Chapter 2

Background

2.1 Introduction

The human visual system is a complex transducer that is believed to have adapted over the last several million years toward providing an efficient representation of the natural world [8]. This chapter provides an overview of those aspects of vision-science research which are most applicable to image compression; in particular, contrast perception and visual summation. For a thorough treatment of the subject, the reader is referred to Refs. [9], [10], and [11].

This chapter organized as follows: Section 2.2 provides a brief overview of the general anatomy and neurophysiology of the human visual system, a description of its psychophysical characterization, and a summary of visual processing of natural images. Section 2.3 describes wavelet-based image coding (discrete wavelet transform, quantization, entropy coding), and an overview of the JPEG-2000 standard.

2.2 Spatial Vision

Three major lines of research have established the majority of what is currently known about the human visual system (HVS). Anatomical studies have successfully mapped out several key areas involved in vision. Physiological studies have provided much insight into the response properties and tuning characteristics of the neurons within these areas. And, psychophysical research, which infers HVS
properties based on experiments using human subjects, has proved a crucial tool in linking neuroanatomy and neurophysiology with human perception and cognition.

2.2.1 Structure of the HVS

This section provides an overview of the visual areas which process the input signal as it ascends the visual hierarchy. Figure 2.1 depicts a block diagram which shows the feedforward connections between key visual processing areas. These areas include the eye, lateral geniculate nucleus (LGN), striate cortex (V1), and higher-level areas such as V2, V3, V4, medial temporal cortex (V5, MT), inferior temporal cortex (IT), and posterior parietal cortex.
Eye

The human eye serves to project and convert light into neural activity. Light enters the cornea, passes through the aqueous humor, then through the lens into the vitreous humor, and finally onto the photoreceptors (rods and cones, discussed next) located at the back of the retina. The ciliary muscles are responsible for accommodating the lens so as to focus the light rays onto the fovea, the region of the retina containing the greatest density of cones and thus the highest acuity for spatial and color vision [10][9].

Retina

The retina is composed of five layers of neurons: photoreceptors, horizontal cells, bipolar cells, amacrine cells, and ganglion cells [11]. The photoreceptors—of which two types exist, rods and cones—are responsible for transducing light into neural output: The rods are responsible for vision in low-light (scotopic) conditions, whereas the cones function under normal (photopic) lighting and are responsible for color vision. The horizontal, bipolar, and amacrine cells comprise the plexiform layer of the retina: Bipolar cells receive input from the photoreceptors and provide output to the ganglion cells; horizontal and amacrine cells provide lateral connections between photoreceptors, and between bipolar cells and ganglion cells, respectively [12].

Retinal ganglion cells constitute the first stage of complex spatial processing. Via extracellular recording techniques, the receptive fields of ganglion cells have been shown to consist of an antagonistic center-surround organization [13]. The receptive field of an on-center ganglion cell contains a central, disc-shaped excitatory region, flanked by an annulus-shaped inhibitory region; whereas the receptive
field of an off-center ganglion cell exhibits a reciprocal spatial organization. The spatial extents of these receptive fields come in a variety of sizes, which gives rise to spatial-frequency selectivity. Furthermore, because of the antagonistic center-surround organization, ganglion cells respond to ratios of luminance, or contrast (see Section 2.2.2), rather than to absolute luminance (though the exact functional role of these neurons is still an area of active research [14][15]). Ganglion cells have also been classified according to the layer in LGN (discussed next) to which their outputs are directed: M cells, which receive input from both rods and cones, demonstrate high sensitivity to contrast, but reduced sensitivity to color. P cells, which receive input from only cones, demonstrate high sensitivity to color, but reduced sensitivity to contrast.

LGN

The outputs of retinal ganglion cells travel through the optic nerve, and then pass through the optic chiasm. Here, outputs from ganglion cells in the nasal region of each retina are directed to opposite hemispheres. The neural outputs then synapse upon either superior colliculus, a region in the brain stem which is believed to be responsible primarily for eye movements [11]; or lateral geniculate nuclei (LGN), a portion of the thalamus composed of six laminar sheets of neurons [10][12].

As with ganglion cells, neurons in LGN exhibit center-surround-type receptive fields. Two of LGN’s six layers, the magnocellular layers, contain neurons with large receptive fields that are responsible for coding achromatic contrast; these neurons receive input from the M ganglion cells. Neurons in the other four layers, the parvocellular layers, contain smaller receptive fields and accordingly demonstrate a high degree of spatial resolution; these parvocellular neurons, which receive
input from the P ganglion cells, are highly sensitive to color, but show reduced contrast sensitivity and lower temporal resolution than magnocellular neurons.

**Primary Visual Cortex**

Output from LGN is directed to the visual areas of cortex (*visual cortex*), with the majority of the projections synapsing in *primary visual cortex* (V1; also called *striate cortex* or *area 17*). V1, which is located in the posterior region of the occipital lobe, is the largest area of visual cortex that contains approximately 200 million neurons, more than 100 times the amount found in LGN [10].

As with LGN, V1 is composed of six layers, the fourth of which is itself divided into four sub-layers (4A, 4B, 4Cα, and 4Cβ). Most input from LGN is directed at layer 4C (magnocellular input to layer 4Cα, parvocellular input to layer 4Cβ), which then sends the processed signals on to layers 2, 3, and 4B [12]. Layers 2 and 3 contain *blob* and *interblob* regions; the blobs, which contain color-selective neurons, receive both parvocellular and magnocellular input (via layer 4Cβ; primarily parvocellular input); whereas the interblobs, which contain neurons that are sensitive to orientation but largely insensitive to color, receive only parvocellular input (via layer 4Cβ). Layer 4B receives magnocellular input (via layer 4Cα) and contains neurons which demonstrate both orientation selectivity and selectivity for direction of motion [9].

In their seminal work, Hubel & Weisel [16][17] investigated via extracellular recordings the receptive fields of neurons in primary visual cortex of cat and monkey; they classified V1 cells as *simple*, *complex*, or *hypercomplex* based on neural tuning characteristics and degrees of nonlinearity:

- **Simple cells**, which comprise approximately 25% of V1 neurons, contain
elongated receptive fields which demonstrate selectivity primarily to bars or edges of specific widths and orientations thus giving rise to both spatial-frequency and orientation tuning. Hubel & Weisel proposed that the receptive fields of these so-called edge and bar detectors might be constructed based on the outputs of several LGN neurons. Simple cells are “simple” in the sense that their responses to complex patterns can, to a first approximation, be predicted based on their responses to the pattern’s constituent spots of light [18].

- **Complex cells**, on the other hand, exhibit response properties which are highly nonlinear. The vast majority of complex cells respond to oriented bars and edges, often with a preference for a particular direction of motion; however, complex cells do not respond to the individual spots of light which comprise the bars and edges. Furthermore, complex cells demonstrate a degree of position invariance wherein small perturbations in the location of a stimulus does not effect changes in response. The receptive fields of complex cells have been proposed to be constructed based on the outputs of several simple cells with similar spatial-frequency and orientation tuning characteristics, thus giving rise to position invariance and direction-of-motion selectivity [17].

- **Hypercomplex cells** exhibit an extra degree of selectivity in that they also demonstrate a preference for stimuli of limited length. Exposing a hypercomplex cell to a bar or edge beyond the neuron’s preferred length results in a decreased response, a phenomenon termed *end stopping*. Many now believe that hypercomplex cells are end-stopped simple or complex cells [10].
Other Cortical Areas

Output from V1 is sent to V2 and on to other higher-level areas of visual cortex, though not necessarily in a serial fashion. Cytochrome-oxidase staining techniques have shown that V2 is composed of pale, thin, and thick stripes, which receive input from interblobs, blobs, and layer 4B in V1, respectively, and which are believed to process form, color, and depth information, respectively [12]. Output from the thin stripes of V2 are sent to V4 the majority of whose own output is directed at inferior temporal cortex (IT); this path, V1 (blobs) → V2 (thin stripes) → V4 → IT, is often called the “what” system as it is believed to be responsible for form perception and recognition. Output from the thick stripes of V2 project to medial temporal cortex (V5, MT; responsible for processing stereo and motion), which in turn projects to medial superior temporal cortex (MST; responsible for visual tracking), and then to posterior parietal cortex (PP); this path, V1 (layer 4B) → V2 (thick stripes) → V5 → MST → PP, is known as the “where” system which is believed to handle localization.

Unfortunately, beyond this task-based classification, much less is known about the mechanisms of these extrastriate cortical areas. Lee et. al [19] have proposed that higher levels work in conjunction with V1 to perform complex tasks such as pattern analysis and object recognition. Rao & Ballard [20] have suggested that higher levels function as predictive coders whose feedback connections to V1 carry the prediction and whose feedforward connections from V1 convey the prediction’s error. Other researchers have argued that higher levels have evolved to efficiently encode the joint activity of V1 neurons [21][22][23]. However, on the whole, the functional mechanisms of extrastriate visual cortex (and even much of V1; see [24]) remain largely unknown.
2.2.2 Contrast Sensitivity

In addition to anatomical- and neurophysiological-based approaches, much insight into human vision has been provided via psychophysical techniques. For example, the HVS can be experimentally characterized by finding contrast detection thresholds (or simply thresholds) for detection of various spatial patterns, most commonly sine-wave gratings (see Figure 2.3). The contrast detection threshold for a given target is defined as the minimum contrast necessary for a human observer to visually detect the target.

Definitions of Contrast

Various measures of contrast have been employed in psychophysical studies, including simple contrast, Weber contrast, Michelson contrast, and root-mean-squared contrast (RMS contrast):

- **Simple contrast** is defined as
  \[
  C_{\text{simple}} = \frac{L_{\text{max}}}{L_{\text{min}}} \tag{2.1}
  \]
  where \(L_{\text{min}}\) and \(L_{\text{max}}\) denote the target’s minimum and maximum luminance, respectively.

- **Weber contrast** is defined as
  \[
  C_{\text{simple}} = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{min}}} \tag{2.2}
  \]
  where \(L_{\text{min}}\) and \(L_{\text{max}}\) denote the target’s minimum and maximum luminance, respectively.

- **Michelson contrast** is defined as
  \[
  C_{\text{Michelson}} = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{max}} + L_{\text{min}}} = \frac{1}{2} \left( \frac{L_{\text{max}} - L_{\text{min}}}{\mu_L} \right) \tag{2.3}
  \]
where \( L_{\text{min}} \) and \( L_{\text{max}} \) denote the target’s minimum and maximum luminance, respectively; and where \( \mu_L \) denotes the mean luminance.

- **RMS contrast** is defined as

\[
C_{\text{RMS}} \equiv \frac{1}{\mu_L} \left( \frac{1}{N} \sum_{i=0}^{N} [L_i - \mu_L]^2 \right)^{1/2} = \frac{\sigma_L}{\mu_L} \tag{2.4}
\]

where \( \mu_L \) denotes the mean luminance, \( \sigma_L \) denotes the standard deviation of the target’s luminances, \( L_i \) denotes the target’s luminance at spatial location \( i \), and \( N \) denotes the total number of spatial locations. For cases in which the mean luminance of the target differs from the mean luminance of the background, RMS contrast is given by

\[
C_{\text{RMS}} \equiv \frac{1}{\mu_{\text{bkgnd}}} \left( \frac{1}{N} \sum_{i=0}^{N} [L_i - \mu_{\text{target}}]^2 \right)^{1/2} = \frac{\sigma_{\text{target}}}{\mu_{\text{bkgnd}}} \tag{2.5}
\]

where \( \mu_{\text{bkgnd}} \) and \( \mu_{\text{target}} \) denote the mean luminance of the background and target, respectively; and where \( \sigma_{\text{target}} \) denotes the standard deviation of the target’s luminances.

Most often, the contrast metric is selected based on the experimental stimuli; e.g., Michelson (peak-to-peak) contrast is commonly used for targets consisting of sine-wave gratings, whereas RMS contrast is commonly employed for non-periodic targets (noise, textures, images). Note that the luminance values used in all of these definitions of contrast are defined in standard CIE units of candelas/square meter (cd/m\(^2\)).

**Contrast Sensitivity Function**

Contrast detection thresholds were first measured by Schade [25] in an experiment that presented human observers with achromatic sine-wave gratings of various
The key result of Schade’s experiment was the discovery that contrast sensitivity varies with the spatial frequency of the grating, a relationship described by a contrast sensitivity function (CSF) with peak sensitivity near 4-6 cycles/degree. In humans, the reduction in sensitivity at high spatial frequencies has been attributed to limitations both in the optics of the eye and in receptor spacing, and to quantum noise; whereas at low spatial frequencies, reduced sensitivity is believed to occur in part by limited receptive-field sizes (though see the following section and Refs. [1], [26]).

Schade measured HVS contrast sensitivity to sine-wave gratings based on the
notion that any stimulus can be described as a superposition of sine-waves. This idea was extended by Campbell & Robson [27] who measured contrast detection thresholds for both sine-wave and square-wave gratings. Because a square wave is composed of numerous sine waves (in constructive and destructive phases), the physical contrast of a square wave will always be lower than the physical contrast of its fundamental sine-wave component (by a factor of approximately 1.3 in Ref. [27]). Campbell et al. found contrast detection thresholds for the square-wave gratings to be approximately 1.3 times lower than those found for the sine-wave gratings; they concluded from this result that the visual system performs a Fourier-type decomposition of a stimulus in which the (sine-wave) components are detected independently via tuned spatial-frequency channels. This paradigm is known as the multi-channel model of human vision.

**Multi-Channel Models**

Campbell & Robson’s suggestion of a multi-channel HVS model prompted a number of psychophysical studies, for, if the contrast threshold a complex stimulus could be predicted from the individual thresholds of its Fourier components, the CSF is not a modulation transfer function (MTF) of the visual system as a whole, rather, it is a consequence of the tuning characteristics of the individual channels. Evidence in support of this notion has since been provided by adaptation and summation experiments [9]. For example, the CSF measured for an observer adapted to a grating of a single spatial frequency shows attenuation only within a limited range of frequencies (approximately 1 octave) around that of the grating’s. A true MTF, on the other hand, would show attenuation throughout the frequency spectrum. Brady & Field [26] explained the shape of the CSF in the
context of a model with equally-sensitive spatial-frequency channels; reduction in detection performance for high spatial frequencies was attributed to extrinsic noise that dominates the response of channels tuned to high frequencies therefore resulting in decreased signal-to-noise ratios for these higher-frequency channels.

Further psychophysical evidence in support of multi-channel models has emerged from summation experiments, wherein the detectability of a compound target (e.g., a plaid grating composed of two sine waves; see Figure 2.3) is compared to the detectability of its components. If the components within the compound target are detected by independent channels, then the compound target should not be detectable until at least one of its components reaches its own detection threshold. Current evidence [28][11][29][30][31][32][33] suggests this to be the case as long as the two components are separated in spatial frequency by at least one octave or in orientation by at least $30^\circ-45^\circ$. Similar experiments have shown channels tuned to other dimensions such as color and direction of motion [9].

### 2.2.3 Summation of Responses

Although it is generally accepted that the human visual system (HVS) decomposes visual input via a bank of dimensionally localized channels [27], it is less clear how the responses of these channels are combined or summed to form what is ultimately seen. Summation experiments begin to address this issue by comparing the detectability of a compound target (e.g., a target consisting of two sine-wave gratings of different orientations) to the detectability of its individual components, i.e., its components presented as simple targets. If the compound target is more easily detected than the simple targets, the visual responses to the compound target’s components are believed to have summed [11]. Figures 2.3(a) and 2.3(b)
Figure 2.3: Simple and compound sine-wave gratings. (a) Simple 8 cycles/picture horizontally oriented sine-wave grating, \( t_1 \). (b) Simple 8 cycles/picture vertically oriented sine-wave grating, \( t_2 \). (c) Compound sine-wave grating composed of these two simple targets generated via \( t_1 + t_2 - \mu_t \), where \( \mu_t \) denotes the mean value of \( t_1 \) (and \( t_2 \)).

depict simple targets consisting of, respectively, horizontally and vertically oriented sine-wave gratings; Figure 2.3(c) depicts a compound target composed of these two simple targets.

Relative Contrast Threshold

Let \( t \) denote a simple target. The relative contrast of \( t \), \( RC(t) \), is defined as

\[
RC(t) \equiv \frac{C(t)}{CT(t)} \tag{2.6}
\]
where $C(t)$ is the contrast of $t$ and $CT(t)$ is the contrast threshold of $t$ [11]. Let $t$ denote a compound target composed of two simple targets $t_1$ and $t_2$. The \textit{relative contrast threshold} of $t_i$, $RCT(t_i; t)$, $i = 1, 2$, is defined as

$$RCT(t_i; t) \equiv \frac{CT(t_i|t)}{CT(t_i)} \quad (2.7)$$

where $CT(t_i|t)$ is the contrast threshold of target $t_i$ measured when $t_i$ was presented as part of $t$; and $CT(t_i)$ is the contrast threshold of $t_i$ measured when $t_i$ was presented alone (i.e., as a simple target) [11].

When $RCT(t_i; t) = 1$, $CT(t_i|t) = CT(t_i)$, suggesting that the detectability of $t_i$ is not affected by the presence of the other component $t_j$ ($j \neq i$) of $t$. When $RCT(t_i; t) < 1$, $CT(t_i|t) < CT(t_i)$, suggesting that the detectability of $t_i$ is enhanced by the presence of $t_j$. When $RCT(t_i; t) > 1$, $CT(t_i|t) > CT(t_i)$, suggesting that the detectability of $t_i$ is reduced by the presence of $t_j$.

### Relative Sensitivity and Quick Pooling

Summation is typically quantified via a single parameter denoting either \textit{relative sensitivity} [32] (RS; also called \textit{summation index} [11] or \textit{threshold ratio} [34]) or a Minkowski summation exponent ($\beta$). For a compound target composed of two components, the Minkowski sum (Quick approximation [35]) is given by

$$RCT(t_1; t)^\beta + RCT(t_2; t)^\beta = 1. \quad (2.8)$$

When the components within the compound target are equally detectable, i.e., when $RCT(t_1; t) = RCT(t_2; t)$, relative sensitivity $RS \equiv 1/RCT(t_i; t) = 2^{1/\beta}$ (see Ref. [11]). Thus, when $\beta = 1$, $RS = 2$, suggesting complete or \textit{linear} summation [i.e., $\frac{1}{2} + \frac{1}{2} = 1$ in Equation (2.8)]; whereas when $\beta = \infty$, $RS = 1$, suggesting no summation. In the latter case, the summation model functions as a maximum
operator, i.e., the visual response to the compound is based only on the channel with the greatest output. Although we focus our current discussion on compound targets composed of two components, note that the Minkowski summation model was derived for the more general context in which compound targets are composed of \( M \) components (see, e.g., Ref. [36]); in this case, when the components within the compound are equally detectable, \( RS = M^{1/\beta} \).

For the compound target depicted in Figure 2.3(c), under the single-channel peak-detector model \( (\beta = 1) \), because the contrast of the compound target is twice that of its components, the compound target would be twice as detectable as either component alone; in this case, relative sensitivity \( RS = 2 \). Under a single-channel energy model [32], \( \beta = 2 \), and thus the compound target would be about \( RS = 1.4 \) times as detectable as either component alone. Under the multi-channel model, the visual system separately detects the components; thus, the compound target would be either slightly more detectable \( (RS \approx 1.2, \beta = 4-5, \text{ termed probability summation}) \) or no more detectable \( (RS = 1, \beta = \infty) \) than either component alone depending on the whether the detectors operate in a statistically independent (assuming uncorrelated noise) or dependent (assuming correlated noise) fashion prior to the decision stage [11].

**Summation-Square Plots**

Summation results are typically reported in the form of a summation square plot as depicted in Figure 2.4. The horizontal axis corresponds to the relative contrast of one of the compound target’s components and the vertical axis corresponds to the relative contrast of the other component. For linear summation \( (RS = 2; \beta = 1) \), relative contrast thresholds would fall on the diagonal line connecting coordinates
Figure 2.4: Example summation-square plot denoting regions of linear, energy, probability, and no summation. The horizontal axis corresponds to the relative contrast of one of the compound target’s components; the vertical axis corresponds to the relative contrast of the other component. For linear (complete) summation ($RS = 2; \beta = 1$), relative contrast thresholds would fall on the diagonal line connecting coordinates $(0,1)$ to $(1,0)$. For no summation ($RS = 1; \beta = \infty$), relative contrast thresholds would fall on the lines formed by connecting $(0,1)$ to $(1,1)$ and $(1,1)$ to $(1,0)$. The majority of summation-at-threshold experiments have found $RS \in [1.2, 1.4]$.

(0,1) to (1,0). For no summation ($RS = 1; \beta = \infty$), relative contrast thresholds would fall on the lines formed by connecting $(0,1)$ to $(1,1)$ and $(1,1)$ to $(1,0)$.

Previous summation studies have compared the detectability of grating- and Gabor-type compound targets to the detectability of their individual components. The general conclusion of these studies is that, at near-threshold contrasts, the compound target is only slightly more detectable than either of its components given that the components are far enough apart along the dimension of interest such that each component is analyzed by a separate channel (i.e., the response of channel $A$ to component $B$ is negligible, and vice-versa). The minor increase
in detectability is commonly attributed to either probability summation [29] or nonlinear summation [35], typically with \( RS \in [1.2, 1.4] \). A detailed summary of previous summation studies is provided in Chapter 3.

### 2.2.4 Visual Masking

Visual masking is a general term that refers to the perceptual phenomenon in which the presence of masking signal (the *mask* or *masker*) reduces a subject’s ability to detect a given test signal (the *target*). In this paradigm, the task of detection becomes one of *masked detection*, and contrast thresholds denote *masked detection thresholds*.

Current explanations of visual masking can generally be divided into three paradigms:

1. *Noise masking*, which attributes the increase in detection thresholds to the corruptive effects of the masker on internal decision variables [37];

2. *Contrast masking*, which attributes threshold elevations to contrast gain control [2] (discussed next);

3. *Entropy masking*, which is imposed solely by an observer’s unfamiliarity with the masker [38].

Because a masker’s contrast is readily computable, contrast masking has been exploited in a variety of image processing applications [39][40][41][42][43]. The extent to which a masker constitutes visual noise, and the extent to which an observer is unfamiliar with a masker, are phenomena which are more difficult to quantify; accordingly, the application of noise and entropy masking to image processing remains an area of active research (see, e.g., [44], [45]).
Contrast Masking and Threshold Elevation

Contrast masking results are commonly reported in form of threshold-versus-contrast (TvC) curves, in which masked detection thresholds are plotted as a function of the contrast of the masker. This notion is depicted in Figure 2.5 in which contrast thresholds for detection of 1.0, 2.0, and 4.0 cycles/degree sine-wave gratings presented against a 2.0 cycles/degree sine-wave-grating masker are plotted as a function of the contrast of the masker [2]. As shown in Figure 2.5, masked detection thresholds generally increase as the contrast of the masker is increased and often demonstrate a region of facilitation (i.e., a decrease in threshold;

In some cases, masked detection thresholds are normalized (divided) by the corresponding unmasked threshold to denote threshold elevations.
Figure 2.6: Standard gain control model of V1 neurons: The initial, linear response of a V1 neuron is computed based on a weighted inner product of the neuron’s receptive field and an input image; this response is then subjected to a pointwise nonlinearity followed by divisive inhibition (normalization) from other neurons.

“dipper effect”) at lower masker contrasts, depending on the dimensional relationship between the target and the masker (e.g., differences in spatial frequency and orientation).

Gain Control

Both visual masking and the nonlinear responses of cortical neurons have been described based on a widely accepted model of gain control (see Figure 2.6) [46][47][48][49][50][51]. Under this model, the response of a neuron is computed via: (1) a weighted inner product between the input image and the neuron’s receptive field, (2) a point-wise nonlinearity, and (3) divisive inhibition from other neurons which comprise a so-called “inhibitory” or “normalization” pool. These inhibitory neural interactions are believed to serve as a regulatory means of response normalization whereby the gain of the detecting neuron is attenuated so as to prevent a saturated response and therefore maximize the differential sensitivity of the neuron [51].

Contrast masking, in particular, has been modeled by using variations of the standard gain-control model [49] wherein detection thresholds are predicted based
on the difference between the model’s response to the masker alone (e.g., an original image) and the response to the mask+target (e.g., an image with compression-induced distortions). Indeed, models of this type generally perform well at predicting the detectability of simplistic targets presented against artificial backgrounds (e.g., a sine-wave target presented against a sine-wave masker). However, image-processing applications are typically more concerned with the detectability of specific targets presented against naturalistic, structured backgrounds (e.g., the visibility of a watermark in an image). In particular, when applied to natural images, threshold elevations for structured areas are consistently over-estimated by gain control models. Furthermore, masking has primarily been investigated in the context of detection, whereas image-processing applications—image compression, in particular—often deal with suprathreshold targets (e.g., visible compression artifacts).

2.2.5 Suprathreshold Sensitivity

As mentioned in Section 2.2.2, contrast sensitivity is traditionally defined as the inverse of contrast detection threshold. This definition has drawn considerable opposition however, because the natural world is dominated by clearly visible or suprathreshold stimuli. Several studies have shown that the perceived contrast of a suprathreshold target depends much less on its spatial frequency than what is predicted by the CSF. For highly suprathreshold gratings, perceived contrast is relatively invariant to the spatial frequency of the gating, a finding termed contrast constancy [3].

Contrast constancy was first reported by Georgeson & Sullivan [3] who conducted contrast-matching experiments in which observers adjusted the contrast of
Figure 2.7: Contrast-matching results of Ref. [3]. The horizontal and vertical axes correspond, respectively, to the spatial frequency and contrast of the sine-wave gratings which subjects adjusted to match in perceived contrast to a fixed, 5.0 cycles/degree sine-wave grating. Filled symbols aligned at 5.0 cycles/degree denote the “to-be-matched” contrasts of the fixed grating; open symbols denote the contrasts of the adjustable gratings when matched in perceived contrast to the fixed grating. Note that the vertical axis represents increasing contrast in the downward direction. (From Ref. [3]; cf Figure 2.2.)

a sine-wave grating of spatial frequency \( f \) until it perceptually matched the contrast of a fixed sine-wave grating of spatial frequency 5.0 cycles/degree. When matched by observers in perceived contrast, the differences between the physical contrasts of any two gratings could be predicted from the contrast sensitivity function only at near-threshold contrasts. As the contrast of the fixed (“to-be-matched”) grating became increasingly suprathreshold, perceived contrasts approached physical contrasts, resulting in a flattened “suprathreshold contrast sensitivity” curve; see Figure 2.7.

Georgeson et al. attributed contrast constancy to an intrachannel gain control mechanism that, at suprathreshold contrasts, compensates for reduced sensitiv-
ity both at low spatial frequencies and at high spatial frequencies. In a similar study, Brady & Field [26] attributed contrast constancy to a multi-channel model with equally-sensitive spatial-frequency channels proposed to yield a constant response to the spatial scales of natural scenes. These data provide evidence which confounds the ability of contrast detection experiments to measure an observer’s sensitivity to a stimulus; instead, the CSF at threshold might denote an observer’s ability only to detect the stimulus under the given experimental conditions.

2.2.6 Visual Processing of Natural Images

While the characterization of V1 based on its responses to simple stimuli has proved useful, other researchers have suggested that in order to fully understand the response properties of cortex, one must first understand the signal that is to be encoded: natural images. Field [8] suggested that natural images possess characteristic statistical regularities that have imposed evolutionary constraints on the functional role of cortical neurons, and thus this special class of input has the potential of revealing properties of visual cortex beyond those invoked by using simple stimuli. Indeed, the use natural images in psychophysical studies [52][53][54][55] has revealed both supportive and confounding evidence for previous theories of V1.

Amplitude Spectrum

For example, natural images exhibit characteristic amplitude spectra which generally follow a $f^{-\alpha}$ trend (where $f$ denotes spatial frequency; $\alpha \in [0.7, 1.6]$) [8], a property which is believed to result from the scale-invariant/fractal nature of natural scenes [56, 57]. Knill, Kersten, & Field [56] have shown that human dis-
discrimination of fractal Brownian textures is optimal with $\alpha \in [1.4, 1.8]$. Parraga, Troscianko, & Tolhurst [54] demonstrated that discrimination performance between morphed pairs of natural images was best for stimuli with $\alpha = 1$. Webster & Miyahara [53] have shown that changes in the slope of an image’s amplitude spectrum induces selective effects when contrast thresholds and suprathreshold contrast matches are measured following adaptation to the image.

**Phase Spectrum**

Natural images also possess a coherent phase structure which is the primary contributor to an image’s phenomenal appearance. This fact was first demonstrated by Oppenheim & Lim [58] who synthesized an image from the amplitude spectrum of one image and phase spectrum of another; the resulting image appeared much more similar to the image whose phase structure was used. Thomson, Foster, & Summers [59] have demonstrated that randomization or quantization of this phase structure severely impacts the semblance of an image. Similarly, Bex & Makous [55] have shown that randomizing a natural image’s phase structure at a particular spatial scale decreases detection and contrast-matching performance by the same amount as that found when the spatial scale is completely removed. In addition, Geisler, Perry, Super, & Gallogy [60] have demonstrated that the edge co-occurrence statistics of natural images can be used to predict human performance in a contour-detection task.

**Ecological Theories of Visual Coding**

The predominance of low spatial frequencies in natural images suggests slow changes in intensity, which are reflected by high positive interpixel correlations (i.e., neigh-
boring pixels tend to have similar intensity values). Atick [14] has argued that retinal ganglion cells may have evolved to remove these second-order (pair-wise) correlations, yielding a “whitened” image as input to LGN and V1. More recently, Graham et al. [15] have argued that retinal ganglion cells serve to provide a means of response equalization, which facilitates the use of channels tuned to higher spatial frequencies when exposed to natural-scene input, and which reduces the number of active units required to encode the signal (sparse coding). Similarly, Olshausen & Field [61] have shown that training a neural network on natural images under a sparse-coding constraint yields a basis set which possesses similarities to cortical simple-cell receptive fields. Hyvärinen & Hoyer [62] later discovered similar correspondences between cortical complex-cell receptive fields and a basis set generated by training a multi-layer neural network on natural images. Other nonlinear computational models have been used to demonstrate the phase- and shift-invariance properties of complex cells [63], and the end-stopping property of hypercomplex cells [22].

**Masking by Natural Images**

From an evolutionary standpoint, an organism must be efficient both at detecting a visual target within its natural environment and at blending in with this environment to avoid detection. How effective are natural images at masking visual targets? As mentioned in Section 2.2.4, numerous models of visual masking have been quite successful at predicting detection thresholds for spatial targets placed upon relatively simplistic backgrounds [2][43][47][48][49][50]. Many of these models have been developed and refined to fit various TvC data [47][48][50], the majority of which employ a gain-control-type stage in which the combined response of several
mechanisms is used to regulate the excitatory response of a detecting mechanism. Indeed, variations of this framework have been incorporated into algorithms which have proved successful for image compression [39][40][41] and for predicting image discriminability [42][43].

2.3 Image Coding

Though higher-order HVS mechanisms are not yet well understood, vision research thus far has provided a general tenet for HVS-based image compression. The spatial-frequency decomposition performed by the HVS, for example, suggests the use of a Gabor- or wavelet-type decomposition. In turn, this type of representation allows exploitation of HVS characteristics such as contrast sensitivity and visual masking.

Figure 2.8 depicts the key steps involved in wavelet-based image compression. The original image is first subjected to a discrete wavelet transformation, which is then followed by quantization of the transform coefficients, and then the quantized data are entropy coded; decompression is performed by reversing this process.\(^2\) The following sections describe these steps in more detail.

\(^2\)The discrete wavelet transform and entropy coding are lossless steps and are therefore invertible. Quantization however, is a non-invertible process; see Equations (2.9), (2.12), and (2.13).
2.3.1 Discrete Wavelet Transform

State-of-the-art image compression algorithms attempt to mimic the multi-channel nature of the human visual system by employing a discrete wavelet transform (DWT) front-end which separates an image into spatial-frequency and orientation components, and which approximately decorrelates the image and provides good energy compaction. Although the DWT is not necessarily a good model of the decomposition performed during the early stages of human visual processing, the computational efficiency afforded by the DWT makes it particularly attractive for image compression and analysis (cf Ref. [64]). Specifically, the DWT is typically implemented via a filtering/lifting operation [65], usually in a separable fashion by successively processing the rows and the columns of the image. This operation results in a tiling of the spatial frequency plane, whereupon the image is represented as a series of spatial frequency bands (called subbands).

Figures 2.9 and 2.10 depict the frequency responses of the (one-dimensional)
filters and the corresponding tiling of the spatial frequency plane that results from a five-level DWT (using the 9/7 biorthogonal filters [66][67][5]). As shown in Figure 2.10, an $N$-level DWT will yield $3N + 1$ subbands; each level contains an $LH$ band, an $HL$ band, and an $HH$ band:

- **LH subbands** are low-pass filtered horizontally and high-pass filtered vertically and thus contain horizontal edge information.
- **HL subbands** are low-pass filtered vertically and high-pass filtered horizontally and thus contain vertical edge information.
- **HH subbands** are high-pass filtered in both directions and thus contain both $45^\circ$ and $135^\circ$ edge information.

The frequency content of each subband can be described by its center spatial frequency (given in cycles per pixel or cycles per degree of visual angle; see...
Ref. [67]) and by its predominant orientation. The three subbands in a level are collectively referred to as a *scale*, with finer scales containing higher-frequency information and coarser scales containing lower-frequency information. The coarsest scale contains four subbands, in which the additional band is the *LL* band (represented by the lower left of the tiling shown in Figure 2.10).

### 2.3.2 Quantization

After an image is transformed into its spatial-frequency representation, the coefficients within each subband are quantized. Quantization is a non-invertible process in which a continuous set of input values (e.g., subband coefficients) is approximated by a discrete set of output levels (called *reproduction values*). Let \( c(s) \) denote a coefficient of subband \( s \); and let \( C = \{ C_n \} \) denote a partition of the real line into contiguous, nonoverlapping intervals \( (C_n) \). A scalar quantizer operates by mapping each coefficient \( c(s) \in C_n \) to the reproduction value \( \hat{c}(s) \); the width of the interval \( C_n \) is called the *quantizer step size*, denoted \( \Delta_n \). In particular, if the reproduction values are midway between adjacent intervals, \( \hat{c}(s) \) is given by

\[
\hat{c}(s) = \Delta_n \times \left\lfloor \frac{c(s)}{\Delta_n} + \frac{1}{2} \right\rfloor.
\]  

(2.9)

Furthermore, if the intervals are equispaced (i.e., \( \Delta_n = \Delta \forall n \)), the quantizer is said to be *uniform* with a single quantizer step size \( \Delta \) [68].

Quantization is modeled by the addition of distortions to the original image. Specifically, quantization of a subband coefficient \( c(s) \) induces an error \( d(s) = \hat{c}(s) - c(s) \), which manifests itself in the reconstructed image as a wavelet ba-

---

3In some contexts, the rounding operation in Equation (2.9) is referred to as quantization, and the post-multiplication by the quantizer step size is referred to as *dequantization*. 
Figure 2.11: Quantization of a DWT subband induces artifacts in the reconstructed image; this process is modeled as the addition of distortions to the original image. The distortions depicted in this figure were generated by quantizing the LH subband at the fourth decomposition level (obtained using the 9/7 biorthogonal filters) with a step size $\Delta = 600$.

The basis function (distortion) whose amplitude is proportional to $d(s) \times |\psi(s)|$, where $\psi(s)$ represents the wavelet basis function associated with subband $s$. When all coefficients of subband $s$ are quantized, the resulting distortions constitute a superposition of scaled wavelet basis functions. The (mean-offset) distortions, $e$, are given by

$$ e = \hat{m} - m + \mu_m $$

(2.10)

where $m$ and $\hat{m}$ denote the original and reconstructed images, respectively; and where $\mu_m$ corresponds to the mean pixel-value of $m$. Thus, quantization is modeled as the addition of $e$ to $m$; i.e., an image reconstructed from quantized subbands is given by

$$ \hat{m} = m + e - \mu_m; $$

(2.11)

this process is illustrated in Figure 2.11. Notice in Figure 2.11 that the distortions are spatially correlated with the image.

During compression, the quantizer step size ($\Delta$) for each subband is typically
selected based on rate-distortion criteria using mean-squared error (MSE) as the distortion metric; this is the approach used in baseline JPEG-2000. However, in order to maximally compress an image in a visually lossless manner, the step sizes must be selected such that the resulting distortions are just below the threshold of visual detection. Furthermore, when rate constraints necessitate the generation of suprathreshold distortions, the step sizes must be selected so as to preserve the visual quality of the compressed image. Attaining these goals requires an understanding of both the visual detectability of the distortions, the masking effects imposed upon this detectability when the distortions are viewed within an image, and the effects suprathreshold distortions have on the visual quality of the image; these aspects are discussed in Chapters 3 and 7.

### 2.3.3 Entropy Coding

*Entropy coding*, which constitutes the final stage of most compression algorithms, serves to encode the quantizer’s output into fixed- or variable-length codewords. Unlike quantization, entropy coding is a lossless process which is entirely reversible. The basic strategy of an entropy coder is to assign shorter-length codewords to symbols (or groups of symbols) that appear more frequently and longer-length codewords to symbols (or groups of symbols) that occur less frequently in an attempt to achieve an average codeword length which approaches the entropy of the underlying data source [69].

*Huffman coding* and *arithmetic coding* are the two predominant entropy-coding schemes used in image compression:

- **Huffman codes** are variable-length instantaneous, prefix codes, which allow unique decoding without referring to future symbols. Huffman coding yields
integer-length codewords, and is therefore optimal only for sources in which the symbols exhibit power-of-two probabilities-of-occurrence [69]. For image compression, Huffman coding is often combined with run-length encoding due to the prevalence of long runs of the same symbol (namely zeros) output from the quantizer.

- **Arithmetic codes**, in contrast to Huffman codes, allow symbols to be assigned codewords represented by a fractional number of bits [70]. In this scheme, the codewords are defined based on a partition of the real interval $[0, 1]$; each codeword corresponds to variable-length partition of this interval, and the length of the partition is inversely related to the probability-of-occurrence of the associated symbol. Accordingly, the performance of an arithmetic coder is highly dependent on the accuracy at which these probabilities-of-occurrence can be estimated. For image compression, arithmetic coding is often combined with context modeling wherein the probability estimator is adapted online based on previously encoded data.

Huffman coding and arithmetic coding are used in JPEG and JPEG-2000, respectively. Although JPEG-2000’s arithmetic-coding scheme is more complex than the Huffman-coding scheme used in JPEG, this increased complexity alone affords a 5–10% improvement in compression performance [71]. As discussed in the following section, JPEG-2000 offers other key advantages over its predecessor.

### 2.3.4 JPEG-2000

JPEG-2000 is the next-generation lossy image compression standard, Part I of which was finalized in late 2001 (ISO/IEC FDIS15444-1:2000) [72]. JPEG-2000
Figure 2.12: Block diagram of the steps involved in JPEG-2000 compression.

provides several new features not found in JPEG, including resolution/quality scalability, region-of-interest coding, and both lossless and lossy compression. Figure 2.12 depicts the steps involved in JPEG-2000 compression; the following sections briefly describe these steps. (For more information on JPEG-2000, the reader is referred to Refs. [73] and [72].)

**Tiling**

The first step in JPEG-2000 compression involves dividing the original image into same-sized, nonoverlapping blocks called tiles; each tile is then compressed independently as if it were an entirely separate image. This tiling step facilitates both spatially selective decoding, and the use of JPEG-2000 in reduced-memory applications. Note however, that diving the image into tiles can also give rise to tiling artifacts which appear at the tile boundaries. For most screen-sized images, or when memory is not an issue, only a single tile is used, the dimensions of which are chosen to match those of the entire image.

**DC-level Shifting and Component Transform**

After the image is divided into tiles, *DC-level shifting* is performed on each tile. DC-level shifting, which is applied only to unsigned data, is performed by subtracting $2^{N_{\text{bits}}-1}$ from each data value, where $N_{\text{bits}}$ denotes the number of bits
with which each (gray, R, G, or B) data value is represented. Thus, for an 8-bpp grayscale image, DC-level shifting is performed by subtracting 128 from each pixel; for a 24-bpp RGB color image, 128 is subtracted separately from each color plane of each pixel.

For color images, an optional component transform is applied in which the RGB pixel values of each tile are transformed into luminance and chrominance components. This step, which is performed either via floating-point-based operations (so-called irreversible component transform) or via integer-based operations (reversible component transform, for lossless compression), corresponds roughly to an RGB-to-YCbCr transformation.

**Discrete Wavelet Transform**

Following the preprocessing steps, a discrete wavelet transform is applied to each tile. For lossy compression, JPEG-2000 permits the use of either the 5/3 filters or the 9/7 filters; for lossless compression, only the 5/3 filters are permitted. As described in Section 2.3.1, the DWT provides a subband-based representation of the image (here, of each image tile). These subbands are then subjected to quantization and entropy coding.

**Quantization**

Part I of the JPEG-2000 standard permits only scalar quantization of the coefficients within each subband (Part II allows the use of trellis-coded quantization). Scalar quantization in JPEG-2000 is performed by mapping each coefficient $c(s)$ of subband $s$ to quantization index $q_{\text{index}}(s)$ via

$$q_{\text{index}}(s) = \text{sgn}(c(s)) \left\lfloor \frac{|c(s)|}{\Delta_s} \right\rfloor$$  \hspace{1cm} (2.12)
where \( \Delta_s \) denotes the quantizer step size, and where \( sgn(\cdot) \) is the standard signum operator. At the decoder, dequantization is performed via

\[
\hat{c}(s) = \Delta_s \times q_{\text{index}}(s) \tag{2.13}
\]

where \( \hat{c}(s) \) denotes the reconstructed coefficient.

Two modes of scalar quantization are used in JPEG-2000—explicit quantization and implicit quantization:

- In **explicit quantization**, the step size \( \Delta_s \) used in Equation (2.12) is selected for each subband \( s \) based on some encoder-determined criterion (e.g., meeting a target bit-rate or MSE). In this case, the step sizes must be explicitly signaled in the compressed stream as side information (via QCC markers).

- In **implicit quantization**, only a single step size is selected for the LL subband, and all other step sizes are computed based on this step size (see Annex E.1 of Ref. [72]). In this case, only the single LL-band step size is signaled in the compressed stream (via a QCD marker).

As discussed in the following sections, implicit quantization facilitates the use of post-encoding-based rate control wherein the bulk of quantization is performed by truncating the bit-planes of the quantization indices.

**Tier 1 and Tier 2 Encoding**

After quantization, each subband of each tile is divided into same-sized, non-overlapping blocks; three co-located blocks from each scale (one from each of the LH, HL, and HH subbands) are collectively called a **precinct**. Each block of each precinct is further divided into same-sized, nonoverlapping blocks called **codeblocks**, and then entropy coding is applied independently to each codeblock.
Entropy coding in JPEG-2000 is performed via the EBCOT (Embedded Block Coding with Optimized Truncation) algorithm developed by Taubman [74]. EBCOT operates via two stages, called tiers, in which context modeling and arithmetic coding is performed in the first tier (Tier 1), and bit-stream ordering and truncation (for rate-control) is performed in the second tier (Tier 2). Tier 1 encoding is performed on each bit-plane of each codeblock via three coding passes (significance propagation, magnitude refinement, and cleanup) which collectively generate the compressed stream. This stream is then subjected to Tier 2 encoding in which the data are ordered based on, e.g., rate-distortion criteria via a Post-Compression Rate-Distortion (PCRD) optimization step. In particular, rate-control, if used, is efficiently performed during Tier 2 encoding by including in the final stream only those bit-planes which, e.g., minimize the distortion for the specified rate. Chapters 5, 6, and 8 describe contrast-based modifications of this scheme in which Tier 2 encoding is performed based on visual criteria.

2.4 Summary

This chapter presented an overview of spatial vision and image compression. An understanding of both paradigms is important for generating compressed images that are to be viewed by humans. In particular, to maximally compress an image in a visually lossless manner, the (compound) compression-induced distortions must be just below the threshold of visual detection. Achieving this goal requires an understanding of contrast sensitivity, visual masking, and visual summation of responses to the distortions. Furthermore, for low-rate applications that require highly compressed images which most-often contain suprathreshold distortions, a successful visually lossy compression algorithm must take into account the effects
these distortions impose on the visual quality of the reconstructed image.

The following chapters investigate these issues. Chapter 3 presents a study in which the detectability of simple and compound wavelet subband quantization distortions was measured both in the absence of a masker and when masked by natural images. Chapter 4 presents an algorithm for predicting these masked detection thresholds based on the results of an additional psychophysical experiment in which masked detection thresholds were measured using natural-texture maskers. Chapters 5 and 6 describe the application of these experimental results to visually lossless compression of natural and medical images, respectively. Chapter 7 describes an experiment involving contrast matching of suprathreshold wavelet subband quantization distortions. Chapter 8 describes the application of these suprathreshold results to visually lossy image compression.
Chapter 3

Visual Detection of Wavelet Distortions

3.1 Introduction

As described in Chapter 2, signal detection and discrimination have proved useful in characterizing many aspects of human vision. In this paradigm, a signal (target) is presented against a usually well-defined background (masker), and thresholds indicate the observer’s ability to discriminate the signal+background from the background (i.e., to detect the target in the presence of the masker.) The detectability of a visual target is of particular importance for visually lossless compression wherein the visual target consists of compression-induced distortions which must be kept below the threshold of detection. The application of psychophysical results to image compression however, requires several key assumptions which have yet to be validated:

1. It must be assumed that visual sensitivity to traditional targets, such as sine-wave gratings and Gabor patches, is similar to sensitivity to compression-induced distortions;

2. It must be assumed that summation of visual responses to these traditional targets is similar to summation of responses to distortions;

3. It must be assumed that results assessed in the absence of a masker are valid when distortions are presented against a natural image.
To address these issues, this chapter describes four psychophysical experiments in which detection thresholds were measured for targets consisting of wavelet subband quantization distortions (see Section 2.3.2). To quantify differences in visual responses to wavelet distortions versus traditional targets, Experiment I measured unmasked detection thresholds for simple wavelet subband quantization distortions generated via scalar quantization of individual DWT subbands. To quantify the effects of natural-image backgrounds on detection thresholds, Experiment II measured detection thresholds for simple wavelet subband quantization distortions presented against two natural-image maskers. To investigate summation of visual responses to wavelet distortions on orientation and spatial-frequency dimensions, Experiment III measured unmasked detection thresholds for compound wavelet subband quantization distortions generated via quantization of pairs of subbands. To quantify differences in unmasked summation versus summation in the presence of natural-image maskers, Experiment IV measured detection thresholds for compound wavelet distortions presented against two natural-image maskers.

This chapter organized as follows: Section 3.2 provides a survey of previous summation studies. Section 3.3 describes the methods and stimuli used in the experiments. Results, analyses, and a discussion are presented in Sections 3.4, 3.5, and 3.6. General conclusions are presented in Section 3.7.

### 3.2 Previous Work

#### 3.2.1 Summation on the Spatial-Frequency Dimension

Summation of visual responses to spatial-frequency components is tested by using compound targets containing similarly oriented components of different spatial
frequencies. The following list summarizes previous studies which have tested summation-at-threshold on the spatial-frequency dimension:

- Graham & Nachmias [75] found $RS \approx 1$ ($\beta \approx \infty$) using the method of adjustment and a two-alternative forced-choice paradigm and compound targets composed of two sine-wave components at spatial frequencies $f$ and $3f$; this result was reported to be invariant to the contrast ratio between the two components, invariant to the phase relationship between the two components, and invariant to the experimental paradigm (method of adjustment vs. forced-choice procedure).

- Sachs, Nachmias & Robson [29] found $RS \approx 1$ ($\beta \approx \infty$) using a Yes-No procedure and sine-wave targets composed of spatial frequencies $f_1 + f_2$; the components of the target were detected independently for most frequency ratios $f_1/f_2 \notin [0.80, 1.25]$.

- Watson [34] found $RS = 1.2$ ($\beta = 3.6$) using a two-alternative forced-choice paradigm and Gabor targets separated by roughly 1 octave in spatial frequency.

- Manahilov & Simpson [32] found $RS = 1.37$ ($\beta = 2.2$) using a two-alternative forced-choice paradigm and a compound target composed of 2+6 cycles/degree Gabor patches.

- Meinhardt [33] found $RS = 1.44$ ($\beta = 1.9$) using the method of limits and sine-wave targets composed of spatial frequencies in the range $[1, 5]$ cycles/degree.
3.2.2 Summation on the Orientation Dimension

Summation of visual responses to orientation components is tested by using compound targets containing components with similar spatial frequencies but different orientations. The following list summarizes previous studies which have tested summation-at-threshold on the orientation dimension:

- Carlson, Cohen, & Gorog [28] found $RS = 1.18$ ($\beta = 4.2$) using a Yes-No procedure and compound targets composed of two sine-wave components oriented at $0^\circ$ and $90^\circ$. $RS = 1$ ($\beta = \infty$) was found when the components were oriented at $45^\circ$ and $135^\circ$.

- Manahilov & Simpson [32] found $RS = 1.37$ ($\beta = 2.2$) using a two-alternative forced-choice (2AFC) paradigm and compound targets composed of a pair of 6 cycles/degree Gabor patches oriented vertically and at $45^\circ$.

3.2.3 Summation on the Spatial Dimension

Summation of visual responses to spatial components is tested by using compound targets containing spatially separated components with similar spatial frequencies and similar orientations. The following list summarizes previous studies which have tested summation-at-threshold on the spatial dimension:

- Watson, Yang, Solomon, & Villasenor [67] found $RS = 1.19$ ($\beta \approx 4.0$) using a two-alternative forced-choice paradigm and compound wavelet targets generated by adding values drawn from a uniform distribution to an empty wavelet subband and then performing an inverse DWT.

- Bonneh & Sagi [36] measured masked summation on spatial-extent and
Table 3.1: Summary of results from previous summation studies.

<table>
<thead>
<tr>
<th>Source</th>
<th>Dimension</th>
<th>Targets</th>
<th>Relative Sens.</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graham <em>et al.</em> [75]</td>
<td>Frequency</td>
<td>Sine waves</td>
<td>$\approx 1.0$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>Sachs <em>et al.</em> [29]</td>
<td>Frequency</td>
<td>Sine waves</td>
<td>$\approx 1.0$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>Watson [34]</td>
<td>Frequency</td>
<td>Gabor</td>
<td>1.20</td>
<td>3.6</td>
</tr>
<tr>
<td>Manahilov <em>et al.</em> [32]</td>
<td>Frequency</td>
<td>Gabor</td>
<td>1.37</td>
<td>2.2</td>
</tr>
<tr>
<td>Meinhardt [33]</td>
<td>Frequency</td>
<td>Gabor</td>
<td>1.44</td>
<td>1.9</td>
</tr>
<tr>
<td>Carlson <em>et al.</em> [28]</td>
<td>Orientation</td>
<td>Sine waves</td>
<td>$\approx 1.0$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>Manahilov <em>et al.</em> [32]</td>
<td>Orientation</td>
<td>Gabor</td>
<td>1.37</td>
<td>2.2</td>
</tr>
<tr>
<td>Bonneh <em>et al.</em> [36]</td>
<td>Space</td>
<td>Gabor</td>
<td>1.19</td>
<td>$\approx 4.0$</td>
</tr>
<tr>
<td>Watson <em>et al.</em> [67]</td>
<td>Space</td>
<td>Wavelets</td>
<td>1.19</td>
<td>$\approx 4.0$</td>
</tr>
</tbody>
</table>

spatial-position dimensions using a compound target composed of 12.5 cycle/degree Gabor patches presented on a 30% contrast 12.5 cycles/degree Gabor-patch masker. $RS \in [1.20, 1.23]$ ($\beta \in [3.3, 3.7]$) was found when the extent of the masker was fixed at its maximal size or number; $RS \in [1.16, 1.19]$ ($\beta \in [4.0, 4.6]$) was found on the spatial-position dimension and $RS = 1$ was found on the spatial-extent dimension when the extent of the masker was varied with that of the target.

### 3.2.4 Summary

In summary, numerous studies have investigated summation-at-threshold on spatial frequency, orientation, and spatial dimensions and masked summation on the spatial dimension; the results of these studies are summarized in Table 3.1. These
experiments have revealed relative sensitivities ranging from 1.0 ($\beta = \infty$; no summation) to 1.44 ($\beta = 1.9$) using either a uniform background (no mask) or an unnatural masker. The applicability of these results to image compression—in which targets are necessarily presented against an image masker—remains unclear. The following section describes four experiments designed to investigate this issue.

### 3.3 Experimental Methods

Four experiments were conducted to quantify the effects of natural images on the detectability of wavelet subband quantization distortions. In **Experiment I**, unmasked detection thresholds were measured for simple wavelet distortions. In **Experiment II**, detection thresholds were measured for these same distortions presented against two different natural-image maskers. In **Experiment III** unmasked detection thresholds were measured for compound wavelet distortions composed of either two octaves of spatial frequencies or two orientations. In **Experiment IV** detection thresholds were measured for these compound distortions presented against the same image maskers of Experiment II.

### 3.3.1 Apparatus

Stimuli were displayed on a high-resolution, non-interlaced Hewlett Packard A4033A 19-inch monitor at a display resolution of 36.4 pixels/cm, a frame rate of 75 Hz, and an overall gamma of 2.3. The display yielded minimum, maximum, and mean luminances of, respectively, 0.08, 48.2, and 13.3 cd/m$^2$. A bit-stealing procedure [76] was used to increase the effective luminance resolution to approximately 9-10 bits. Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 58 cm.
3.3.2 Stimuli

Stimuli consisted of 512×512-pixel luminance modulations which subtended 13.9×13.9 degrees of visual angle. Each stimulus was composed of a target and a mask: In all experiments, targets consisted of wavelet subband quantization distortions; in Experiments I and III (unmasked detection), the mask consisted of a uniform gray 10.1 cd/m² image; in Experiments II and IV, the mask consisted of one of two natural images.

Masks

Two natural images, balloon and horse, were used in all experiments as maskers and sources of the distortions; Figure 3.1 depicts these two images. Both images were of size 512×512 and were 8-bits/pixel (bpp) grayscale with digital pixel values in the range 0-255. The displayed images had mean physical luminances of 15.7 cd/m² (balloon) and 10.9 cd/m² (horse).

Targets

Wavelet targets were generated via uniform scalar quantization of one or two DWT subbands, yielding simple or compound distortions, respectively; the former were used in Experiments I and II and the latter were used in Experiments III and IV. The subbands were obtained by transforming a natural image of size 512×512 pixels using five decomposition levels and the 9/7 biorthogonal DWT filters.

- Simple targets were generated by uniformly quantizing an LH (or HL) subband of DWT level 1, 2, 3, 4, or 5, yielding a superposition of horizontally (or vertically) oriented wavelet basis functions centered at spatial frequencies 18.4, 9.2, 4.6, 2.3, or 1.15 cycles/degree, respectively. The quantizer step size
for the subband was selected such that the RMS contrast of the resulting distortions was as requested by the adaptive staircase procedure described in Section 3.3.3.

- **Compound targets** composed of two orientations were generated by uni-
formly quantizing the LH and HL subbands of DWT level 3, 4, or 5, yielding a superposition of horizontally and vertically oriented targets centered at spatial frequencies 4.6, 2.3, or 1.15 cycles/degree, respectively. Horizontally oriented compound targets composed of two (center) spatial frequencies were generated by uniformly quantizing the LH subbands of DWT levels 4 & 3 or 5 & 4, yielding a superposition of targets centered at spatial frequencies of $2.3 + 4.6$ cycles/degree or $1.15 + 2.3$ cycles/degree, respectively. Vertically oriented compound targets composed of the same pairs of (center) spatial frequencies were generated in a similar fashion via uniform quantization of the HL subbands. For each compound target, the relative contrasts of its components were equalized by using an estimate of each simple-target’s threshold (obtained from a previous study using the same images; see Ref. [4]). Quantizer step sizes were selected for the two subbands to meet this relative-contrast criterion and such that the RMS contrast of the compound target was as requested by the adaptive staircase procedure described in Section 3.3.3.

Following quantization of the subband(s), an inverse DWT was applied to generate a reconstructed image (target + mask) of size $512 \times 512$ pixels. For Experiments I and III, the mask (image) was subtracted from the reconstructed image, yielding only the target (wavelet distortions); the target was then added to an equally sized uniform gray image. This technique allowed unmasked presentation of targets while preserving the spatial correlations between the distortions and the original images (cf Figure 2.11).

Figure 3.2 depicts suprathreshold versions of the stimuli used in Experiment I. Figure 3.3 depicts suprathreshold versions of the stimuli used in Experiment II.
Figure 3.2: Representative stimuli used in Experiment I: (a), (b), and (c) depict horizontally oriented wavelet subband quantization distortions at center frequencies of 4.6, 2.3, and 1.15 cycles/degree, respectively. (d), (e), and (f) depict vertically oriented wavelet subband quantization distortions at center frequencies of 4.6, 2.3, and 1.15 cycles/degree, respectively. Distortions were generated by quantizing subbands from image *balloon*. Stimuli containing quantization distortions at center frequencies of 9.2 and 18.4 cycles/degree, and distortions generated from image *horse* are not depicted.
Figure 3.3: Representative stimuli used in Experiment II: Image balloon containing simple wavelet subband quantization distortions at center frequency 2.3 cycles/degree oriented (a) horizontally and (b) vertically; and at center frequency 1.15 cycles/degree oriented (c) horizontally and (d) vertically. Stimuli containing quantization distortions at center frequencies of 4.6, 9.2 and 18.4 cycles/degree, and stimuli containing image horse are not depicted.

Figure 3.4 depicts suprathreshold versions of the stimuli used in Experiment III.

Figure 3.5 depicts suprathreshold versions of the stimuli used in Experiment IV.
Figure 3.4: Representative stimuli used in Experiment III: Compound wavelet subband quantization distortions composed of horizontal+vertical components at center frequency (a) 2.3 cycles/degree and (b) 1.15 cycles/degree. Compound wavelet subband quantization distortions composed of two octaves of frequencies centered at 1.15+2.3 cycles/degree oriented (c) horizontally and (d) vertically. Distortions were generated by quantizing subbands from image horse. Stimuli containing quantization distortions at center frequencies of 4.6, 9.2 and 18.4 cycles/degree, and distortions generated from image balloon are not depicted.

3.3.3 Procedures

Thresholds were measured by using a spatial three-alternative forced-choice procedure. On each trial, observers concurrently viewed three adjacent images placed upon a uniform 10.1 cd/m$^2$ background. Two of the images contained the mask
Figure 3.5: Representative stimuli used in Experiment IV: Image *horse* containing compound wavelet subband quantization distortions composed of horizontal+vertical components at center frequency (a) 2.3 cycles/degree and (b) 1.15 cycles/degree. Image *horse* containing compound wavelet subband quantization distortions composed of two octaves of frequencies centered at 1.15+2.3 cycles/degree oriented (c) horizontally and (d) vertically. Stimuli containing quantization distortions at center frequencies of 4.6, 9.2 and 18.4 cycles/degree, and stimuli containing image *balloon* are not depicted.

alone and the other image additionally contained one of the previously described targets (distortions); the image to which the target was added was randomly selected at the beginning of each trial. Observers indicated via keyboard input which
of the three images contained the target (i.e., which image was distorted). Target RMS contrasts were guided via an adaptive QUEST staircase procedure [77] using software derived from the Psychophysics Toolbox [78][79]. Contrast detection threshold was defined as the 75%-correct point on a Weibull function, which was fitted to the data following each series of 32-trial tracks.

Each experimental session began with three minutes of adaptation to a blank 10.1 cd/m² display. Before each series of trials, observers were briefly shown a high-contrast, spatially randomized version of the distortions to facilitate a signal-known-exactly condition [80][45] (i.e., to minimize subjects’ uncertainty in the orientation and frequency of the target). During each trial, an auditory tone indicated stimulus onset, and auditory feedback was provided to denote correct and incorrect responses. Response time was limited to within 7 seconds of stimulus onset during which all three images remained visible; observers were instructed to examine all three images before submitting a response.

### 3.3.4 Observers

Three adult subjects (DC, SC, and MM¹) participated in the experiments. All observers were familiar with compression-induced distortions; however, only DC had previous exposure to the image maskers. All had either normal and corrected-to-normal visual acuity.

¹Subject MM did not participate in the parts of Experiments II and IV which tested summation on the spatial-frequency dimension.
3.3.5 Contrast Metric

Results are reported here in terms of RMS contrast [cf Equation (2.5)], which has also been used for compound [81], noise [82], wavelet [83], and natural-image [55] stimuli. Let \( t \) denote a target presented against an image mask \( m \); the RMS contrast of \( t \), \( C(t) \), is defined as

\[
C(t) \equiv \frac{1}{\mu_{L(m)}} \left( \frac{1}{N} \sum_{i=0}^{N} \left[ L(t_i) - \mu_{L(t)} \right]^2 \right)^{1/2}
\]  

(3.1)

where \( \mu_{L(m)} \) denotes the average luminance of the mask, \( \mu_{L(t)} \) denotes the average luminance of the target, \( L(t_i) \) denotes the luminance of the \( i^{th} \) pixel of \( t \), and \( N \) represents the total number of pixels.

The luminance values in Equation (3.1) are defined in standard units of cd/m\(^2\) given by [84]

\[
L = (\epsilon + kI)^{\gamma}
\]  

(3.2)

where \( L \) and \( I \) represent luminance and (8-bit) digital pixel value, respectively. The parameters \( \epsilon \), \( k \), and \( \gamma \) represent, respectively, the black-level offset, the pixel-value-to-voltage scaling factor, and the gamma of the display monitor. For the display monitor used in this experiment, these parameters were \( \epsilon = 0.33 \), \( k = 0.019 \), and \( \gamma = 2.3 \).

3.4 Detection of Simple Wavelet Distortions

This section describes two experiments performed to assess the effects of natural images on the detectability of simple wavelet subband quantization distortions. In Experiment I, unmasked detection thresholds were measured for simple wavelet distortions. In Experiment II, detection thresholds were measured for these same distortions presented against two different natural-image maskers.
3.4.1 Experiment I: Unmasked Detection

Experiment I measured unmasked detection thresholds for simple horizontal and vertical wavelet subband quantization distortions at center spatial frequencies of 1.15, 2.3, 4.6, 9.2, and 18.4 cycles/degree. Figure 3.6 depicts thresholds for each observer plotted as a function of the center spatial frequency of the distortions. These data indicate the minimum RMS contrast necessary to detect the target (distortions) in the absence of a masker. Each data point in Figure 3.6 represents the average of at least two blocks of trials; the error bars in each plot denote standard errors of the means.

These results are generally consistent with those of previous contrast threshold measurements; namely, similar to what has been found for gratings, the minimum contrast required to detect wavelet subband quantization distortions varies with spatial frequency, and equal sensitivity is observed for horizontal and vertical distortions. However, the data of Figure 3.6 generally reveal maximum sensitivity (minimum threshold) at 1.15 cycles/degree, whereas maximum sensitivity to sine-wave gratings is traditionally found to occur at 4-6 cycles/degree (see Chapter 2).

The variation in maximum sensitivity to sine-wave gratings versus wavelet sub-band quantization distortions might be attributable to differences in the bandwidths of the stimuli. Whereas sine-wave gratings occupy a single point in frequency, wavelet subband quantization distortions encompass an octave of spatial frequencies. Although current evidence (including results presented in Section 3.5) indicates that HVS spatial frequency channels have a tuning bandwidth of approximately 1–1.5 octaves, other studies employing stimuli with bandwidths near 1 octave have also found low-pass-shaped CSFs. For example, Peli et al. [85] and
Figure 3.6: Contrast detection thresholds for simple wavelet subband quantization distortions generated via quantization of images *balloon* and *horse* measured in the unmasked paradigm (Experiment I). *Black circles*: data for horizontal targets; *gray circles*: data for vertical targets. Error bars indicate ±1 SE. Note that the vertical axis represents increasing contrast in the downward direction.

Watson *et al.* [67] have reported maximum sensitivity at 1-2 cycles/degree for 1-octave Gabor patches and wavelets, respectively.
3.4.2 Experiment II: Detection in Natural-Image Backgrounds

Experiment II measured detection thresholds for simple horizontal and vertical wavelet subband quantization distortions at center spatial frequencies of 1.15, 2.3, 4.6, 9.2, and 18.4 cycles/degree presented against two different natural-image maskers (balloon and horse). Thresholds obtained for each observer are plotted in Figure 3.7 for images balloon and horse as a function of the spatial frequency of the target. These data indicate the minimum RMS contrast necessary to detect the target (distortions) in the presence of the corresponding natural-image masker. Each data point in Figure 3.7 represents the average of at least two blocks of trials; error bars denote standard errors of the means.

These results provide insight into the types of quantization distortions that are readily masked by natural images. Whereas in Experiment I maximum sensitivity occurred at 1.15 cycles/degree, these data show approximately equal sensitivity to 1.15, 2.3, and 4.6 cycles/degree targets. As mentioned in Chapter 2, previous spatial masking experiments have traditionally found greatest elevations in thresholds when the spatial frequency of the mask is nearby that of the target. Thus, the difference in maximum sensitivity between the data of Figures 3.6 and 3.7 might be attributable to the characteristic amplitude spectra of natural images; namely, the lower frequencies that predominate natural images might also reduce the visibility of lower-frequency distortions.

This notion is illustrated in Figure 3.8, which depicts contrast threshold elevations (TEs) between the masked and unmasked conditions (masked thresholds divided by unmasked thresholds). When the mask is a natural image composed predominantly of low spatial frequencies, the greatest elevation in threshold occurs for low frequency distortions (TE ≈ 8-9 at 1.15 cycles/degree for subject
Figure 3.7: Contrast detection thresholds for simple wavelet subband quantization distortions measured in the masked paradigm (Experiment II). Black circles: data for horizontal targets; gray circles: data for vertical targets. Error bars indicate ±1 SE. Light-gray data correspond to unmasked thresholds which are replotted from Figure 3.6. Note that the vertical axis represents increasing contrast in the downward direction.

DC), whereas distortions of high spatial frequencies incur only a minor elevation in threshold (TE ≈ 2-3 at 18.4 cycles/degree for subject DC).
Figure 3.8: Contrast threshold elevations (masked/unmasked) imposed by each natural image on the detectability of wavelet distortions. *Black circles*: data for horizontal targets; *gray circles*: data for vertical targets.

### 3.5 Detection of Compound Wavelet Distortions

Two additional experiments were performed to quantify summation of visual responses to wavelet subband quantization distortions both in the unmasked paradigm and in the presence of a natural-image masker. In *Experiment III* unmasked
detection thresholds were measured for compound wavelet distortions composed of either two octaves of spatial frequencies or two orientations. In Experiment IV detection thresholds were measured for these compound distortions presented against the same image maskers of Experiment II.

3.5.1 Experiment III: Unmasked Summation

Experiment III measured unmasked detection thresholds for compound wavelet targets composed either of two orientations (horizontal+vertical) both at center spatial frequencies of 1.15, 2.3, or 4.6 cycles/degree, or of two octaves of spatial frequencies (2.3 + 4.6 cycles/degree and 1.15 + 2.3 cycles/degree) both oriented horizontally or vertically. Results of this experiment were compared with those of Experiment I to quantify summation-at-threshold on orientation and spatial-frequency dimensions in the absence of a masker. As described in Chapter 2, if the visual responses to the components of a compound wavelet target exhibit summation, then we expect a reduction in detection thresholds for the components within the compound target (measured in Experiment III) versus thresholds measured for those same components presented as simple targets (Experiment I). Using the notation presented in Section 2.2.3: \( CT(t_i) \) were measured in Experiment I; \( CT(t_i|t) \) were measured in Experiment III.

Figures 3.9 and 3.10 depict relative contrast thresholds of Experiment III in the form of summation-square plots [11] (see Figure 2.4) for orthogonal components of equal spatial frequencies and for equally oriented components of different center spatial frequencies, respectively. In each plot, the horizontal axis represents the relative contrast of one component [i.e., \( RC(t_1) \)] and the vertical axis represents the relative contrast of the other component [\( RC(t_2) \)]. The points within each plot
denote relative contrast threshold pairs

\[ (RC(t_1), RC(t_2)) = (RCT(t_1; \mathbf{t}), RCT(t_2; \mathbf{t})) \]

computed using each observer’s average contrast threshold for each component (from Experiment I); each data point represents the average of at least two blocks of trials, and the solid lines in each plot represent fits of Equation (2.8) to each point.

Recall that for linear summation \((RS = 2; \beta = 1)\) the data points would lie on the diagonal line connecting \(RC\) coordinates (0,1) to (1,0), and for no summation \((RS = 1; \beta = \infty)\) the points would lie on the lines formed by connecting (0,1) to (1,1) and (1,1) to (1,0) (see Figure 2.4).

Tables 3.2 and 3.3 list relative sensitivities computed from the data of Figures 3.9 and Figure 3.10, respectively. Mean relative sensitivities (over all three observers), computed by using the procedure described by Watson [34], are as follows:

- **Summation on the orientation dimension** yielded mean relative sensitivities of 1.21 \((\beta = 3.64)\) and 1.15 \((\beta = 5.06)\) for targets generated from images *balloon* and *horse*, respectively; with an overall mean (over both images) of 1.19 \((\beta = 3.97)\).

- **Summation on the spatial-frequency dimension** yielded mean relative sensitivities of 1.15 \((\beta = 5.01)\) and 1.13 \((\beta = 5.62)\) for targets generated from images *balloon* and *horse*, respectively; with an overall mean of 1.14 \((\beta = 5.30)\).

These results, which are consistent with previous summation experiments (see Section 3.2), indicate that in the absence of a masker, visual responses to wavelet subband quantization distortions pool in a fashion similar to responses to sine-wave
Figure 3.9: Relative contrast thresholds (RCTs) measured in Experiment III for compound wavelet subband quantization distortions containing orthogonal components of equal spatial frequencies. Distortions were generated by quantizing the LH and HL subbands of images balloon and horse. The horizontal and vertical axes of each plot represent the relative RMS contrasts of the horizontal and vertical components, respectively; data points indicate relative contrast thresholds. Open circles: RCTs for 4.6 cycles/degree components; gray circles: RCTs for 2.3 cycles/degree components; black circles: 1.15 cycles/degree components. Solid lines indicate fits of Equation (2.8) to each data point.
Figure 3.10: Relative contrast thresholds (RCTs) measured in Experiment IV for compound wavelet subband quantization distortions containing equally oriented components of different center spatial frequencies. Distortions were generated by quantizing the LH\(_n\) (HL\(_n\)) and LH\(_{n+1}\) (HL\(_{n+1}\)) subbands \([n = 3, 4]\) of images balloon and horse. The horizontal and vertical axes represent the relative RMS contrasts of the higher-frequency and lower-frequency components, respectively; data points denote relative contrast thresholds. **Black circles:** RCTs for horizontal 2.3+4.6 cycles/degree; **gray circles:** RCTs for vertical 2.3+4.6 cycles/degree; **black squares:** RCTs for horizontal 1.15+2.3 cycles/degree; **gray squares:** RCTs for vertical 1.15+2.3 cycles/degree. Solid lines indicate fits of Equation (2.8) to each data point.

gratings and Gabor patches. These data provide further evidence that HVS orientation and spatial-frequency channels have approximately 1-octave tuning bandwidths. Thus, when the components of the compound target are far enough apart
Table 3.2: Relative sensitivities for unmasked wavelet distortions generated via quantization of subbands from image *balloon*.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Components</th>
<th>Relative Sens.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.27</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.45</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.14</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.27</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.00</td>
<td>0.10</td>
</tr>
<tr>
<td>SC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.17</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.26</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.31</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.08</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.34</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.35</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.30</td>
<td>0.10</td>
</tr>
<tr>
<td>MM</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.00</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.24</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.33</td>
<td>0.06</td>
</tr>
</tbody>
</table>

along the dimension of interest (here, 1 octave, 90°), doubling the number of components does not halve the contrasts of the components when the compound is at threshold; rather, the contrast of each component is reduced only by a scaling factor of approximately 0.85 (corresponding to $\beta \approx 4.5$).
Table 3.3: Relative sensitivities for unmasked wavelet distortions generated via quantization of subbands from image horse.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Components</th>
<th>Relative Sens.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.23</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.18</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.17</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.29</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.08</td>
<td>0.00</td>
</tr>
<tr>
<td>SC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.29</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.13</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.08</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.08</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.33</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.00</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.20</td>
<td>0.02</td>
</tr>
<tr>
<td>MM</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.16</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.11</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.19</td>
<td>0.00</td>
</tr>
</tbody>
</table>

3.5.2 Experiment IV: Summation in Natural-Image Backgrounds

Experiment IV measured masked thresholds for the same compound wavelet targets of Experiment III (see Section 3.5.1) using natural-image maskers balloon
and horse. Thresholds measured in this experiment were compared to those from Experiment II to quantify summation of visual responses on orientation and spatial-frequency dimensions for wavelet distortions presented against natural-image maskers. Figures 3.11 and 3.12 depict the relative contrast thresholds. The horizontal and vertical axes each represent the relative contrast of a component within the compound target. Data points denote relative contrast threshold pairs

\[(RC(t_1), RC(t_2)) = (RCT(t_1; t), RCT(t_2; t))\]

computed using each observer’s average contrast threshold for each component (from Experiment II). Each data point represents the average of at least two blocks of trials; solid lines represent fits of Equation (2.8) to each point. Relative contrast thresholds for compound targets composed of orthogonal components of equal spatial frequencies are shown in Figure 3.11. Relative contrast thresholds for compound targets composed of equally oriented components of different center spatial frequencies are shown in Figure 3.12.

Tables 3.4 and 3.5 list relative sensitivities computed from the data of Figures 3.11 and 3.12. Mean relative sensitivities (over all three observers), computed by using the procedure described by Watson [34], are as follows:

- **Summation on the orientation dimension** yielded mean relative sensitivities (over all three observers) of 1.53 ($\beta = 1.62$) and 1.56 ($\beta = 1.57$) for targets generated from images balloon and horse, respectively; with an overall mean (over both images) of 1.54 ($\beta = 1.60$).

- **Summation on the spatial-frequency dimension** yielded mean relative sensitivities of 1.71 ($\beta = 1.29$) and 1.65 ($\beta = 1.39$) for targets generated from images balloon and horse, respectively; with an overall mean of 1.68 ($\beta = 1.34$).
Figure 3.11: Relative contrast thresholds (RCTs) measured in the masked paradigm (Experiment IV) for compound wavelet subband quantization distortions containing orthogonal components of equal spatial frequencies. Distortions were generated by quantizing the LH and HL subbands of the natural images described in Section 3.3. The horizontal and vertical axes of each plot represent the relative RMS contrasts of the horizontal and vertical components, respectively; data points denote relative contrast thresholds. Open circles: RCTs for 4.6 cycles/degree components; gray circles: RCTs for 2.3 cycles/degree components; black circles: RCTs for 1.15 cycles/degree components. Solid lines indicate fits of Equation (2.8) to each data point.
Figure 3.12: Relative contrast thresholds (RCTs) measured in the masked paradigm (Experiment IV) for compound wavelet subband quantization distortions containing equally oriented components of different center spatial frequencies. Distortions were generated by quantizing the \( \text{LH}_n \) (\( \text{HL}_n \)) and \( \text{LH}_{n+1} \) (\( \text{HL}_{n+1} \)) subbands \([n = 3, 4]\) of the natural images described in Section 3.3. The horizontal and vertical axes represent the relative RMS contrasts of the higher-frequency and lower-frequency components, respectively; data points denote relative contrast thresholds. Black circles: RCTs for horizontal 2.3+4.6 cycles/degree; gray circles: RCTs for vertical 2.3+4.6 cycles/degree; black squares: RCTs for horizontal 1.15+2.3 cycles/degree; gray squares: RCTs for vertical 1.15+2.3 cycles/degree. Solid lines indicate fits of Equation (2.8) to each data point.

These data indicate that when wavelet subband quantization distortions are presented against a natural-image masker, summation of visual responses is significantly greater than what was found for these same distortions in the unmasked...
Table 3.4: Relative sensitivities for masked wavelet distortions generated via quantization of subbands from image *balloon*.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Components</th>
<th>Relative Sens.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.68</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.63</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.59</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.45</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.50</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>2.25</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>2.09</td>
<td>0.27</td>
</tr>
<tr>
<td>SC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.71</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.36</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.76</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.66</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.69</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.77</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.55</td>
<td>0.16</td>
</tr>
<tr>
<td>MM</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.36</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.42</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.42</td>
<td>0.48</td>
</tr>
</tbody>
</table>

paradigm. These high relative sensitivities are perhaps attributable to within-channel summation, which may arise due to the spatial correlations that exist between the distortions and the image maskers. Similarly, increased summation might be attributable to *off-frequency looking* [37][86] a phenomenon in which the
Table 3.5: Relative sensitivities for masked wavelet distortions generated via quantization of subbands from image *horse*.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Components</th>
<th>Relative Sens.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.92</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.35</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.65</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.64</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.90</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.78</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.73</td>
<td>0.36</td>
</tr>
<tr>
<td>SC</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.72</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.26</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.33</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>H 1.15+2.3 cycles/degree</td>
<td>1.86</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>V 1.15+2.3 cycles/degree</td>
<td>1.27</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>H 2.3+4.6 cycles/degree</td>
<td>1.47</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>V 2.3+4.6 cycles/degree</td>
<td>1.73</td>
<td>0.06</td>
</tr>
<tr>
<td>MM</td>
<td>H+V 1.15 cycles/degree</td>
<td>1.63</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>H+V 2.3 cycles/degree</td>
<td>1.60</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>H+V 4.6 cycles/degree</td>
<td>1.81</td>
<td>0.64</td>
</tr>
</tbody>
</table>

channel with the greatest signal-to-noise ratio (SNR) is used for detection, even though that channel might be less-optimally tuned to the target. In this case, if the image masker lowers the SNR of the channel tuned to one component, the other channel (i.e., the channel optimal for the other component, and “switched to” for
the first component) might elicit detection for both components, again resulting in within-channel summation.

3.6 Analysis and Discussion

Previous studies have investigated the detectability of simple and compound stimuli presented against no mask or against well-defined maskers. In this study, we have examined the detectability of simple and compound wavelet subband quantization distortions presented against no mask and two natural-image maskers. Detection thresholds measured in the unmasked paradigm were compared with detection thresholds measured when the distortions were presented against two natural-image backgrounds.

3.6.1 Detection of Simple Wavelet Distortions

Unmasked Detection

Contrast sensitivity functions (CSFs) for sine-wave gratings traditionally demonstrate a band-pass profile, with a peak at 4-6 cycles/degree [27][85]. Contrast sensitivity functions for Gabor patches usually demonstrate a low-pass profile, peaking at 0.5-3 cycles/degree, depending on the bandwidth of the grating and the temporal nature of stimulus presentation [85]. Here, the results of Experiment I indicate that contrast sensitivity to simple wavelet subband quantization distortions measured in the absence of a masker displays a maximum at spatial frequencies at least as low as 1.15 cycles/degree. The resulting CSF demonstrates a low-pass profile (see Figure 3.6) that is consistent with what has been reported in the literature for targets of similar bandwidth [85][67][87]. For example, the
average -3 dB cut-off frequency for CSFs surveyed by Peli et al. [85] was approximately 3-5 cycles/degree for 1-octave Gabor patches (estimated by eye from Figure 5 of Ref. [85]). The results of Experiment I indicate -3 dB cut-offs well within this range: 3-4 cycles/degree for DC and SC; 4-5 cycles/degree for MM.

**Masked Detection**

When simple wavelet subband quantization distortions are presented against one of the two natural images used in this study, the results of Experiment II indicate that the masker imposes both frequency-selective and image-selective effects of detection thresholds. Average threshold elevations (for subject DC) were approximately 4-6, with minimum and maximum elevations of approximately 2-3 and 8-9 for 18.4 cycles/degree and 1.15 cycles/degree distortions, respectively. This finding is consistent with the results of Webster et al. [53] who measured detection thresholds for gratings following adaptation to various natural images. Webster et al. found a maximum threshold elevation (post-adapt/pre-adapt) at 0.25-1 cycles/degree, a minimum threshold elevation at approximately 16 cycles/degree, and a threshold elevation profile not unlike those depicted in Figure 3.8 (see Figure 6 in Ref. [53]). Chapter 4 describes a model which provides a reasonable fit to these data.

### 3.6.2 Summation of Responses to Wavelet Distortions

**Unmasked Summation**

Relative sensitivities measured in the absence of a masker were consistent with those found in previous summation-at-threshold studies. Summation on the orientation dimension was independent of the spatial frequencies tested (1.15, 2.3, and 4.6 cycles/degree), and summation on the spatial-frequency dimension was
independent of the orientations tested (horizontal and vertical). Thus, for the ori-
entations and range of frequencies tested here, our data suggest that responses to
wavelet subband quantization distortions pool in a fashion similar to responses to
sine-wave and Gabor-patch targets, despite the fact that wavelet subband quanti-
ization distortions occupy a band of spatial frequencies as opposed to traditional,
single-frequency gratings.

**Masked Summation**

As mentioned in Section 3.2, Bonneh *et al.* [36] found $RS \approx 1.2$ ($\beta \approx 4$) on
the spatial dimension using Gabor-patch targets which were spatially correlated
with Gabor-patch maskers. Here, we have found much greater summation, $RS \approx$
1.5-1.7 ($\beta \approx 1.3-1.6$), on orientation and spatial frequency dimensions for wavelet
targets which were spatially correlated with natural-image maskers. A possible
explanation for our finding is the presence of within-channel summation, wherein a
single channel is used to detect both components of the compound target. Within-
channel summation may result from the spatial correlations that exists between
the distortions and the images and/or from off-frequency looking.

Unlike unnatural maskers, natural images contain relatively broadband and
spatially localized frequency content. The effectiveness of a natural image at mask-
ing a particular target, thus, depends not only on the nature of the target, but also
on where in the image this target is located. In the context of wavelet subband
quantization, because the basis functions (i.e., the wavelets) are localized both in
frequency and in space, so too are the quantization distortions. Quantizing an
LH subband, for example, will induce distortions that are localized to the hori-
zontal edges within the image. Similarly, quantizing an HL subband will induce
distortions that are localized to the vertical edges within the image. When both of these subbands are quantized together, the nature of the combined distortions will therefore depend on the image’s local frequency content. It is therefore reasonable to assume that there exist regions within the image that are particularly effective at masking simple targets, but are only marginally effective at masking compound targets, and vice versa.

For example, consider the two segments of a natural image [4] depicted in Figure 3.13. Because the horizontal and vertical content of the antenna is spatially separated, so too are the components of the compound target formed when both of these subbands are quantized. The image segment with the people however, contains overlapped edges of multiple orientations; in this case, the components of the compound target are spatially superimposed. Thus, depending on where in the image observers looked, it is possible for this one natural image masker to give rise a range of summation values, including within-channel summation.

It remains unclear whether within-channel summation is a result of spatial correlations or off-frequency looking, or a combination of both paradigms. The effects of the spatial correlations might be investigated by measuring summation in the presence of a phase-randomized natural-image masker (e.g., $1/f$ noise), whereas off-frequency looking might be investigated by assessing summation on the orientation dimension to components at a particular spatial scale in the presence of an image masker whose corresponding spatial scale has been filtered out. Off-frequency looking might also be tested by using compound targets containing components which are well-separated on the spatial-frequency dimension such that the channel “switched to” for one component is not the same channel used for detecting the other component.
Figure 3.13: Two regions of the image *wall* used in a previous masked-summation study [4] yielding nearly linear summation. The top region, “antenna,” contains horizontal and vertical components that are spatially separated, whereas the bottom region, “people,” contains overlapped edges of multiple orientations. The bottom two thumbnails depict these regions reconstructed from quantized LH and HL subbands at the third decomposition level.

### 3.6.3 Reconstructed Images

We have used the masked detection thresholds and summation results to generate images containing compound distortions induced via quantization of 10 subbands.
(5 LH bands, 5 HL bands). Because 10 subbands are quantized simultaneously, the resulting compound distortion contains $M = 10$ components, and we accordingly assume $RS = M^{1/\beta} = 10^{1/\beta}$ when the distortions are proportioned such that they are equally detectable (see Chapter 2). Thus, to produce a reconstructed image in which the compound distortion $t$ is at threshold, the contrast $C(t_i)$ of each (simple distortion) component $t_i$ is selected as

$$C(t_i) = 10^{-1/\beta} CT(t_i)$$

(3.3)

where $CT(t_i)$ denotes the contrast threshold of $t_i$ as measured in Experiment II (see Figure 3.7).

Figures 3.14 and 3.15 depict image balloon reconstructed from quantized subbands (obtained via the 9-7 biorthogonal filters and 5 decomposition levels) using the contrast thresholds measured for subject DC (for image balloon) and two summation rules: $\beta = 1.5$ (Figure 3.14) and $\beta = \infty$ (Figure 3.15). Figures 3.16 and 3.17 depict image horse reconstructed from quantized subbands using the contrast thresholds measured for subject DC (for image horse) and two summation rules: $\beta = 1.5$ (Figure 3.16) and $\beta = \infty$ (Figure 3.17).

In Figures 3.14 and 3.16 the subbands were quantized such that the RMS contrasts of the distortions are approximately 22% of the thresholds specified in Figure 3.7 (i.e., $\beta = 1.5$). In Figures 3.15 and 3.17 the subbands were quantized such that the RMS contrasts of the distortions are exactly as specified by the thresholds of Figure 3.7 for subject DC (i.e., $\beta = \infty$). Notice that the distortions in the images generated using $\beta = \infty$ are suprathreshold. Pay particular attention to the sky region near the top of balloon and near the horse’s ear, the interior of the symbol “11” located in the center of the balloon’s envelope, and the fur on the horse’s side. These images were designed to be viewed on an sRGB [88] display.
Figure 3.14: Image *balloon* reconstructed from quantized LH and HL subbands using $\beta = 1.5$. The distortions should not be visible in this image; see text for details.

from a distance of approximately three picture heights; to facilitate viewing, these images are also available online [89].

### 3.7 Summary

In this chapter, we have investigated the detectability of wavelet subband quantization distortions. Contrast detection thresholds were measured for both simple and compound distortions presented in the unmasked paradigm and against two natural-image maskers. The distortions were generated via uniform scalar quanti-
Figure 3.15: Image balloon reconstructed from quantized LH and HL subbands using $\beta = \infty$. The distortions in this image should be suprathreshold; pay particular attention to the sky region near the top of balloon and the interior of the symbol “11” located in the center of the balloon’s envelope; see text for details.

zation of individual subbands and pairs of subbands. Unmasked detection thresholds for simple distortions yielded contrast sensitivity functions similar to those reported for 1-octave Gabor patches. Detection thresholds for simple distortions presented against two natural-image backgrounds revealed that thresholds were elevated across the frequency range of 1.15–18.4 cycles/degree, with the greatest elevation for low-frequency distortions. Unmasked thresholds for compound distortions revealed relative sensitivities of 1.1-1.2, suggesting that summation of responses to wavelet distortions is similar to summation to gratings. Masked thresh-
olds for compound distortions revealed relative sensitivities of 1.5-1.7, suggesting greater summation when distortions are masked by natural images.

Note that we have not tested combinations of spatial frequencies other than those described for Experiments III and IV (i.e., $N = 2$, H+V, 1.15+2.3 cycles/degree, 2.3+4.6 cycles/degree). It is possible that other combinations of distortions require the use of different summation exponents. Additional testing is also required to determine summation involving distortions generated via quantization of the HH subbands. Furthermore, the use of these results for general wavelet-based image compression requires a model which can predict thresholds

Figure 3.16: Image horse reconstructed from quantized LH and HL subbands using $\beta = 1.5$. The distortions should not be visible in this image; see text for details.
Figure 3.17: Image *horse* reconstructed from quantized LH and HL subbands using $\beta = \infty$. The distortions in this image should be suprathreshold; pay particular attention to the sky region near the horse’s ear and the fur on the horse’s side; see text for details.

on a per-image basis; the following chapter describes such an algorithm.