета.

Возникает вопрос: сколько всего точек можно записать на голограмме так, чтобы в восстановленном изображении они были разрешены? Количество таких точек определяет максимальный объём информации в двоичной системе единиц (в «битах»). Поскольку угловое расстояние между точками должно быть не меньше $\Delta\varphi_{\min} \approx \frac{\lambda}{a}$, то максимальное число точек есть $\left(\frac{\pi}{\Delta\varphi}\right)^2 \approx \left(\frac{a}{\lambda}\right)^2$. На голограмме размером 1×1 см при $\lambda = 5 \cdot 10^{-5}$ см можно записать $\sim 10^9$ двоичных единиц информации. Это огромная ёмкость, и хранение больших объёмов информации является одним из перспективнейших применений голографии.

В заключение этого параграфа отметим ещё одну интересную возможность, связанную с голографией: кодирование информации, за-

писанной на голограмму. Введя в канал опорного луча специальную маску, можно сделать его волновой фронт чрезвычайно сложным. При этом для восстановления изображения нужно иметь точную копию этой маски. Поместив на пути опорного луча маску с пропускаемостью $e^{i\psi(x,y)}$, получим

$$t_з = f(x,y)e^{i[v_0 y - \psi(x,y)]}.$$

Если теперь ту же маску использовать в процессе восстановления: $f_p = e^{i\psi(x,y)}$, то получаем

$$f_з = t_з(x,y)f_p = f(x,y)e^{iv_0 y}.$$

В качестве маски со сложным фазовым законом пропускания можно взять, например, матовое стекло. Ясно, что, не имея точной копии маски (а подобрать её практически невозможно), нельзя восстановить изображение.

С задачей кодирования информации тесно связана задача распознавания полезного сигнала (см. также § 5.7). Используем в качестве опорной волны полезный сигнал $f_0(x,y)$, а в качестве предмета при записи голограммы — точечный источник света. Ясно, что если в волне, освещающей проявленную голограмму, присутствует полезный сигнал, т. е. $f_p = f_0(x,y) + \sum f_n(x,y)$, где $f_n(x,y)$ — шумы или какие-либо посторонние сигналы, то мы получим восстановленное изображение точечного источника — яркую точку, по появлению которой можно судить о присутствии полезного сигнала в волне $f_p(x,y)$. Ниже рассмотрен ряд задач, в которых, с одной стороны, маски-транспаранты, полученные голографическим способом, используются для решения ряда задач пространственной фильтрации, а с другой стороны, принципы пространственной фильтрации используются для формирования голограмм без опорного пучка.

§ 5.6. Синтез оптического фильтра с заданным импульсным откликом

Требуется создать оптическую систему с заданным законом преобразования входного сигнала в выходной, т. е. с заданным импульсным откликом $h(x,y)$. Будем полагать, что в нашем распоряжении имеется транспарант с функцией пропускания $h(x,y)$ (следует отметить, что для решения многих задач пространственной фильтрации необходимо иметь оптическую систему с действительным импульсным откликом, поэтому транспарант с функцией пропускания $h(x,y)$ легко может быть изготовлен, например, обычным фотографическим методом).

На рис. 5.14 изображена схема получения голограммы, которая затем должна быть установлена в частотной плоскости оптической системы. Как видно из рис. 5.14, поле опорной волны на голограмме есть $f_0 = a_0 e^{-iv_0 y}$, где $v = k\sin\theta$, а поле предметной волны $f_s(x,y)$ есть преобразование Фурье (с аргументами $\frac{kx}{f}$, $\frac{ky}{f}$) функции $h(x,y)$:

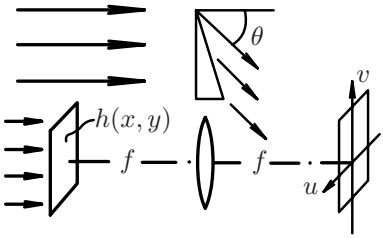


Рис. 5.14

$f_s(x,y) = H(\frac{kx}{f}, \frac{ky}{f})$. Введя в плоскости голограммы «координаты» $u = \frac{kx}{f}$, $v = \frac{ky}{f}$, получаем для функции пропускания голограммы:

$$T(u,v) = \left|H(u,v) + a_0 e^{-if\sin\theta\cdot v}\right|^2.$$

Раскрывая последнее выражение, находим

$$T(u,v) = a_0^2 + |H(u,v)|^2 + a_0 H(u,v) e^{if\sin\theta\cdot v} + a_0 H^*(u,v) e^{-if\sin\theta\cdot v}. \quad (5.24)$$

Полученная описанным способом голограмма называется *фильтром Ван-дер-Люгта*.

Итак, установим голограмму-маску с функцией пропускания (5.24) в частотную плоскость Ф схемы Катрона (рис. 5.1), и пусть на вход системы поступает подлежащий преобразованию сигнал $f(x,y)$. Тогда поле на входе в частотную плоскость есть $F(u,v)$, а поле на выходе частотной плоскости (на выходе из маски с функцией пропускания (5.24)) имеет вид $G(u,v) = F(u,v)T(u,v)$ (это поле, как известно, является преобразованием Фурье искомого выходного сигнала, § 5.1). Используя (5.24), находим

$$G(u,v) = a_0^2 F(u,v) + F(u,v)|H(u,v)|^2 + \\ + a_0 F(u,v) H(u,v) e^{if\sin\theta\cdot v} + a_0 F(u,v) H^*(u,v) e^{-if\sin\theta\cdot v}. \quad (5.25)$$

Преобразование Фурье (и, следовательно, сам выходной сигнал $g(x,y)$) состоит из четырёх слагаемых. Нас интересует третье слагаемое в спектре $G_3(u,v) = a_0 F(u,v) H(u,v) e^{if\sin\theta\cdot v}$ и соответствующее слагаемое в выходном сигнале:

$$g_3(x,y) = a_0 \int F(u,v) H(u,v) e^{if\sin\theta\cdot v} e^{i(ux+vy)}\, du dv. \quad (5.26)$$

Сдвинув систему координат в выходной плоскости по оси y на величину $y_0 = -f\sin\theta$, т. е. введя переменную $y_1 = y + f\sin\theta$, получим вместо (5.26)

$$g_3(x,y_1) = a_0 \int F(u,v)H(u,v)e^{i(ux+vy_1)}\,dudv. \tag{5.27}$$

Согласно (5.27), спектр составляющей $g_3(x,y_1)$ есть произведение спектра входного сигнала $F(u,v)$ и фурье-образа $H(u,v)$ функции $h(x,y)$, и, следовательно, сам сигнал g_3 является свёрткой входного сигнала $f(x,y)$ и пропускаемости транспаранта $h(x,y)$:

$$g_3(x,y_1) = f(x,y_1) \otimes\otimes h(x,y_1) = \iint f(\xi,\eta)h(x-\xi,y-\eta)\,d\xi d\eta. \tag{5.28}$$

Подчеркнём, что в системе координат, центр которой находится на оптической оси, интересующий нас сигнал смещен по оси y на величину $y_0 = -f\sin\theta$.

Четвёртому слагаемому в спектре соответствует слагаемое в выходном сигнале $g_4(x,y)$:

$$g_4(x,y) = a_0 \iint F(u,v)H^*(u,v)e^{-if\sin\theta\cdot v}e^{i(ux+vy)}\,dudv.$$

В системе координат, сдвинутой по оси y в противоположную сторону на ту же величину y_0: $y_2 = y - f\sin\theta$, получаем

$$g_4(x,y_2) = a_0 \iint F(u,v)H^*(u,v)e^{i(ux+vy_2)}\,dudv. \tag{5.29}$$

Поскольку $H^*(u,v)$ является преобразованием Фурье функции $h^*(-x,-y)$, то вместо (5.29) имеем

$$g_4(x,y_2) = f(x,y_2) \otimes\otimes h^*(-x,-y_2) = \iint f(\xi,\eta)h^*(\xi-x,\eta-y_1)\,d\xi d\eta. \tag{5.30}$$

Если угол наклона опорного пучка θ достаточно велик, то функции $g_3(x,y)$ и $g_4(x,y)$, будучи сдвинутыми в разные стороны по оси y на величину $y_0 = f\sin\theta$, не перекрываются и могут наблюдаться независимо. Необходимо потребовать также, чтобы эти функции не перекрывались с фоном, создаваемым первыми двумя слагаемыми в поле $g(x,y)$.

Задача. Показать, что условием полного разделения слагаемых g_3, g_4, а также фона, обусловленного слагаемыми g_1 и g_2, является неравенство: $f\sin\theta \gtrsim \Delta y_f + 3/2\Delta y_h$, где Δy_f и Δy_h — соответственно размеры функций $f(x,y)$ и $h(x,y)$ по оси y.

Можно сделать следующий вывод: выходной сигнал в окрестности точки $(0, -f\sin\theta)$ определяется формулой (5.28), т. е. при наблюдении сигнала в окрестности точки $(0, -f\sin\theta)$ наша система имеет импульсный отклик $h(x,y)$, и, следовательно, задача синтеза системы с заданным импульсным откликом решена.

Попутно мы синтезировали фильтр с откликом $h^*(-x, -y)$ (т. е. с частотной характеристикой $H^*(u,v)$), поскольку при наблюдении сигнала в окрестности точки $(0, -f\sin\theta)$ последний определяется формулой (5.30).

§ 5.7. Принцип согласованной фильтрации в оптике и задача распознавания образов

Имеется ряд практических задач, решение которых может быть основано на принципе согласованной фильтрации. В оптической локации и связи среди множества различных сигналов, поступающих на вход приемной системы, требуется обнаружить сигнал заранее известной формы или выделить сигнал известной формы на фоне сильных шумов; криминалистов интересует задача отыскания среди множества имеющихся в картотеке отпечатков пальцев заданного отпечатка (принадлежащего данному человеку), лингвистов может интересовать частота повторения какой-либо определённой буквы в тексте и т. д.

Что же представляет собой согласованный фильтр в оптике? Пусть сигнал $f_0(x,y)$ поступает на вход системы Катрона (рис. 5.1). Тогда поле в частотной плоскости есть $F_0(u,v)$ — преобразование Фурье сигнала $f_0(x,y)$ (напомним, что $u = \frac{kx}{f}$, $v = \frac{ky}{f}$). Установим в частотной плоскости транспарант с пропускаемостью

$$T(u,v) = F_0^*(u,v). \tag{5.31}$$

Поле на выходе транспаранта (спектр выходного сигнала) определяется формулой

$$G(u,v) = F_0(u,v)\cdot T(u,v) = |F_0(u,v)|^2,$$

и, таким образом, *волна на выходе частотной плоскости имеет плоский фазовый фронт*, совпадающий с плоскостью транспаранта, поскольку $G(u,v)$ — положительная действительная функция. Эта волна сфокусируется линзой $Л_2$ в начало координат выходной плоскости, где образуется яркое светящееся пятнышко. Смысл установки транспаранта с пропускаемостью (5.31) состоит в том, что такой транспарант *в точности компенсирует фазовые искривления волны*, создаваемой в частотной плоскости сигналом $f_0(x,y)$.

Для другого входного сигнала $f(x,y)$ фазовые искривления, вообще говоря, не будут скомпенсированы маской (5.31), волновой фронт на выходе частотной плоскости не будет плоским, и соответствующая волна не сфокусируется в яркое пятнышко в выходной плоскости.

Частотная характеристика оптической системы, в частотной плоскости которой помещён транспарант (5.31), очевидно, есть

$$H(u,v) = T(u,v) = F_0^*(u,v).$$

Таким образом, можно сказать, что оптическая система с частотной характеристикой $H(u,v) = F_0^*(u,v)$ «согласована» с входным сигналом $f_0(x,y)$ в том смысле, что выходной сигнал в точке ($x = 0$, $y = 0$) оказывается максимальным, если на вход подаётся сигнал $f_0(x,y)$. Изменение формы входного сигнала приводит к уменьшению отклика в точке ($x = 0$, $y = 0$) (разумеется, речь идёт о сравнении сигналов разной формы, но имеющих одну и ту же энергию: $\int |f(x,y)|^2\,dxdy = \text{const}$). Отметим, что если согласованный сигнал $f_0(x,y)$ сдвинуть во входной плоскости, не меняя его формы (т. е. подать на вход $f(x,y) = f_0(x - x_0, y - y_0)$), то яркая точка на выходе образуется не в начале координат, а в точке (x_0, y_0) (докажите это!).

Приведённые выше качественные соображения вовсе не являются доказательством того, что среди множества сигналов $f(x,y)$ с фиксированной энергией $\int |f(x,y)|^2\,dxdy = E_0 = \text{const}$, максимальный отклик в точке $x = 0$, $y = 0$ (т. е. максимальное значение $g(0,0)$) даёт сигнал $f_0(x,y)$, спектр которого связан с частотной характеристикой системы $H(u,v)$ равенством $F_0(u,v) = H^*(u,v)$. Больше того, на первый взгляд это утверждение кажется неверным. Можно было бы рассуждать следующим образом: пусть сигнал $f_0(x,y)$ создаёт в частотной плоскости системы поле $F_0(u,v) = A(u,v)e^{i\psi(u,v)}$. Поместим в частотной плоскости маску с пропускаемостью $T_1(u,v) = e^{-i\psi(u,v)}$, и, следовательно, наша система имеет частотную характеристику $H_1(u,v) = T_1(u,v) = e^{-i\psi(u,v)}$. Это совершенно прозрачная маска, амплитудное пропускание которой $a(u,v) \equiv 1$, т. е. она полностью пропускает весь световой поток, компенсируя лишь фазовые искривления падающей волны. Волна на выходе частотной плоскости имеет комплексную амплитуду $G_1(u,v) = F_0(u,v)T_1(u,v) = A(u,v)$, т. е. также имеет плоский фазовый фронт, и поэтому должна хорошо фокусироваться в начало координат выходной плоскости. Причём создаётся впечатление, что маска с пропускаемостью $T(u,v) = e^{-i\psi(u,v)}$ «лучше», чем маска $T(u,v) = F^*(u,v) = A(u,v)e^{-i\psi(u,v)}$, поскольку первая в отличие от второй совершенно не поглощает света. В действительности, следует помнить, что поля на выходе частотной

плоскости $G_1(u,v) = A(u,v)$ и $G(u,v) = F_0(u,v) \cdot T(u,v) = A^2(u,v)$ различны (имеют различный закон изменения амплитуды по фронту волны) и, следовательно, различным образом фокусируются в выходную плоскость, хотя интегральный световой поток в первом случае больше. Строгое доказательство принципа согласованной фильтрации основано на неравенстве Шварца. Выходной сигнал системы можно представить в виде

$$g(x,y) = \iint F(u,v) e^{i(ux+vy)}\, du dv$$

и, следовательно,

$$g(0,0) = \iint F(u,v) H(u,v)\, du dv.$$

Согласно неравенству Шварца:

$$|g(0,0)|^2 = \left| \iint F(u,v) H(u,v)\, du dv \right|^2 \leqslant$$

$$\leqslant \int |F(u,v)|^2\, du dv \cdot \int |H(u,v)|^2\, du dv. \quad (5.32)$$

В неравенстве (5.32) равенство достигается при $F(u,v) = cH^*(u,v)$, где c — произвольная комплексная постоянная.

Рассмотрим в качестве примера оптическое устройство, обеспечивающее поиск заданной информации и распознавание образов (например, поиск буквы «И» в тексте, изображённом на рис. 5.15б). Прежде всего, необходимо по схеме Ван-дер-Люгта (рис. 5.14) синтезировать оптический фильтр, импульсный отклик которого является образом буквы «И». Другими словами, транспарант с функцией пропускания $h(x,y)$, изображённый в схеме рис. 5.14, должен представлять собой непрозрачный экран с отверстием в виде буквы «И» (рис. 5.15а). Нас интересует четвёртое слагаемое в функции пропускания полученной голограммы (5.24). Оно в нашем случае будет иметь вид: $a_0 u^*(u,v) e^{-if \sin\theta\, v}$, где через $u(u,v)$ обозначен спектр буквы «И». Установив голограмму в частотную плоскость системы Катрона, подадим текст, изображённый на рис. 5.15б, на вход системы. Выходной сигнал будем наблюдать в окрестности точки ($x = 0$, $y = f \sin\theta$). Он имеет вид, изображённый на рис. 5.15в. Яркие точки, возникающие в плоскости изображения, указывают на наличие буквы «И» в соответствующих местах текста.

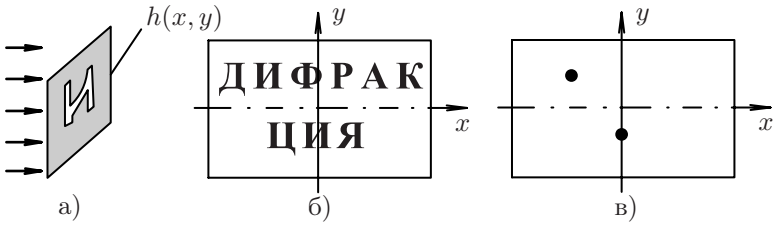


Рис. 5.15

§ 5.8. Устранение аберраций в оптической системе

Всякая реальная оптическая система вносит не только дифракционные искажения, связанные с конечностью размеров объективов, но также и искажения, обусловленные несовершенством их формы (аберрации). Аберрации приводят к тому, что частотная характеристика системы, даже при использовании бесконечно больших линз, не равна тождественно единице, и, следовательно, спектр изображения отличен от спектра предмета. Пусть частотная характеристика системы, имеющей аберрации, есть $H(u,v)$. Действие искажений, вносимых объективами, эквивалентно действию транспаранта с функцией пропускания $H(u,v)$, установленного в частотной плоскости. Таким образом, оптическую систему с аберрациями можно заменить идеальной системой с транспарантом $H(u,v)$ в частотной плоскости. Из этой посылки мы и будем исходить.

Установим в частотной плоскости оптической системы два транспаранта: один из них, изготовленный голографическими методами, с пропускаемостью $T_1(u,v) = H^*(u,v)$, а другой, чисто действительный транспарант, изготовленный обычными фотографическими методами, с пропускаемостью $T_2(u,v) = |H(u,v)|^{-2}$. Пропускаемость двух совмещённых транспарантов, очевидно, равна

$$T(u,v) = T_1(u,v) \cdot T_2(u,v) = \frac{H^*(u,v)}{|H(u,v)|^2}. \tag{5.33}$$

Маска с пропускаемостью (5.33) обладает замечательным свойством: она автоматически компенсирует аберрационные искажения, или, другими словами, искажения, вносимые «гипотетическим» транспарантом $H(u,v)$. Действительно, полная пропускаемость (маски (5.33) и «транспаранта» $H(u,v)$) равна

$$T_0(u,v) = H(u,v) \cdot \frac{H^*(u,v)}{|H(u,v)|^2} \equiv 1.$$

Это и есть частотная характеристика исправленной системы:

$$H_0(u,v) = T_0(u,v) \equiv 1.$$

Рассмотрим теперь задачу исправления уже полученного, «испорченного» аберрациями изображения (пример так называемой «апостериорной», послеопытной обработки информации). Поместим такое «испорченное» изображение, имеющее спектр

$$F_1(u,v) = F(u,v)H(u,v),$$

на вход нашей оптической системы, установив в частотной плоскости транспарант с функцией пропускания $T^2(u,v)$, где $T(u,v)$ определяется формулой (5.33) (этот транспарант — просто две совмещённые маски $T(u,v)$). При отсутствии этого транспаранта мы получили бы изображение со спектром $F_1(u,v) \cdot H(u,v) = F(u,v)H^2(u,v)$ (аберрации вторично портят уже испорченное изображение, или по-другому, спектр испорченного предмета на входе в частотную плоскость есть $F_1(u,v)$, а на выходе частотной плоскости (где установлен гипотетический транспарант $H(u,v)$) имеем $F_1(u,v) \cdot H(u,v)$). Полная пропускаемость (маски $T^2(u,v)$ и гипотетического «транспаранта» $H(u,v)$) и есть частотная характеристика нашей системы:

$$H_0(u,v) = T^2(u,v) \cdot H(u,v) = \frac{H^*(u,v) \cdot H^*(u,v)}{|H(u,v)|^4} \cdot H(u,v) = \frac{H^*(u,v)}{|H(u,v)|^2}.$$

Спектр скорректированного изображения есть произведение спектра «предмета» $F_1(u,v)$ и функции $H_0(u,v)$, и мы получаем

$$G(u,v) = F_1(u,v) \cdot \frac{H^*(u,v)}{|H(u,v)|^2} = F(u,v).$$

Таким образом, маска с пропускаемостью $T^2(u,v)$ автоматически компенсирует искажения, вносимые оптической системой на двух этапах: при формировании испорченного изображения и при преобразовании этого испорченного изображения в исправленное изображение со спектром $F(u,v)$.

§ 5.9. Голограмма без опорного пучка

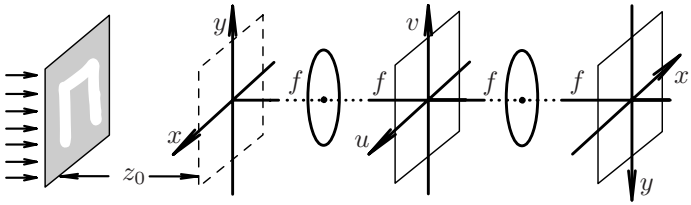


Рис. 5.16

Идеи пространственной фильтрации могут быть использованы для получения голограммы без специального опорного пучка. Одна из возможных схем формирования такой голограммы изображена на рис. 5.16.

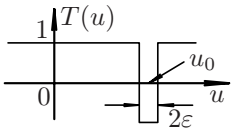


Рис. 5.17

Предмет, расположенный на расстоянии z_0 от входной плоскости оптической системы, создаёт в этой плоскости поле $f(x,y)$ и, следовательно, поле в частотной плоскости Ф есть $F(u,v)$. Рассмотрим для простоты случай, когда спектр $F(u,v)$ содержит δ-функцию на частоте $(u_0,\ v_0)$: $F(u,v) = F_0(u,v) + A\delta(u-u_0,v-v_0)$. Установим в частотной плоскости маску, пропускаемость которой равна

$$T(u,v) = \begin{cases} -1 & \text{при } (u-u_0)^2 + (v-v_0)^2 \leqslant \varepsilon^2, \\ 1 & \text{при всех остальных } u,\ v \end{cases}$$

(маленькая прозрачная пластинка, установленная в точке $(u_0,\ v_0)$ частотной плоскости), вносит фазовую задержку в π. Очевидно, поле на выходе частотной плоскости есть

$$G(u) = F_0(u,v) - A\delta(u-u_0,v-v_0) = F(u,v) - 2A(u-u_0,v-v_0),$$

и, следовательно, поле в выходной плоскости оптической системы, где установлена голограмма, равно

$$g(x,y) = f(x,y) - 2Ae^{i(u_0x+v_0y)}.$$

Пропускаемость проявленной голограммы (пропорциональная величине gg^*) определяется равенством

$$t(x,y) = \left| f(x,y) - 2Ae^{i(u_0x+v_0y)} \right|^2,$$

которое по форме ничем не отличается от формулы (5.21), определяющей пропускаемость голограммы с наклонной опорной волной

$$f_0(x,y) = 2Ae^{i(u_0x+v_0y+\pi)}.$$

Реконструкция изображения осуществляется по обычной схеме, при этом мнимое изображение предмета возникает на расстоянии z_0 слева от голограммы, а действительное изображение — на том же расстоянии справа от голограммы.

§ 5.10. Голографический синтез оптического изображения из спектра плоских волн

Рассматриваемый ниже способ формирования голограммы основан на способности фотопластинки накапливать действие во времени: почернение проявленной пластинки и, следовательно, её пропускаемость (в определённом диапазоне интенсивностей и времён экспозиции) пропорциональны световому потоку за полное время экспозиции τ: $t(x) \sim I(x) \cdot \tau$. Если интенсивность волны, падающей на фотопластинку-голограмму, со временем изменяется, то

$$t(x) \sim \int_0^\tau I(x,\tau)\, d\tau.$$

Указанное свойство можно использовать, последовательно во времени налагая на одну и ту же голограмму плоские волны, составляющие суммарную волну от предмета, и таким образом синтезировать полное изображение.

Расположим в частотной плоскости Ф в схеме рис. 5.16 непрозрачный экран с двумя малыми отверстиями. Одно из этих отверстий в точке u_0 фиксировано, а положение второго отверстия u_n меняется от одного экспонирования к другому.

Интенсивность света на голограмме, установленной в выходной плоскости, при n-м экспонировании равна

$$I_n(x) = \left|F(u_0)\Delta u\, e^{iu_0x} + F(u_n)\Delta u\, e^{iu_nx}\right|^2. \tag{5.34}$$

Пропускаемость проявленной голограммы пропорциональна световому потоку за полное время экспозиции: $t(x) \sim \sum I_n\Delta\tau_n$. Полагая времена $\Delta\tau$ одинаковыми, получаем $t(x) \sim \sum I_n$. Используя (5.34), находим

$$t(x) = \sum_n \left|F(u_0)e^{iu_0x} + F(u_n)e^{iu_nx}\right|^2 \Delta u^2. \tag{5.35}$$

Слагаемое $F(u_0)\Delta u \cdot e^{iu_0 x}$ остаётся неизменным при всех экспонированиях и играет роль наклонной опорной волны. Раскрывая (5.35), получаем

$$t(x) = \sum_n |F(u_0)|^2 \Delta u^2 + \sum_n |F(u_n)|^2 \Delta u^2 + \\ + F^*(u_0) e^{-iu_0 x} \Delta u \sum_n F(u_n) e^{iu_n x} \Delta u + \\ + F(u_0) e^{iu_0 x} \Delta u \sum_n F^*(u_n) e^{-iu_n x} \Delta u. \quad (5.36)$$

Первые два слагаемых в (5.36) представляют собой постоянную составляющую в пропускаемости и дают при восстановлении плоской, нормально падающей волной плоскую же волну, распространяющуюся по нормали к голограмме. Третье слагаемое можно переписать в виде

$$t_3(x) = F^*(u_0)\Delta u e^{-iu_0 x} \cdot f(x),$$

т. к.

$$\sum_n F(u_n) e^{iu_n x} \Delta u \approx \int F(u) e^{iux}\, du = f(x).$$

Оно приводит к появлению мнимого изображения слева от голограммы (в том же месте, что и при реконструкции голограммы Лейта–Упатниекса). Четвёртое слагаемое

$$t_4(x) \approx F(u_0)\Delta u e^{iu_0 x} \cdot f^*(x)$$

ответственно за появление действительного изображения.

§ 5.11. Математические операции, осуществляемые оптическими системами

С помощью оптических систем можно выполнять самые различные математические операции над функциями двух переменных $f(x,y)$. Напомним, что функции, подлежащие преобразованию, должны быть записаны в виде комплексной пропускаемости транспаранта, установленного во входной плоскости оптической системы. Рассмотрим ряд примеров.

1. *Преобразование Фурье функции двух переменных*

Поле в задней фокальной плоскости линзы является преобразованием Фурье функции $f(x,y)$ (комплексной пропускаемости транспаранта, установленного в передней фокальной плоскости и освещаемого

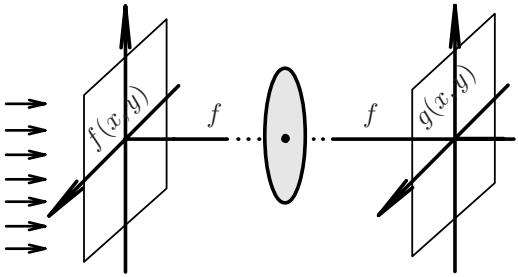


Рис. 5.18

плоской волной (рис. 5.18)):

$$g(x,y) = F\left(\frac{kx}{f}, \frac{ky}{f}\right).$$

2. *Интегрирование функций двух переменных*

Приёмник, установленный в начале координат выходной плоскости (рис. 5.18), зарегистрирует величину

$$g(0,0) = F(0,0) = \iint f(x,y)\,dxdy.$$

3. *Многоканальный анализатор спектра*

На плоском транспаранте можно записать большое число функций одной переменной:

$$f_n(x) = f(x,y_n).$$

Внутри каждого канала $(y_n - \Delta y,\ y_n + \Delta y)$ пропускаемость транспаранта является функцией одной переменной x, а координата y_n определяет номер канала. На рис. 5.19 изображена оптическая система, состоящая из двух вплотную расположенных линз: цилиндрической $Л_1$ и сферической $Л_2$. В направлении оси x такая система осуществляет преобразование Фурье (так как в этом направлении цилиндрическая линза является просто плоскопараллельной пластинкой: $f_1 = \infty$). В направлении оси y система имеет фокусное расстояние $f_{об} = f/2$ (оптические силы линз складываются), и, следовательно, входная и выходная плоскости оказываются сопряжёнными; поэтому в направлении оси y система создаёт изображение (т. е. канал — полоска y_n во входной плоскости переходит в канал — полоску y_n в выходной плоскости). Сказанное поясняет рис. 5.20, где оптическая система показана в двух проекциях: вид сверху (а) и сбоку (б).

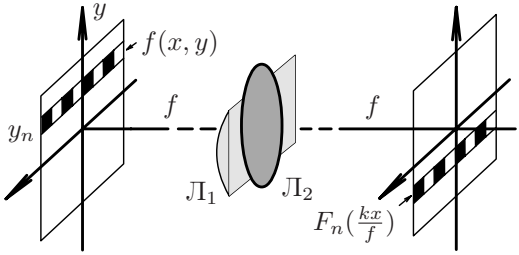
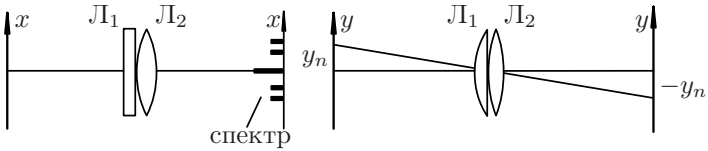


Рис. 5.19



а) б)

Рис. 5.20

Вопрос: чем лимитируется число возможных каналов?

4. *Свёртка функций*

Расположив во входной плоскости схемы Катрона транспарант с пропускаемостью $f_1(x,y)$, а в частотной плоскости — транспарант $F_2\left(\frac{kx}{f},\frac{ky}{f}\right)$, получаем на выходе свёртку функций:

$$g(x,y) = f_1(x,y) \otimes\otimes f_2(x,y).$$

5. *Дифференцирование функций одной переменной*

Расположим в частотной плоскости схемы Катрона маску с пропускаемостью:

$$T(u) = iu = \begin{cases} u \cdot e^{i\frac{\pi}{2}} & \text{при } u > 0, \\ u \cdot e^{-i\frac{\pi}{2}} & \text{при } u < 0 \end{cases}$$

(маска вносит фазовую задержку в π при $u < 0$ и обладает линейно изменяющимся поглощением). На выходе частотной плоскости получаем

$$G(u) = F(u) \cdot iu,$$

и, следовательно,

$$g(x) = \int G(u)e^{iux}\,du = \int iuF(u)e^{iux}\,dx = \frac{d}{dx}\int F(u)e^{iux}\,dx,$$

окончательно

$$g(x) = \frac{d}{dx} f(x).$$

6. *Сложение функций* можно осуществить, используя голографический метод синтеза при последовательных экспонированиях. При экспонировании транспаранта $f_1(x,y)$ получаем

$$I_1 = |f_{1S}(x,y) + f_0(x,y)|^2.$$

Аналогично

$$I_2 = |f_{2S}(x,y) + f_0(x,y)|^2.$$

Полная пропускаемость проявленной голограммы $I = I_1 + I_2$ содержит слагаемое

$$f_0^*(x,y)[f_{1S}(x,y) + f_{2S}(x,y)],$$

которое при восстановлении даёт изображение $f_1(x,y) + f_2(x,y)$. Введя в канал опорного луча при втором экспонировании фазовую задержку в π, можно получить изображение $f_1(x,y) - f_2(x,y)$.

§ 5.12. Цифровая процедура решения фазовой проблемы в оптике

Решению фазовой проблемы в оптике (восстановлению фазовой структуры светового поля по измерениям картины интенсивности) посвящено огромное количество исследований. В частности, напомним о методах восстановления фазового рельефа волнового поля, основанных на разработке различного рода итерационных процедур (например, алгоритм Гершберга—Сэкстона) [13].

Мы предлагаем процедуру, в которой фазовая структура светового поля определяется с помощью решения системы алгебраических уравнений, причём заданными (известными) параметрами являются значения интенсивности, измеренные в дискретном наборе точек плоскости изображения и в дискретном наборе точек фурье-плоскости оптической системы.

Итак, пусть объект — транспарант с комплексной пропускаемостью $f(x) = a(x)e^{i\varphi(x)}$ — расположен во входной плоскости оптической системы (передней фокальной плоскости объектива) и освещается монохроматическим параллельным пучком света с длиной волны λ. При этом световое поле в задней фокальной плоскости объектива описывается функцией $F(\Omega)$ — фурье-образом функции $f(x)$: $f(x) \leftrightarrow F(\Omega)$ (Ω — "пространственная частота", связанная с координатой ξ в фурье-плоскости очевидным равенством $\Omega = k\xi/f_0$, где f_0 — фокусное расстояние объектива, $k = 2\pi/\lambda$ — волновое число).

Предлагаемая процедура восстановления фазовой структуры $\varphi(x)$ предполагает известными (зарегистрированными фотодетектором) значения интенсивности в дискретном наборе точек x_k входной плоскости $I(x_k) = f(x_k)f^*(x_k)$ и значения интенсивности $F(\Omega)F^*(\Omega)$ в дискретном наборе точек $\Omega_n = k\xi_n/f_0$ фурье-плоскости.

Поясним предлагаемый метод с помощью простого примера. Пусть картина интенсивности в фурье-плоскости представляет собой три яркие точки (дифракционные пятна) с координатами Ω_0, Ω_1 и Ω_2 и интенсивностями A_0^2, A_1^2, A_2^2. Таким образом, измеренные значения пространственных частот, а также значения интенсивностей соответствующих спектральных компонент предполагаются известными. Неизвестными и подлежащими определению являются значения фаз α_0, α_1 и α_2 спектральных компонент: $F(\Omega_0) = A_0e^{i\alpha_0}$, $F(\Omega_1) = A_1e^{i\alpha_1}$, $F(\Omega_2) = A_2e^{i\alpha_2}$. Собственно функция пропускания объекта имеет, очевидно, вид

$$f(x) = A_0e^{i\alpha_0}e^{i\Omega_0 x} + A_1e^{i\alpha_1}e^{i\Omega_1 x} + A_2e^{i\alpha_2}e^{i\Omega_2 x}. \qquad (5.37)$$

Пусть в нашем примере $\Omega_0 = 0$, $\Omega_1 = -\Omega_2 = \Omega$, $A_1 = A_2 = A$ и, кроме того, интенсивности боковых гармоник малы, $(A/A_0)^2 \ll 1$. Тогда для интенсивности в выходной плоскости оптической системы, которая тождественно повторяет картину интенсивности во входной плоскости, имеем

$$I(x) = f(x)f^*(x) = A_0^2 + 2A_0A\cos[(\alpha_0-\alpha_1)-\Omega x] + 2A_0A\cos[(\alpha_0-\alpha_2)+\Omega x]. \qquad (5.38)$$

Интерес представляют лишь относительные фазы $(\alpha_1 - \alpha_0)$ и $(\alpha_2 - \alpha_0)$, поэтому, полагая фазу гармоники с частотой Ω_0 равной нулю $(\alpha_0 = 0)$, получаем из (5.38):

$$I(x) = f(x)f^*(x) = A_0^2 + 2A_0A\cos(\alpha_1+\Omega x) + 2A_0A\cos(\alpha_2-\Omega x). \qquad (5.39)$$

Для определения фаз α_1 и α_2 достаточно решить систему двух уравнений, которая получается из (5.39) для двух измеренных значений картины интенсивности в плоскости изображений, например, для точек $x_1 = 0$ и $x_2 = \pi/(2\Omega)$.

Рассмотрим две ситуации.

1. Пусть картина интенсивности $I(x)$ в выходной плоскости имеет вид

$$I(x) = I_0(1 + 2m\cos\Omega x). \qquad (5.40)$$

Для точек x_1 и x_2 получаем

$$x_1 = 0: \quad I_0(1 + 2m) = A_0^2\left(1 + 2\frac{A}{A_0}\cos\alpha_1 + 2\frac{A}{A_0}\cos\alpha_2\right), \qquad (5.41)$$

$$x_2 = \frac{\pi}{2}: \quad I_0 = A^2 \left(1 - 2\frac{A}{A_0}\sin\alpha_1 + 2\frac{A}{A_0}\sin\alpha_2\right). \tag{5.42}$$

Система уравнений (5.41)–(5.42) имеет решение $\alpha_1 = \alpha_2 = 0$ $(= \alpha_0)$, при этом, как следует из сравнения с (5.40), $A/A_0 = m/2$ и $A_0^2 = I_0$, т. е. в этом случае исходная структура $f(x) = a(x)e^{i\varphi(x)}$ является чисто амплитудной. Используя (5.37), находим

$$f(x) = a(x) = A_0 + Ae^{i\Omega x} + Ae^{-i\Omega x} = A_0(1 + m\cos\Omega x), \quad \varphi(x) = 0,$$

(при малых m $I(x) = a^2(x)$ совпадает с (5.40)).

2. Рассмотрим вторую ситуацию. Пусть наблюдаемая картина интенсивности в фурье-плоскости — всё те же три дифракционных пятна с теми же координатами $\Omega_0 = 0$, $\Omega_1 = -\Omega_2 = \Omega$ и теми же интенсивностями A_0 и $A_1 = A_2 = A$, $(A/A_0)^2 \ll 1$. При этом в плоскости изображений модуляция интенсивности отсутствует, т. е. структура является чисто фазовой: при любом x в (5.39) $I(x) = I_0 = \text{const}$. Например, для тех же точек $x_1 = 0$ и $x_2 = \pi/(2\Omega)$ вместо (5.41)–(5.42) имеем

$$I_0 = A_0^2 \left(1 + 2\frac{A}{A_0}\cos\alpha_1 + 2\frac{A}{A_0}\cos\alpha_2\right); \tag{5.43}$$

$$I_0 = A_0^2 \left(1 - 2\frac{A}{A_0}\sin\alpha_1 + 2\frac{A}{A_0}\sin\alpha_2\right). \tag{5.44}$$

Система (5.43)–(5.44) имеет решение $\alpha_1 = \alpha_2 = \pi/2$ (напомним, что $\alpha_0 = 0$). Получаем из (5.37)

$$f(x) = A_0 + A\exp\left[i\left(\Omega x + \frac{\pi}{2}\right)\right] + A\exp\left[i\left(-\Omega x + \frac{\pi}{2}\right)\right] = A_0(1 + im\cos\Omega x),$$

т. е. при $(A/A_0)^2 \ll 1$ находим: $f(x) = A_0\exp(im\cos\Omega x)$, где $m = A/A_0$ — глубина модуляции фазы. Мы получили, используя предложенную процедуру, хорошо известный результат — закон фазовой модуляции $\varphi(x) = m\cos\Omega x$.

Рассмотренный пример легко обобщается. В самом общем случае представим спектр $F(\Omega)$ финитной функции $f(x) = 0$ при $x \geqslant D$ в виде ряда:

$$F(\Omega) = \sum_n F\left(\frac{n\pi}{D}\right)\frac{\sin(\Omega D - n\pi)}{\Omega D - n\pi} \tag{5.45}$$

(т. н. обратная теорема Котельникова). Взяв обратное преобразование Фурье равенства (5.45), получаем

$$f(x) = \sum_n F\left(\frac{n\pi}{D}\right)\exp\left(i\frac{n\pi}{D}x\right); \quad f(x) \equiv 0 \text{ при } |x| \geqslant D.$$

Далее находим картину интенсивности:

$$I(x) = f(x)f^*(x) = \sum_n \sum_m F\left(\frac{n\pi}{D}\right) F^*\left(\frac{m\pi}{D}\right) \exp\left[i(n-m)\frac{\pi x}{D}\right]. \tag{5.46}$$

Важно отметить, что для восстановления фазовой структуры поля $f(x)$ достаточно иметь конечное число слагаемых в (10). Максимальное значение $n_{\max} = N$ можно оценить с помощью соотношения неопределённостей $\Omega_{\max}\delta \approx 2\pi$, где δ — минимальный размер неоднородностей (деталей), который необходимо сохранить при восстановлении поля $f(x)$:

$$\frac{N\pi\delta}{D} \approx 2\pi, \qquad \text{т. е. } N \approx \frac{2D}{\delta}. \tag{5.47}$$

Измерение интенсивности $I(x_k)$ в N точках плоскости изображения объекта $f(x)$ и в N точках фурье-плоскости $|F(n\pi/D)|^2 = A_n^2$ позволяет в принципе восстановить фазы α_n спектральных компонент $F(n\pi/D) = A_n e^{i\alpha_n}$.

Цифровые методы, основанные на использовании огромных возможностей современных компьютеров, позволяют решить систему N алгебраических уравнений, которая получается из (10) для N значений интенсивности $I(x_k)$. Каждое из этих уравнений при заданных x_k представляет собой сумму, состоящую из пар комплексно-сопряжённых слагаемых, поэтому

$$I(x_k) = 2\sum_n \sum_m A_n A_m \cos\left[(\alpha_n - \alpha_m) + (n-m)\frac{\pi x_k}{D}\right], \quad k = 1,2,\ldots,N. \tag{5.48}$$

При практической реализации измеряемые значения интенсивности $I(x_k)$ усредняются по размеру Δx чувствительного элемента регистрирующего фотодетектора (например, ПЗС-матрицы). Последний должен быть меньше δ — минимального размера деталей неоднородностей поля $f(x)$, подлежащего восстановлению. Наконец, следует отметить, что предлагаемый метод предъявляет серьёзные требования к точности измерения значений интенсивности $I(x_k)$ в отсчётных точках. Действительно, даже малая глубина модуляции m интенсивности (в первом примере) приводит к скачкообразному изменению фаз слагаемых пространственных гармоник ($\alpha_1 = \alpha_2 = \pi/2$ во втором примере и $\alpha_1 = \alpha_2 = 0$ в первом). Впрочем, высокие требования к точности регистрации интенсивности характерны для всех методов восстановления фазового рельефа светового поля по картинам интенсивностей.

Литература

1. *Рытов С.М.* О методе фазового контраста в микроскопии // УФН. 1950. Т. 41. С. 425.

2. *Зверев В.А.* Радиоптика. — М.: Советское радио, 1975.

3. *Гудмен Дж.* Введение в фурье-оптику. — М.: Мир, 1970.

4. *Папулис А.* Теория систем и преобразований в оптике. — М.: Мир, 1971.

5. *Строук Дж.* Введение в когерентную оптику и голографию. — М.: Мир, 1967.

6. *Сороко Л.М.* Основы голографии и когерентной оптики. — М.: Наука, 1971.

7. *Свет В.Д.* Оптические методы обработки сигналов. — М.: Энергия, 1971.

8. *Дьяков В.А., Тарасов Л.В.* Оптическое когерентное излучение. — М.: Советское радио, 1974.

9. *Микаэлян А.Л.* Голография. — М.: Знание, 1968.

10. *Аблеков В.К., Зубков П.И., Фролов А.В.* Оптическая и оптоэлектронная обработка информации. — М.: Машиностроение, 1976.

11. *Франсон М., Сланский С.* Когерентность в оптике. — М.: Наука, 1967.

12. *Юу Ф.Т.С.* Введение в теорию дифракции, обработку информации и голографию. — М.: Советское радио, 1979.

13. *Локшин Г.Р.* Основы радиооптики. — Долгопрудный: Интеллект, 2009.

14. *Кингсеп А.С., Локшин Г.Р., Ольхов О.А.* Основы физики. Т I. Ч. III. Физика колебаний и волн. Волновая оптика. — М.: Физматлит, 2001.

Учебное издание

Локшин Геннадий Рафаилович

ДИФРАКЦИЯ. ПРОСТРАНСТВЕННАЯ ФИЛЬТРАЦИЯ

Редактор *И. А. Волкова*. Корректор *О. П. Котова*.

Подписано в печать 23.05.2016. Формат $60 \times 84^{1}/_{16}$.
Усл. печ. л. 9,75. Уч.-изд. л. 9,25. Тираж 250 экз. Заказ .

Федеральное государственное автономное образовательное
учреждение высшего профессионального образования
«Московский физико-технический институт (государственный университет)»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
Тел. **+7(495)408–5822, +7(499)744–6512**. E-mail: **rio@mipt.ru**

Отдел оперативной полиграфии «Физтех-полиграф»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
Тел. **+7(495)408–8430**. E-mail: **polygraph@mipt.ru**