February 1997

A Methodology for Designing Image Similarity Metrics Based on Human Visual System Models

Thomas Frese
Purdue University School of Electrical and Computer Engineering

Charles A. Bouman
Purdue University School of Electrical and Computer Engineering

Jan P. Allebach
Purdue University School of Electrical and Computer Engineering

A Methodology for Designing Image Similarity Metrics Based on Human Visual System Models

Thomas Frese
Charles A. Bouman
Jan P. Allebach

TR-ECE 97-2
February 1997

School of Electrical
and Computer Engineering
Purdue University
West Lafayette, Indiana 47907-1285

A Methodology for Designing Image Similarity Metrics Based on Human Visual System Models

Thomas Frese, Charles A. Bouman and Jan P. Allebach

Technical Report TR-ECE 97-2

School of Electrical and Computer Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN 47907-1285

Contents

List of Figures

List of Tables

Abstract

In this report we present an image similarity metric for content-based image database search. The similarity metric is based on a multiscale model of the human visual system. This multiscale model includes channels which account for perceptual phenomena such as color, contrast, color-contrast and orientation selectivity. From these channels, we extract features and then form an aggregate measure of similarity using a weighted linear combination of the feature differences. The choice of features and weights is made to maximize the consistency with similarity ratings made by human subjects. In particular, we use a visual test to collect experimental image matching data. We then define a cost function relating the distances computed by the metric to the choices made by the human subject. The results indicate that features corresponding to contrast, color-contrast and orientation selectivity can significantly improve search performance. Furthermore, the systematic optimization and evaluation strategy using the visual test is a general tool for designing and evaluating image similarity metrics.

1 Introduction

In recent years the use of digital imagery has become an important part of computer and telecommunication systems. As a result of advanced computer technology, image databases containing thousands of images have evolved in many applications. The efficient use of these databases requires new database organization and image retrieval methods.

Currently, it is common practice to manually annotate image databases by describing each image with a small set of keywords. However, the manual annotation is not only expensive but also incomplete. Since most images are far too complex to be completely described, the keywords are typically limited to represent the most important objects in the image. Furthermore, it is difficult to be consistent with the choice of keywords for each image in the database. As a result, image databases often must be re-annotated for different users with different search requirements. A more basic problem with text annotations is the inability of the the keyword representation to adequately capture visual aspects of images such as color and spatial arrangement. While the language description is usually connected to the recognition of objects in the image, a large part of human visual perception does not rely on recognition or interpretation. Humans have a graphical memory which stores the appearance of images, and often when we cannot remember the actual objects in an image, we can extract information from the memorized visual appearance. It seems therefore unnatural to search and compare images by representing them in the language domain.

The above difficulties have led to the development of content-based retrieval methods for image database search. Over the past ten years, there has been considerable research activity in the field, resulting in hundreds of publications and several conferences devoted to this topic [1, 2]. Much of this research has been motivated by an increasing number of applications for image database search. With the development of multimedia systems and large data networks, an increasing number of interest groups have gained access to thousands of images. For example, the medical community is forming contiinent-wide inter-hospital networks to allow content-based image search for the diagnosis of rare diseases [3]. Law enforcement agencies are interested in face recognition for subject identification [4]. Further important areas of interest are architecture, art history, astronomy, geology, multimedia, satellite imagery and TV production. Due to the exponential growth of the World Wide Web and the introduction of electronic imaging equipment to the consumer electronics market, we expect a rising demand for image database organization tools for commercial applications such as electronic mail order catalogs as well as for home usage such as electronic home photography.

2 Previous Approaches

Previous approaches have defined image similarity metrics using classical image processing techniques. The first content-based search algorithms were intended to retrieve CAD drawings from technical databases [5, 6, 7]. For these tasks, it was assumed that larger images were manually segmented into objects which could be searched for in the database. However, these drawings contained well defined objects which makes the task quite different from retrieving natural images.

To obtain more general metrics of image similarity, people have defined discriminants based on color histograming, color clustering or Bayesian color segmentation [8, 9]. Color histograms of the entire image have the advantage of being invariant to spatial perturbations and of being computationally inexpensive. However, color histogram methods are too invariant to be consistent with the human perception of image similarity. Different spatial arrangements of similar objects or different perspectives of a similar scent: may have the same color histogram but appear very different to a human observer.
Other approaches have been based on shape and curvature features [10, 11]. These concepts work well for binary images which contain clearly distinct objects. However, in natural images it is usually not possible to extract and match the edges of meaningful objects.

Generally, as [12] points out, metrics based on a single discriminant can only capture some but not all aspects of image similarity since the precomputed databast: representation is incomplete. More recent approaches use image compression techniques to generate perceptually complete image representations. While some of the published similarity metrics directly compare the compression coefficients of, for example, the wavelet compressed [13] images, others use the compressed representation to extract features for different aspects of similarity [12, 14]. MIT's photobook [12] for example, extracts different features for appearance, texture and shape using image processing methods such as the Karhuenen-Loeve transformation.

Although these metrics perform well for the task of comparing images which contain distinct objects, it is questionable how well they relate to human image similarity perception of natural images. The impressive progress made in the field of image quality assessment by employing models of the early human visual system suggests that the use of visual system models for general image comparison may significantly improve similarity metric performance. In this work we propose an approach to image similarity using features extracted from a simple multiscale model of the human visual system. In order to identify features which relate well to human perception, we developed a feature selection and optimization strategy based on experimental image matching data.

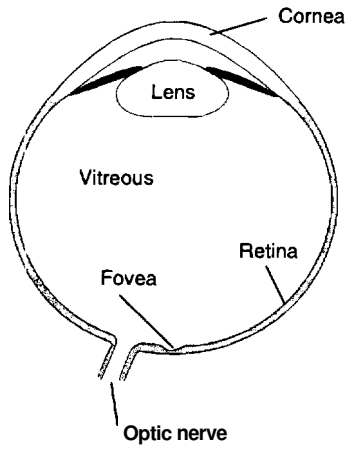

Figure 1: The optics of the human eye. Incoming light is focused by the cornea and lens onto the retina which contains the photoreceptor~.

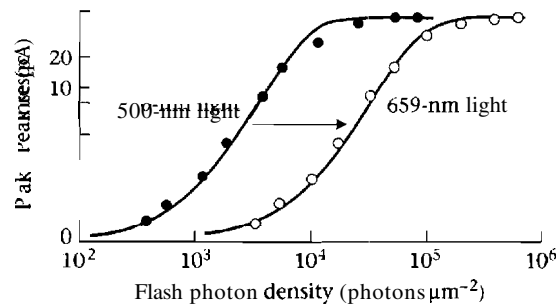

Figure 2: Nonlinear light to photocurrent conversion. The figure shows the photocurrent peak response as a function of the photon intensity of a flash stimulus. Figure from [15].

3 The human visual system and existing models

A natural approach to deriving image similarity metrics which relate well to human perception is to extract features from models of the human visual system. In order to derive such a metric, we need to understand the basics of the human visual system.

3.1 The early human visual system

Figure 1 shows a cross section of the imaging components of the human eye. The incoming light is focused by the lens and passed through the transparent vitreous before it falls on the retina. The retina is a thin layer of tissue containing the photoreceptors and several layers of interconnected neurons. The photoreceptors contain photopigment which converts the incoming light to electric current. The light to current conversion is nonlinear as shown in Fig. 2. The functional relationship within the non-saturated range has long been modeled as being logarithmic, however certain perceptual effects cannot be explained by the log relationship, so that more recent results indicate a power law function of approximately $x^{1/3}$.

Figure 3 shows a simplified block diagram of the retina. The photoreceptors can be distinguished into two basic types which are called rods and cones. While the rods are achromatic sensors, the cones are color sensitive. Based on the spectral sensitivity of their photopigment, the cones can further be subdivided into long (L), medium (M) and short (S) wavelength cones. Figure 4 shows the relative cone sensitivities as a function of wavelength. The cones are comparatively large and highly concentrated in the center of the retina whereas the smaller rods are more concentrated in the periphery. Note that the number of S cones is much smaller than the number of L and M cones. A potential reason for this is that due to

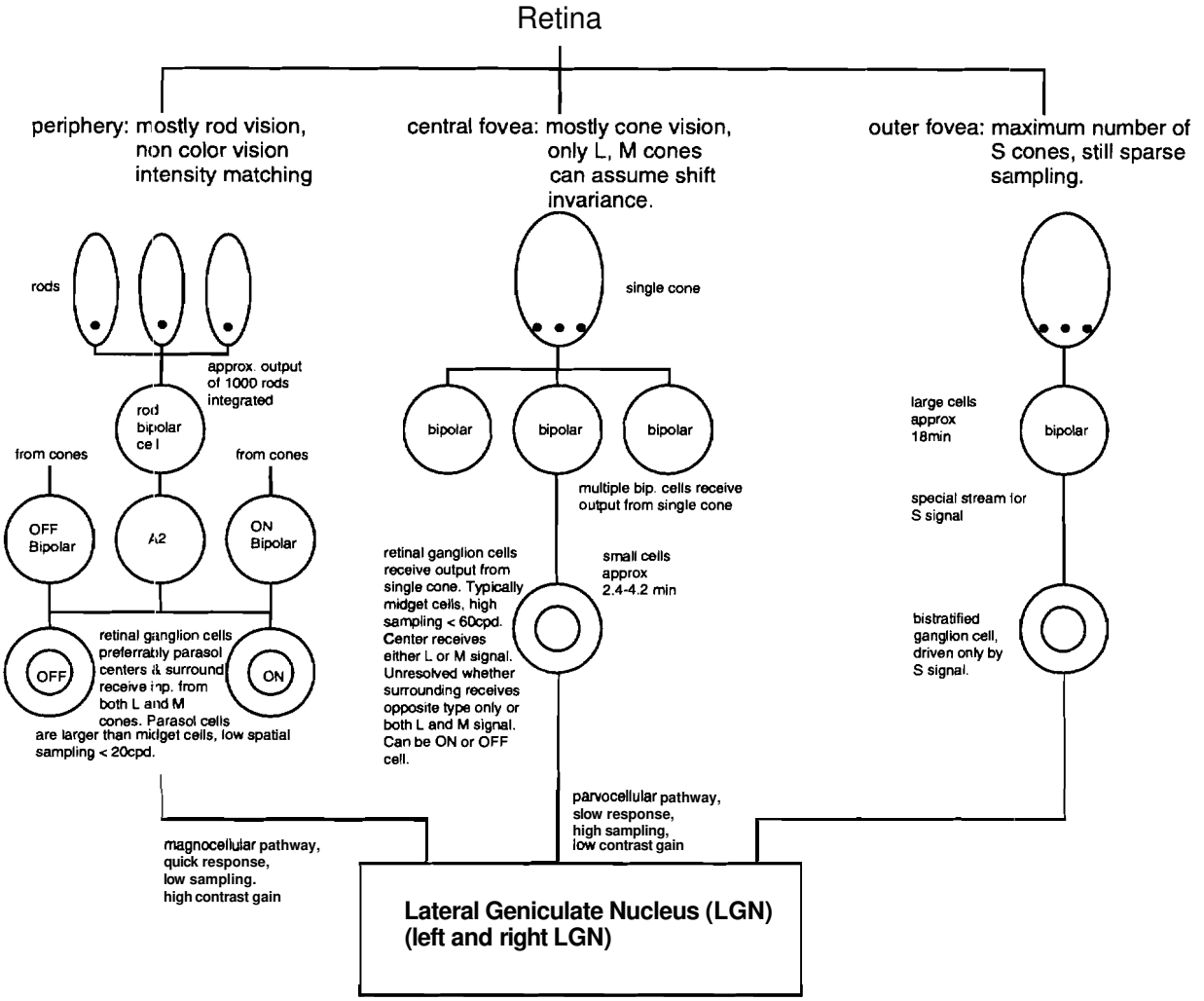

Figure 3: A functional diagram of the human retina. The figure shows the three basic streams which have been identified in the retina. The path in the center corresponds to the L- and M-cone vision in the central fovea which is most important for our considerations. The cones convert the incoming light into photocurrent which is received by multiple bipolar cells per cone. The bipolar cells send their signals to the retinal ganglion cells. These cells have a center surround organization as indicated by the concentric circles. Most of the retinal ganglion cells send their output to the LGN from where it is further distributed to higher areas of the visual cortex. The path on the right processes the signal of the S-cones and has a similar organization to the center path. The path on the left represents the monochrome rod vision.

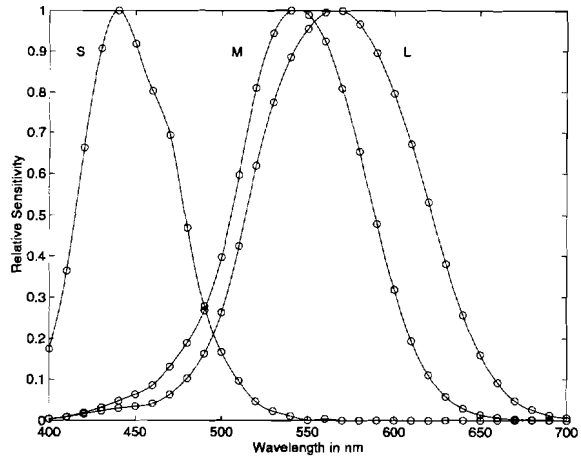

Figure 4: The relative sensitivities of the L, M and S cones. After Boynton, 1979

chromatic aberration, the optics of the eye blur short wavelengths so that a finer sampling of light in this frequency range would not improve vision.

Although the rod sampling is finer than the cone sampling, the monochrome rod-vision has lower spatial resolution than the chromatic cone-vision. This is due to the fact that the photocurrents of approximately 1000 rods are integrated by a single rod bipolar cell to obtain high light sensitivity. In contrast to this, the output of a single cone is sent to multiple bipolar cells which receive input from this cone only, making cone vision sensitive to high spatial frequencies.

The bipolar cells pass their signals on to retinal ganglion cells which are most important for our considerations. Ganglion cells are characterized by their receptive field which is defined as the retinal area in which light influences the cell response. The receptive fields of retinal ganglion cells have a 'center-surround' organization, which means they react differently to stimuli to the center as compared to stimuli to the periphery of their receptive field. When no stimulus is applied to such a cell, it emits a random sequence of electric pulses at a rate of approximately 50 spikes per second. When a stimulus is applied to the center, the cell reacts excitory by emitting more pulses per second. If a stimulus is applied to the surround, the reaction is inhibitory, reducing the number of spikes per second below the equilibrium level. A cell with this behavior is called 'on-center, off-surround' cell. The retina also contains ganglion cells, which have 'off-center, on-surround' organization, that is their center response is inhibitory and their surround response is excitory. Figure 5 shows the steady-state receptive field response for an on-center, off-surround ganglion cell. The plot indicates that the cell response is essentially the difference between the luminance of a center stimulus and the mean luminance over the receptive field. If a stimulus falling on the center of an ON-cell is lighter than the average illumination, the cell response is positive whereas if it is darker the response is negative. Since the difference operation is localized to the receptive field of a single cell, the computed signal is usually called 'local contrast'. Measurements in human visual systems research in fact confirm that the response of ganglion cells to contrast patterns is linear in terms of contrast ([15], page 132). From these

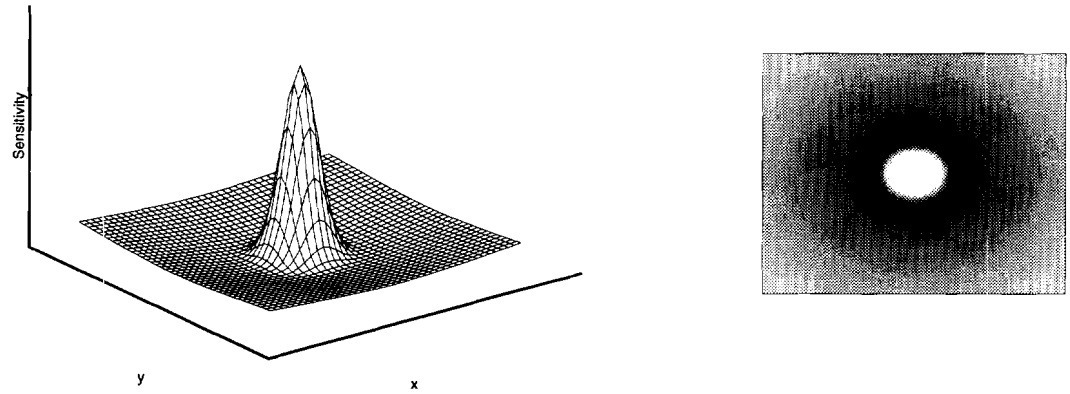

Figure 5: Qualitative receptive field of an on-center, off-surround retinal ganglion cell. The cell responds excitory to stimuli to its center and inhibitory to stimuli to its periphery.

observations, vision scientists conclude that contrast is the most important quantity encoded in the streams of our visual system. The physiological importance of contrast is consistent with psychological observations. The human eye is regularly confronted with changes in luminance over many magnitudes. We usually ignore these changes since most information we perceive is encoded in the spatial relationships of the reflectance of objects. For example, a scene under daylight illumination and its photograph viewed at artificial illumination appear very similar. although the difference in luminance can be several orders of magnitude.

In term:; of the spatial frequency response, ganglion cells have a band-pass characteristic. If the stimulus to the receptive field of a ganglion cell is a contrast pattern of very low spatial frequency, the luminance across the field of the cell will be constant, resulting in equilibrium output. If, on the other hand, the pattern has very high spatial frequency, the output will also be the equilibrium response since the dark and light areas within center and surround average out to mean illuminance. An important measurement which illustrates this band-pass behavior is the smallest perceivable luminance difference as a function of spatial frequency. This function is called the contrast-sensitivity function and plays a key role in determining whether image distortions are perceivable or not.

Most of the retinal ganglion cells send their signals to a brain area called the Lateral Geniculate Nucleus (LGN). There is one LGN on each side of the head, but each LGN receives output from both eyes. Each LGN sends output to approximately 20 different areas of the brain, whose functionality in terms of vision is not well known. Research has concentrated on one specific area called the Primary visual cortex (V1). The Primary visual cortex can be divided into 6 main layers by the characteristics of the neuron responses in each layer. For example, 80% of the neurons in layers 1-3 are binocularly driven while most of the neurons in layers 4-6 respond to one eye only. The cells in V1 are commonly classified in 'simple cells' with linear response and 'complex cells' with non-linear response. Important for our purposes is the observation of simple cells in V1 which respond to stimuli of specific

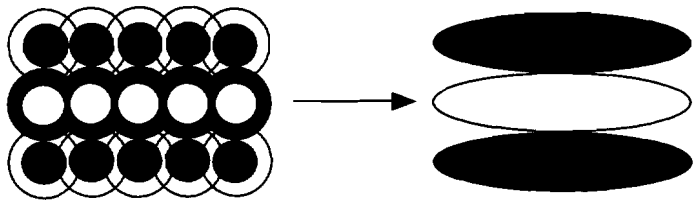

Figure 6: Orientation selective fields. The sum of the responses of the cell array on the left yields an oriented receptive field as shown on the right. The white areas correspond to locations of excitory response whereas the black areas represent locations of inhibitory response. Assuming that the black and the white areas have the same total area, this filter responds predominantly to horizontal stimuli.

orientation!; only. These cells are similar to the retinal ganglion cells in that they have receptive fields which are divided in excitory and inhibitory areas. However.,the cells in V1 are not radially symmetric but have oriented receptive fields as shown on the right of Fig. 6. These fields are assumed to be the result of a linear combination of the signals from several radially symmetric cells in the LGN as shown on left of the figure. Based on this observation, most image processing models of the human visual system have included banks of filters of different orientations.

The knowledge about the further processing of the visual signals in higher areas of the brain is very limited. Vision science has identified brain areas which are believed to be important for vision. An area called V4 has been shown to respond strongly to color-contrast stimuli. Another area called medial temporal (MT) responds predominantly to stimuli containing movement of objects. However, these models are still speculative and the underlying methods of research are controversial.

In summary, this section explained the order in which visual stimuli are processed by the early stages of the human visual system. The first stage of lens optics is followed by a nonlinear light to current conversion in the photoreceptors. Retinal ganglion cells then compute contrast signals which are passed on to the LGN, V1 and higher areas of the visual cortex. While the knowledge about most of the higher brain areas is very limited, layers in V1 have been shown to contain orientation selective neurons.

The reader should keep in mind, that even the knowledge about the early stages of the HVS is highly incomplete. This simplified introduction has only explained some of the known aspects which are commonly modeled by image processing models of the HVS. Vision science indicates that there are many more streams and cells in the retina which have not been investigated. Due to these limitations, extracted features for a model of the human visual system should not only be physiologically motivated, but also be consistent with behavioral measurements. Furthermore, the goal of any engineering model cannot be to mimic the human visual system, but to extract similar features which might relate well to human perception.

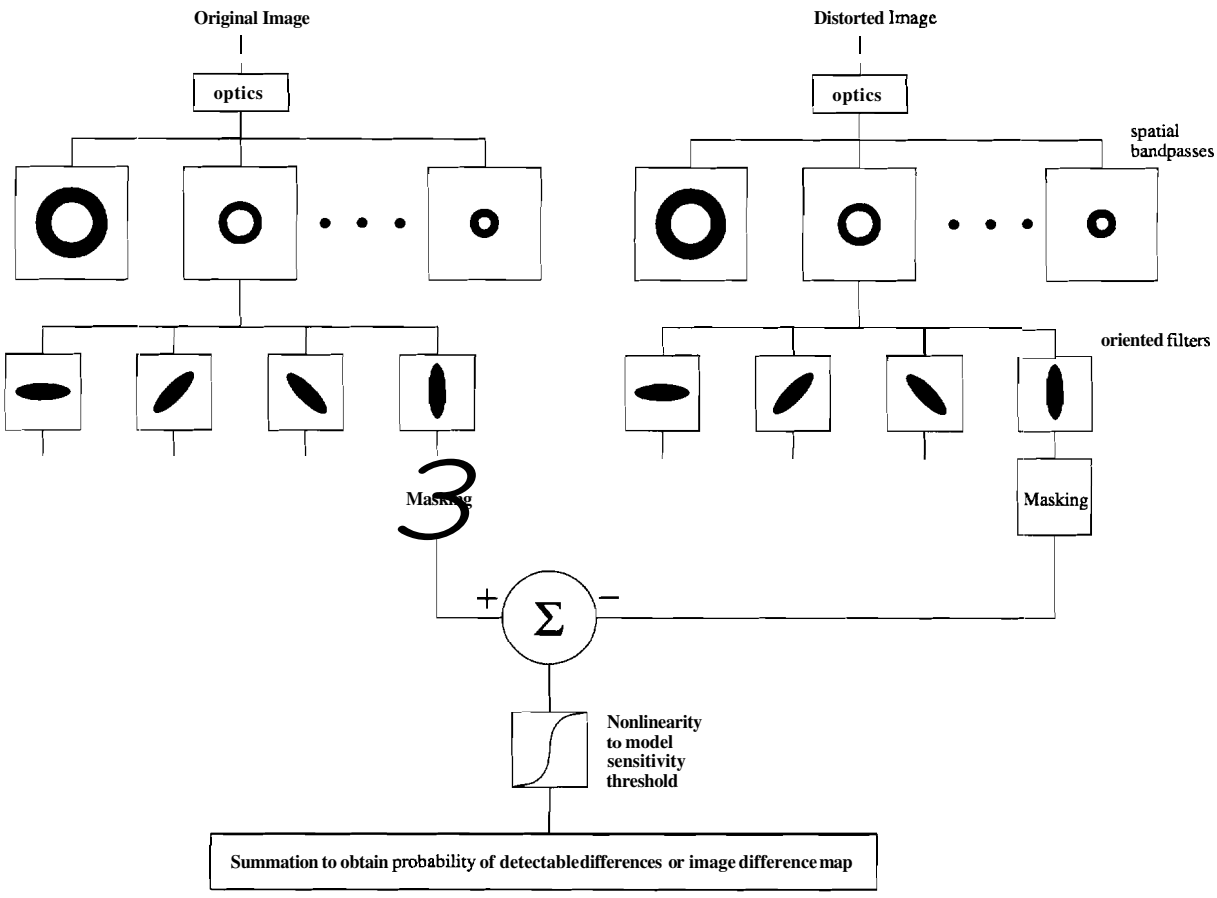

Figure 7: Achromatic HVS channel model used for image quality evaluation. Both input images are filtered by a bank of spatial bandpass filtered followed by filters of different orientations for each channel. After accounting for masking effects, the image representations are subtracted and passed through a non-linearity to model the detection threshold. The results are either displayed as image difference maps or summed up to yield a single quality measure.

3.2 Existing models

Recently, models of the early human visual system have been developed to design quality metrics for applications such as halftoning and perceptually lossless compression [16, 17, 18, 19, 20]. These channelized models exploit multiscale pyramid structures to better account for known visual phenomena and essentially measure the similarity between the original and distorted versions of an image.

Figure 7 shows the typical architecture of an achromatic HVS model for image quality evaluation. The first stage models the optical transfer function of the lens and in some models contains an amplitude nonlinearity to model the transfer function of the photoreceptors. It is followed by a bank of radially symmetric band-pass filters which model the function of the retinal ganglion cells. The band-pass filters compute a multiscale contrast representation which allows modeling the contrast sensitivity function by assigning different weights to each frequency band [20]. Furthermore, a multiscale representation is supported by vision research since the effect of pattern adaption cannot be explained by a single resolution theory [15].

In order to model orientation selective effects such as the variation of contrast sensitivity as a function of orientation as well as masking, each bandpass is followed by a bank of orientational selective filters. A common implementation is to combine the band-pass and oriented filters using a bank of second derivative of Gaussian filters with different sizes and orientations [21]. Other approaches use filters specified in the frequency domain [18], Gabor filters or a Gaussian pyramid model [16].

After performing the filtering operations separately for the original and the distorted image, the representations are subtracted pixel-wise in each channel. This subtraction is either preceded or followed by a stage which accounts for masking. Masking is the effect of reduced visibility of a contrast pattern if a strong background contrast stimulus is present. A thin line, for example, might be visible on a uniform background, however, if it is displayed close to a thick line with stronger contrast, it might become undetectable. In terms of the contrast sensitivity function, masking means a threshold elevation for the masked stimulus. The modeling of the masking mechanism varies considerably between the existing models. While some models extract the masker signal from the original image and perform the actual threshold elevation after the subtraction [20], others perform the entire masking operation before subtracting the original and the distorted image [16, 18].

In most models, the next processing stage is a sigmoid shaped non-linearity which represents the probability of detecting stimuli at threshold levels. This modeling is based on the observation, that the human detection performance varies even for a single subject presented with the same stimulus multiple times. It is therefore reasonable to model the detection threshold as a cumulative probability function instead of using a binary threshold to decide whether a difference is perceivable or not.
Finally, the detection probabilities are either displayed as image difference maps or summed up to obtain an overall image distortion measure.

Models like these have been shown to be quite successful in predicting the perceptibility of differences between images. The main innovation of these models is the introduction of a systematic evaluation method for lossy image processing techniques such as halftoning and image compression. A compression algorithm, for example, can assign less bits to high spatial! frequencies since the human observer is not very sensitive to high frequency distortions. Further compression can be gained in the vicinity of strong contrast stimuli, since details will be masked for the observer.

However, the models for quality assessment are not directly applicable to tlne image search problem because they are designed to measure threshold level differences. In contrast, the search problem requires a metric that describes differences well above threshold.

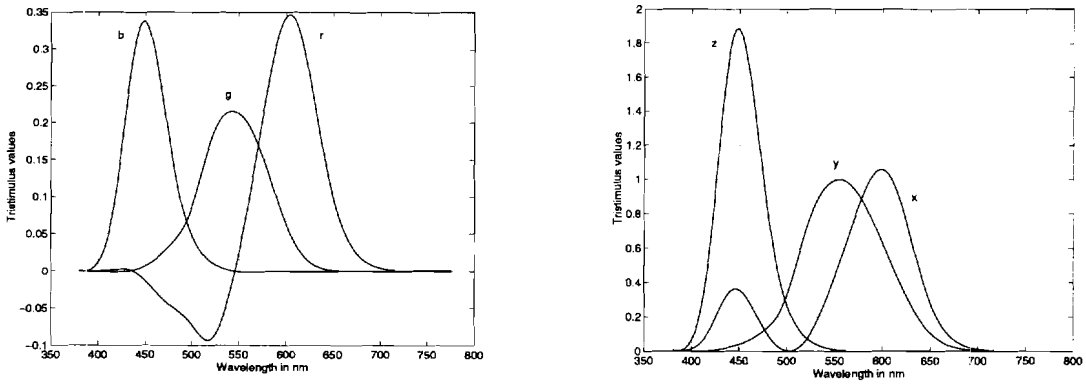

Figure 8: Color-matching functions of the RGB and XYZ color-spaces. After data from [24].

4 Image dis-similarity metric

In this work, we propose an approach to image similarity using features extracted from a multiscale-channelized model. The model is based on existing monochrome models of threshold behavior and the Lab uniform color-space. It includes channels which account for perceptual phenomena such as color, contrast, color difference and orientation selectivity. From these channels, we extract features and then form an aggregate measure of similarity using a weighted linear combination of the feature differences. The specific choice of features and weights is made to maximize the consistency with similarity ratings made by human subjects.

4.1 Choice of the color-space

The choice of the color-space for an image similarity metric is critical in order to obtain color differences which correspond well to human perception. In particular, the color-space must be uniform, i.e. the intensity difference between two colors must be consistent with the color difference estimated by a human observer. Unfortunately, the RGB color-space is not well suited for this task because the relationship between the RGB tristimulus values and perceived color intensity is highly nonlinear. The development of a sufficiently uniform color-space is complex since the human color perception varies with illumination and stimuli. Although complex vision models have been proposed [22, 23], a sufficiently uniform color-space and color difference formula have not yet been identified. Due to the urgent need for perceptually uniform color difference equations, in 1976, the Commission Internationale de l'Eclairage (CIE) recommended the use of two approximately uniform color-spaces called the 1976 CIE $L^*u^*v^*$ and the 1976 CIE $L^*a^*b^*$ color-spaces. Both color-spaces are based on the 1931 CIE XYZ color-space which was designed to match light of any wavelength composition with non-negative primary intensities. The linear transformation from CIE RGB to XYZ

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{pmatrix} 0.490 & 0.310 & 0.200 \\ 0.177 & 0.813 & 0.011 \\ 0.000 & 0.010 & 0.990 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

Figure 8 shows the color-matching functions of the RGB and the XYZ color-spaces. The three curves indicate the intensities of the color-space primaries which are necessary to match monochromatic light sources of the wavelengths drawn on the x-axis. In order to match light of short wavelengths, the R tristimulus value in the RGB space is negative whereas the XYZ intensities are strictly non-negative throughout the visible range. More importantly, the Y tristimulus value is centered in the middle of the spectrum and falls off slowly towards both ends. Consequently, Y corresponds to achromatic luminance whereas X encodes primarily the red† and Z the blue part of the spectrum. The CIE equations for transforming the XYZ color-space into Lab are

$$\begin{aligned} L^* &= 116\left(\frac{Y}{Y_w}\right)^{1/3} - 16 \\ a^* &= 500\left[\left(\frac{X}{X_w}\right)^{1/3} - \left(\frac{Y}{Y_w}^{1/3}\right)\right] \\ b^* &= 200\left[\left(\frac{Y}{Y_w}\right)^{1/3} - \left(\frac{Z}{Z_w}^{1/3}\right)\right] \end{aligned} \tag{2}$$

where X_w, Y_w and Z_w are the tristimulus values of the white point. Since a^* is computed as the nonlinear difference between X and Y, it encodes a red-green opponent signal. Similarly, b^* is obtained by subtracting Y and Z and therefore represents a yellow-blue opponent signal. Since L^* depends only on Y, it encodes achromatic luminance. The recommended color-difference equation for the Lab color-space is given by the Euclidean metric

$$\Delta E^*_{ab} = \left[(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2\right]^{1/2}. \tag{3}$$

A standard test for the uniformity of a color-space is to draw diagrams of the tristimulus values for color patches which are perceptually equally spaced with respect to hue, saturation and brightness. Figure 9 shows such a diagram for the Lab color-space using color patches of Munsell value 5 from the Munsell Book of Colors. The symmetry of the diagram indicates a high uniformity of the Lab color-space. However, since human color perception varies substantially with spatial frequency and illumination, this uniformity is only valid for stimuli at low spatial frequencies viewed under daylight illumination.

While the Lab color-space is widely used in engineering applications, most vision models use color-spaces which are directly based on the cone photopigment absorptions [22, 24]. Since furthermore opponent signals are evident in the neural pathways of the LGN, vision

†In order for light to appear red, it must contain spectral components at short wavelengths. Due to the the overlap of the L and M cone sensitivity functions, light of long wavelengths always invokes both the red and the yellow opponent stream. To perceive the light as being red, the yellow signal must be canceled by an opposing blue component.

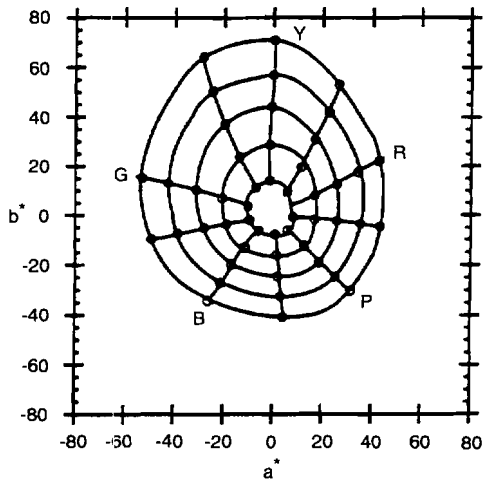

Figure 9: Uniformity of the Lab color-space at low spatial frequencies. The figure shows the a* and b^* values for constant hue and constant chroma. Ideally, the lines for constant hue should be straight and the contours for constant chroma should be concentric circles. Figure from [25].

research [26, 27] has used psychological experiments to identify separate pathways for luminance, red-green and yellow-blue signals in the human visual system. These studies have proposed new opponent-color spaces which are based on linear transformations of the cone absorptions. In particular, the spaces were designed to have separable contrast sensitivity functions for each channel. Consequently, these spaces are well suited for chromatic threshold detection tasks. However, little data about their perceptual uniformity has 'been published.

For the task of general image similarity assessment, we consider the uniformity of the color-space to be more important than its properties in terms of the contrast sensitivity function. In order to obtain general image similarity metrics, we need to measure image differences which we can assume to be considerably above threshold. For these differences to be perceptually relevant, the color-space must be as uniform as possible. The contrast-sensitivity function, however, is a threshold measure which only determines the perceptibility of differences.

While the Lab color-space was designed to be approximately uniform at low spatial frequencies, the color-spaces with a separable contrast-sensitivity function take into account the dependency of human color vision on spatial frequency. In general, this is a much more accurate description of human color perception. However, work on these color-spaces is still underway and little information about their uniformity at low spatial frequencies has been published. Since we consider the perceptual uniformity at low spatial frequencies to be most important, we decided to use the Lab color-space for our model. We will also see, that due to the consistency of the $x^{1/3}$ Lab conversion power with the characteristics of the photoreceptors, the selection of Lab yields a simple contrast calculation with desirable properties in our model.

4.2 HVS Model Architecture

The HVS channel model proposed in this work is derived from existing models for image quality assessment. However, the purpose of this model is quite different from that of the quality assessment models. Recall, that the purpose of the models in section 3.2 is to decide

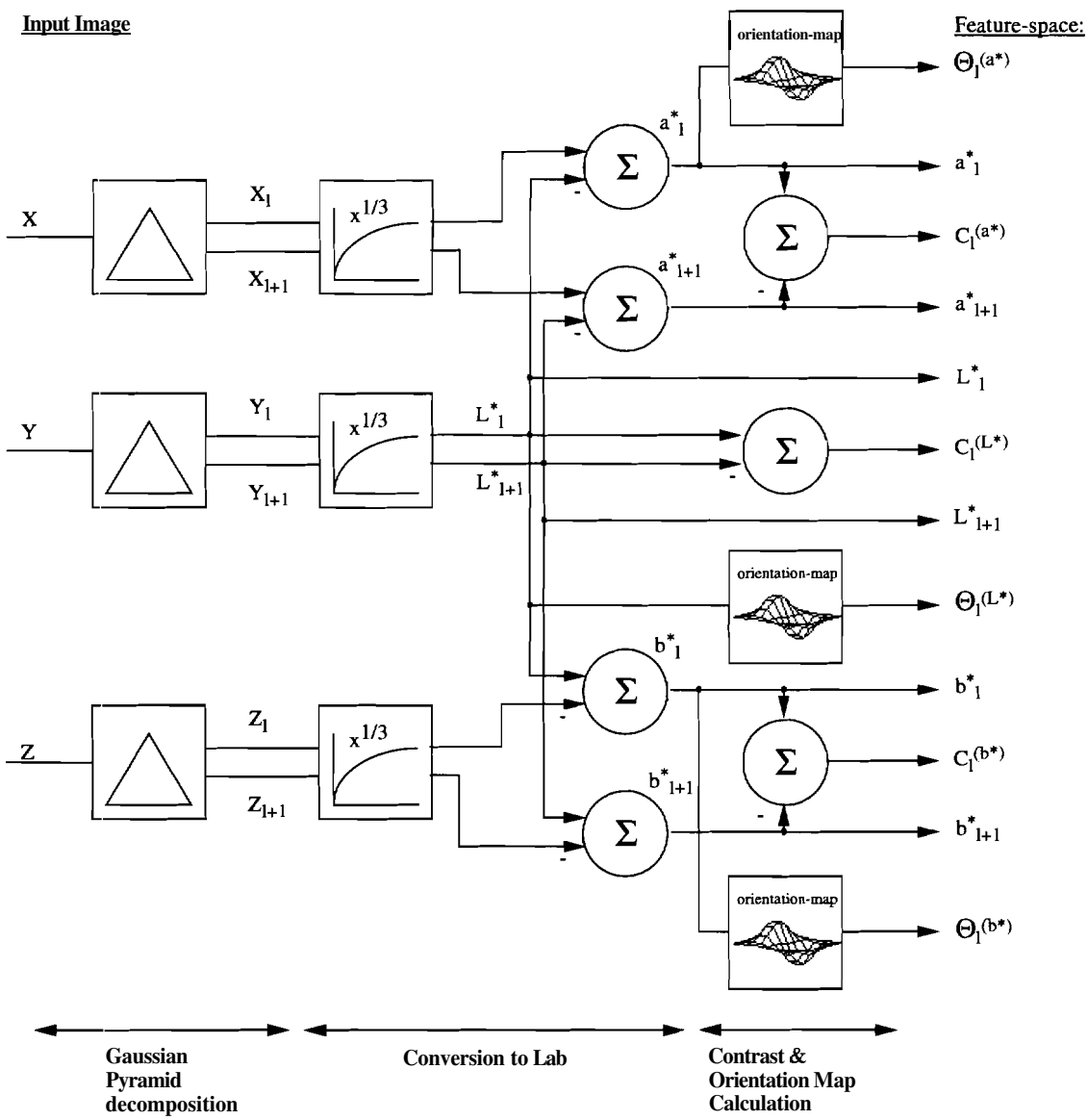

Figure 10: HVS Channel Model for Image Similarity Metric. For reasons of simplicity, the diagram shows only two pyramid levels and contrast calculations using adjacent pyramid levels.

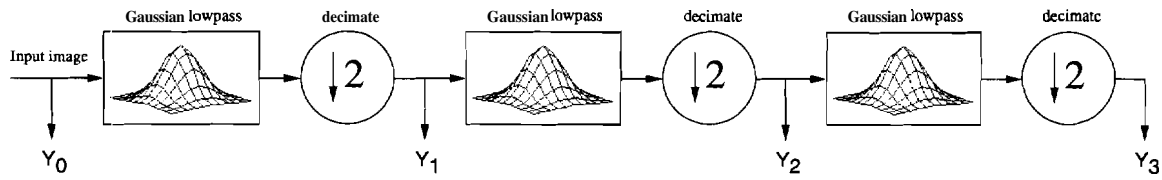

Figure 11: Pyramid decomposition of a four level Gaussian pyramid. The image is successively low-pass filtered and decimated.

whether differences between images are perceptible or not. The purpose of our model on the other hand is to assess continuous perceptual distances between images. While the quality models estimate the limits of the early stages of vision to predict 'what' is perceptible, our model seeks a representation which predicts 'how' we perceive differences. In. particular, our model concentrates on the perceptual uniformity of the image representation instead of on masking and contrast sensitivity.

The basic structure of the proposed model is illustrated in Fig. 10. We first de-gamma correct the RGB input image and convert it to XYZ color-coordinates using (I). We then apply a Gaussian pyramid decomposition to each of the three color channels [28]. The pyramid decomposition is computed by successively low-pass filtering the original image with a Gaussian kernel and decimating by two as shown in Fig.11. The low-pass kernel can be expressed as a sampled 2D-Gaussian kernel with standard deviation a = a, = a,. If we define the sample spacing $a = 1/\sigma$, the kernel can be expressed as

$$h(m,n) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}((\frac{m}{\sigma})^2+(\frac{n}{\sigma})^2)} = \frac{\alpha^2}{2\pi} e^{-\frac{1}{2}((\alpha m)^2+(\alpha n)^2)}. \tag{4}$$

The indices m and n are the discrete-space coordinates in x and y direction. Let l denote the pyramid level ranging from $l = 0$ representing the input image and l = L – 1 for the last level in the pyramid. Furthermore, let M_l and N_l denote the size of image at pyramid level l. The computation of the luminance image Y_l at level l is then given by

$$Y_l'(m,n) = \sum_{i=1}^{M_{l-1}} \sum_{j=1}^{N_{l-1}} h(m-i, n-j) Y_{l-1}(i,j) \tag{5}$$

$$Y_l(m,n) = Y_l'(2m, 2n) \tag{6}$$

where the first equation is the low-pass filtering and the second is the decimation by a factor of 2. Due to the decimation operation, the image size at level l is half the size of that at $l - 1$. If the size of the original image is given by M_0 x N_0, we can express the level sizes as

$$M_l = \frac{M_{l-1}}{2} = \frac{M_0}{2^l}$$

$$N_l = \frac{N_{l-1}}{2} = \frac{N_0}{2^l}. \tag{7}$$

The pyramid levels X_l and Z_l are computed analogously using the X and Z coordinates of the input image. The result is a multiscale representation of the image where each pyramid level l

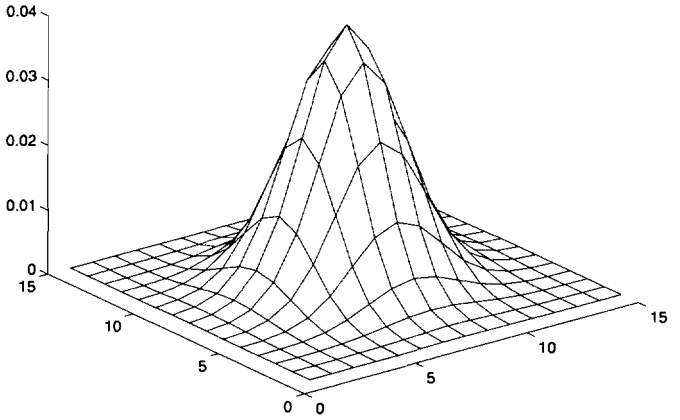

Figure 12: Gaussian filter kernel for pyramid decomposition. This kernel with sample spacing $\mathbf{a} = 0.5$ and of size 15x15 is used to successively low-pass filter the image.

has the resolution of the original image divided by 2^l. Since the convolution of two Gaussians is also Gaussian, the decomposition is equivalent to directly filtering the input image with different Gaussian kernels for each level and decimating the result. However, the pyramid decomposition is much more efficient because most filtering operations are performed on image sizes much smaller than the size of the original image. The implementation of the pyramid decomposition requires choosing the size and the sample spacing of the filter kernel. If the sample spacing is too small, the filter has a very low cutoff frequency, resulting in a large resolution difference between adjacent pyramid levels. If, on the other hand, the sample spacing is too large, the filter has a very high cutoff frequency, resulting in aliasing produced by the decimation operation. Using images containing radially symmetric sine-waves, we experimentally determined a reasonable sample spacing of $\alpha = 0.5$ and a corresponding kernel size of 15x15. The implemented Gaussian filter kernel is shown in Fig. 12. Since we experiment with image sizes of approximately 185x280 pixels, we chose the number of pyramid levels to be $L = 5$. Consequently, the size of the lowest level as given by (7) is M_4 x N_4 = 11 x 17. Since the kernel size of the Gaussian filter is 15 x 15, further levels would be dominated by border effects and are therefore not meaningful to compute.

The pyramid decomposition is followed by a conversion of each pyramid level to Lab color-space. Since the white point of the screen given by $[R_w G_w B_w]^T = [1\,1\,1]^T$ transforms into $[X_w Y_w Z_w]^T = [1\,1\,1]^T$, equation (2) becomes

$$L_l^*(m,n) = 116 Y_l^{1/3}(m,n) - 16 \tag{8}$$

$$a_l^*(\mathrm{m,n}) = 500(X_l^{1/3}(m,\mathrm{n}) - Y_l^{1/3}(m,\mathrm{n})) \tag{9}$$

$$b_l^*(m,n) = 200(Y_l^{1/3}(m,n) - Z_l^{1/3}(m,n)) \tag{10}$$

Figures 13 and 14 show the resulting Lab pyramids for two example images. Note that the artificially introduced red-green and blue-yellow colors in the a_l^* and b_l^* images represent

Figure **13:** Gaussian pyramid decomposition. The figure shows the Lab pyramid decomposition of the color image in the upper left corner. From top to bottom, the rows contain the pyramids for the L*, a^* and b^* channels.

Figure 14: Gaussian pyramid decomposition. The figure shows a second example of an Lab pyramid decomposition. From top to bottom, the rows contain the pyramids for the L*, a* and b* channels.

the signs of these opponent signals. In the following, we will refer to the L_l^* as *luminance channels* and to a_l^* and b_l^* as *color channels.* We will see that we can use these channels not only in the traditional way of comparing luminance and color between images, but also to compute color and color-contrast representations in the HVS model.

4.2.1 Contrast Representation

A common definition of the contrast of a luminance stimulus Y relative to the background luminance Y_B is given by Weber's contrast C_W

$$C_W = \frac{Y - Y_B}{Y_B} \tag{11}$$

For small $C_W \approx 0$, this can be approximated by the logarithmic difference

$$C_W \approx \ln Y - \ln Y_B. \tag{12}$$

That this is in fact an approximation, can be seen by expanding the Taylor series for the logarithm

$$\begin{aligned} \ln Y - \ln Y_B &= \ln \frac{Y}{Y_B} \\ &= \left(\frac{Y}{Y_B} - 1\right) - \frac{1}{2}\left(\frac{Y}{Y_B} - 1\right)^2 + \ldots \\ &\approx \left(\frac{Y}{Y_B} - 1\right) \\ &= \frac{Y - Y_B}{Y_B} \\ &= C_W. \end{aligned} \tag{13}$$

The scientist Weber defined C_W based on the observation, that the human sensitivity to this contrast is approximately constant with respect to background luminance. He formulated this relationship as Weber's law, which states that if C_{WS} denotes the minimum contrast necessary for a stimulus to be just noticeable, then the contrast sensitivity defined as $1/C_{WS}$ is not a function of background luminance, i.e.

$$\frac{1}{C_{WS}} = \text{const.} \tag{14}$$

However, measurements indicate that the contrast sensitivity to C_W is not completely invariant to background luminance, but increases as shown in Fig. 15. Therefore, the contrast definition (11) does not accurately describe the background luminance dependence of human contrast sensitivity. Furthermore, Weber's contrast and its logarithmic approximation have the disadvantage, that if the background luminance approaches zero, the contrast goes to infinity, which is inconsistent with human perception.

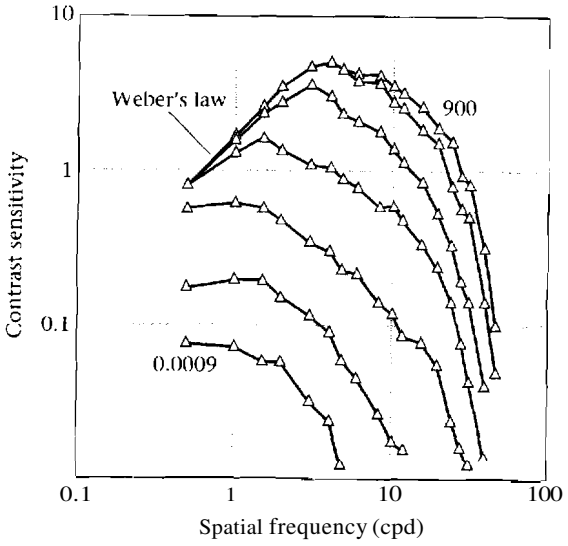

Figure 15: Deviations from Weber's law. The curves show human contrast sensitivity functions for different background luminances. At low background intensities, the contrast sensitivity increases with luminance. At high intensity levels, the functions converge and Weber's law becomes a good approximation. Figure from [15], after data from van Nes and Bouman.

Motivated by physiological aspects of the human visual system, we define contrast differently. The transfer function of the photoreceptors followed by the subtraction performed by the ganglion cells, suggests that contrast be defined as

$$C = Y^{1/3} - Y_B^{1/3}. \tag{15}$$

This definition is not only physiologically motivated , but also avoids the disadvantages of Weber's law contrast. In particular, if Y_B approaches zero, C becomes $Y^{1/3}$ which is more consistent with human perception. Furthermore, if we use the Taylor expansion

$$x^{1/3} = 1 + \frac{1}{3}(x-1) - \frac{1}{9}(x-1)^2 + \ldots \approx 1 + \frac{1}{3}(x-1) \quad \text{for } x \approx 1 \tag{16}$$

we can approximate (15) for $C \approx 0$ as

$$\begin{aligned} C &= Y^{1/3} - Y_B^{1/3} \\ &= Y_B^{1/3}\left[(Y/Y_B)^{1/3} - 1\right] \\ &\approx Y_B^{1/3}\left[1 + \frac{1}{3}(Y/Y_B - 1) - 1\right] \\ &= \frac{1}{3}Y_B^{1/3}(Y/Y_B - 1) \\ &= \frac{1}{3}Y_B^{1/3}\left[\frac{Y - Y_B}{Y_B}\right] \\ &= \frac{1}{3}Y_B^{1/3}C_W. \end{aligned} \tag{17}$$

For small luminance differences, the power law contrast therefore becomes Weber's contrast weighted by $Y_B^{1/3}$ and a constant. If we now set the sensitivity to this contrast to be constant

with respect to Y_B, we obtain

$$\begin{aligned} \frac{1}{C} &= \text{const} \\ \frac{1}{C_W Y_B^{1/3}} &= \text{const} \\ \frac{1}{C_W} &= Y_B^{1/3} \times \text{const} \end{aligned} \tag{18}$$

which qualitatively accounts for the background luminance dependence of $1/C_W$.

Another reason for choosing the contrast definition of C is its consistency with the XYZ to Lab color-space conversion. Since the conversion power of 1/3 in the contrast calculation is the same as the conversion power in the color-space conversion, we can compute C as

$$C = Y^{1/3} - Y_B^{1/3} = \frac{1}{116}(L^* - L_B^*) \tag{19}$$

where L* and L_B^* are the values of L* corresponding to Y and Y_B as obtained by (2). For the similarity metric, we will only be interested in relative contrast differences between images. Therefore, we can ignore the multiplicative constant of $1/116$ and define a scaled power law contrast $C^{(Y)}$ as

$$C^{(Y)} = 116\,C = \text{L}^* - L_B^*. \tag{20}$$

This implies a very simple contrast calculation for our model. Due to the spatial averaging of the low-pass filters in the pyramid decomposition, local background luminances for a stimulus in Y_l are given by the lower pyramid levels Y_k with $k > l$. We can therefore calculate contrast representations at pyramid level l as the difference between the luminance channel at level l and any luminance channel at level $l + \text{i}$ as long as $1 \leq \text{i} < L - l$. We denote these *contrast channels* as $C_{l,i}^{(Y)}(m, \text{n})$:

$$\begin{aligned} C_{l,i}^{(Y)}(m,n) &= 116(Y_l^{1/3}(m,n) - Y_{l+i}^{1/3}(m/2^i, n/2^i)) \\ &= L_l^*(\text{m,n}) - L_{l+i}^*(m/2^i, n/2^i) \end{aligned} \tag{21}$$

The indices $m/2^i$ and $n/2^i$ are a consequence of the i decimations between L_l^* and L_{l+i}^*. Since these indices are in general non-integers, we use bilinear interpolation of $L_{l+i}^*(m, \text{n})$ by a factor of 2^i before performing the subtraction.

The contrast computed by the subtraction of different levels in L_l^* is similar to the Difference of Gaussians model which was developed by vision scientists to model the receptive field of retinal ganglion cells. The model assumes that the cell center and surround responses are separable and have Gaussian shape. As shown in Fig. 16, subtracting the large variance surround response from the small variance center response yields a transfer function similar to the steady-state response of a ganglion cell.

Although it is common to compute contrast involving a difference of low-,passlevels, the proposed contrast calculation is quite different from that in existing HVS channel models. Most models perform a linear subtraction of adjacent low-pass channels in the same color-space in which the multiscale representation is obtained [20, 19]. This linear luminance

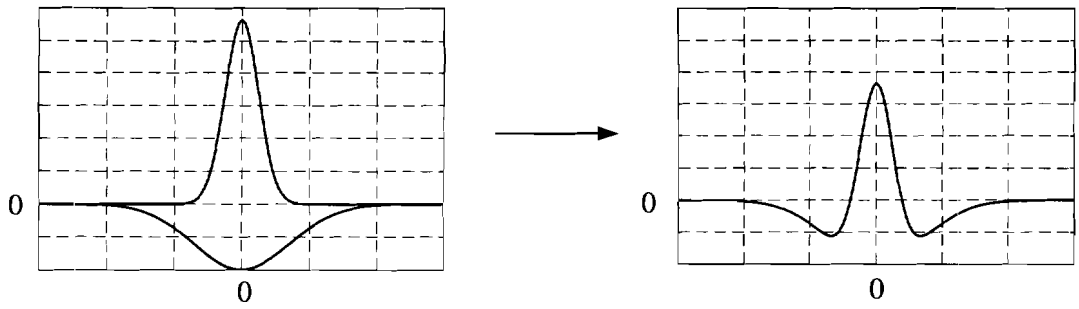

Figure **16:** The Difference of Gaussians concept. The subtraction of two Gaussian density functions with different standard deviations can be used to model the receptive field of a retinal ganglion cell

difference is then divided by a low-pass channel two levels lower in resolution to yield an approximation to Weber's contrast introduced by Peli [29]. A different approach proposed by Zetzsche [16] is to compute a Ratio of Gaussians by dividing adjacent pyramid levels which directly results in a simpler Weber-like contrast approximation. To the contrary, our model performs the multiscale and the contrast computations in different spaces which are separated by a nonlinearity. As explained above, we believe that this structure corresponds well to the human visual system and has several desirable properties.

An important question is, how spatially 'local' the surround response should be, or, in terms of our model, how large the difference i between the subtracted luminance levels in **(21)** should be. While most of the HVS models for image quality assessment compute contrast only between adjacent low-pass levels, we decided to compute all possible $C_{l,i}^{(Y)}(m,n)$. The importance of the different level differences i can then be determined in the feature selection and optimization process which forms the final similarity metric.

So far, we have examined the contrast calculation in terms of its properties for small luminance differences only. A different perspective is given by the fact that we compute the power law contrast in Y as a luminance difference in L*. Therefore, we preserve the uniformity of the Lab color difference equation **(3).** Consequently, contrast comparisons using the Euclidean metric should relate well to human perception.

In summary, we defined a power law contrast calculation in Y which is motivated by the characteristics of the visual processing in the early HVS. This contrast has desirable properties for low background luminances and can qualitatively explain deviations of human vision from Weber's law. Due to the structure of our model, this contrast can be computed as luminance differences in L* which preserves the perceptual uniformity of the Lab color difference equation for contrast comparisons.

4.2.2 Color contrast

In analogy to the luminance contrast, we introduce color contrast channels. Although the notion of color contrast is not well established, we believe that a comparison of color transi-

tions in images could be an important aspect of similarity perception. If, for example, two images differ only by a uniform color shift, they may still appear very similar since the color transitions from one object to another within the images are identical. Furthermore, it is fairly easy to show, that the appearance of colors can depend on the background on which they are shown. For example, black lines on a white background appear black, however, if we introduce a green background, the same lines will appear red. In fact, vision research has evidence that human color perception is based on the relative relationships between the L, M and S-cone absorptions instead of their absolute absorption rates [15]. Given the opponent organization of the visual system, a relative color comparison would be most probable to be performed in each of the opponent streams.

In order to compute contrast in the opponent coordinates a^* and b*, we can define the *color-contrast* channels $C_{l,i}^{(a^*)}(m,\mathrm{n})$ and $C_{l,i}^{(b^*)}(m,n)$ analogously to (21). This yields

$$C_{l,i}^{(a^*)}(m,n) = a_l^*(m,n) - a_{l+i}^*(m/2^i, n/2^i) \tag{22}$$

$$C_{l,i}^{(b^*)}(m,n) = b_l^*(m,n) - b_{l+i}^*(m/2^i, n/2^i). \tag{23}$$

By subtracting pyramid levels in each opponent coordinate, we compute the opponent color difference between the center signal at location (m,n) and the local average around this location. As in the luminance contrast computation, these differences are part of the Lab color difference equation and consequently have the uniformity of the Lab space.

However, the interpretation of the color-contrast signals in terms of the XYZ space is different from that of the luminance contrast. If we denote the multiplicative factor of 500 in (2) as c, a^* is given by

$$a^* = c(X^{1/3} - Y^{1/3}) \tag{24}$$

which can be interpreted **as** a power law contrast between X and Y. If now a^* denotes the tristimulus value of a foreground stimulus and a_B^* the tristimulus value of the background, we compute the color-contrast $C^{(a^*)}$ as

$$\begin{aligned} C^{(a^*)} &= c(X^{1/3} - Y^{1/3}) - c(X_B^{1/3} - Y_B^{1/3}) \\ &= c[(X^{1/3} - X_B^{1/3}) - (Y^{1/3} - Y_B^{1/3})]. \end{aligned} \tag{25}$$

If we approximate this for $\mathrm{X} \approx X_B$ and $\mathrm{Y} \approx Y_B$ analogous to (17), we obtain

$$C^{(a^*)} \approx \frac{c}{3}\left[X_B^{1/3}\frac{X - X_B}{X_B} - Y_B^{1/3}\frac{Y - Y_B}{Y_B}\right]. \tag{26}$$

Clearly, $C^{(a^*)}$ is the difference of power law contrasts in X and Y. Note however, that we must not interpret this as a difference between 'red' and luminance, but bet-ween 'red' and 'green' contrast. This is due to the fact, that in differences of X and Y the overlap in the color matching functions of X and Y in Fig. 8 cancel. Analogously to (26), the color-contrast $C^{(b^*)}$ can be approximated for $\mathrm{Y} \approx Y_B$ and $Z \approx Z_B$ as

$$C^{(b^*)} \approx c'\left[Y_B^{1/3}\frac{Y - Y_B}{Y_B} - Z_B^{1/3}\frac{Z - Z_B}{Z_B}\right] \tag{27}$$

These equations are meaningful, if we assume that the HVS computes the contrast of each opponent color separately before the subtraction. Such an assumption seems reasonable if we consider that the contrast calculation is performed by the retinal ganglion cells whereas the first opponent signals have been shown to exist in the LGN.

The implementation of the contrast channel computation (23) in our model is identical to the computation for luminance contrast. This means, that we compute color-contrast channels for all possible level differences i using bilinear interpolation. Whether the color contrast channels are important for the model, will be determined in the feature selection process.

4.2.3 Quantization of contrast features

Implementing the contrast and color-contrast channels, we found that an amplitude quantization considerably improves the performance of features derived from these channels. More specifically, the quantization of the contrast and color-contrast channels results in an increase of the number of contrast and color-contrast features which are selected by the feature selection process to be described in section 4.4. The best performance was obtained using only the three quantization levels (-1, 0, 1). In particular, if $T_{C_{l,i}^{(Y)}}$ is the quantization threshold for channel $C_{l,i}^{(Y)}(m,\mathrm{n})$, the quantized contrast channel $\tilde{C}_{l,i}^{(Y)}(\mathrm{m},\mathrm{n})$ is obtained by

$$\tilde{C}_{l,i}^{(Y)}(m,n) = \begin{cases} -1 & : \; C_{l,i}^{(Y)}(m,n) \leq -T_{C_{l,i}^{(Y)}} \\ 1 & : \; C_{l,i}^{(Y)}(\mathrm{m},\mathrm{n}) \geq T_{C_{l,i}^{(Y)}} \\ 0 & : \; \text{otherwise} \end{cases} . \tag{28}$$

In order to obtain the thresholds $\mathrm{T}_{C_{l,i}^{(Y)}}$, we calculated the channel variances $\sigma^2_{C_{l,i}^{(Y)}}$ over a set of 200 images q as

$$\sigma^2_{C_{l,i}^{(Y)}} = \frac{1}{200 M_l N_l} \sum_q \sum_{m=1}^{M_l} \sum_{n=1}^{N_l} \left(C_{l,i}^{(Y)}(m,n) - \mu_{l,i}^{(Y)} \right)^2 . \tag{29}$$

The channel thresholds are then calculated from $\mathrm{a}_{C_{l,i}^{(Y)}}$ as

$$T_{C_{l,i}^{(Y)}} = B \sigma^2_{C_{l,i}^{(Y)}} . \tag{30}$$

The constant B was experimentally determined as $\mathrm{B} = 0.3$. Using these values, most of the background noise of the unquantized channels is eliminated whereas important foreground contours are preserved. The calculation of the quantized color-contrast channels $\tilde{C}_{l,i}^{(a^*)}(m,\mathrm{n})$ and $\tilde{C}_{l,i}^{(b^*)}(m,\mathrm{n})$ is analogous to (28). The corresponding thresholds $\mathrm{T}_{C_{l,i}^{(a^*)}}$ and $\mathrm{T}_{C_{l,i}^{(b^*)}}$ are determined as in (30), using the same constant B.

The main interpretation of the increased performance obtained by quantizing the color and color-contrast channels is the elimination of background noise. In most images, the contrast values corresponding to background texture are small, however, they usually make out most of the area of the image. For similarity assessment, functions corresponding to the

local contrast differences between images have to be integrated over the image area. Due to the large background area, the integration of the small contrast values yields a large overall error which does not correspond to human perception. The contrast quantization minimizes this effect 'by setting the background contrast to zero.

The elimination of background noise explains that quantizing small contrast values to zero improves performance. However, this does not account for our observation that the best performance is obtained by quantizing values above noise level to a single value. An explanation in terms of our model might be, that the coarse quantization limits errors introduced by comparisons of foreground contours which are not lined up in the two images. A different interpretation in terms of human vision could be, that much of human similarity perception is based on shape recognition. The shape recognition, however, is not a continuous function of contrast. Instead, it relies on the binary decision whether contrast contours are the boundaries of objects or not. In this sense, we can interpret the three level quantization as a decision procedure about the importance of contrast contours. Note that such a decision is quite similar to the contrast sensitivity function. While the contrast sensitivity function describes the physiological perceptibility of contrast stimuli, this decision describes what we could call the psychological perceptibility. For future work, this motivates replacing the binary quantization threshold by a sigmoid-shaped nonlinearity as used in existing HSV models to compute detection probabilities. Note, that although the contrast is coarsely quantized, the perceptual uniformity of the underlying space is important in order to determine meaningful thresholds. Moreover, if a sigmoid-shaped nonlinearity will be used in the future, the uniformity becomes very important in order for the resulting detection probabilities to be linear.

Figures 17 and 18 show the final contrast and color-contrast representations of our two example images. The pyramids on the right contain the quantized contrast channels corresponding to the luminance and color channels on the left. Clearly, the contrast computations extract luminance transitions in the L* channel and color-transitions in the a* and b* channels. The bandpass behavior of the contrast levels is particularly apparent in Fig. 18, where for example the girl's hair in the highest level is shown by its boundary whereas it becomes a massive object in subsequent levels.

4.2.4 Orientation selective channels

In parallel to computing the contrast representation, we derive orientation selective channels from the pyramids in Lab. The representation of orientation selectivity in this model differs considerably from that in existing models. The main purpose of the filter banks in existing models is to account for masking. Since the masking effect decreases with angular difference between masker and masked signal, it is important to separate the stimuli into groups of similar orientation. At the same time, however, the shape and location of the stimuli must be preserved in order to identify masker and masked stimulus. The filter bank representation achieves this by computing separate channels for the angular intervals. Each of these channels contains an image representation whose stimuli amplitudes and locations are consistent with the input image. For the image search problem, however, the goal of orientation selective channels is to extract the dominant orientation at each location in the image. In order

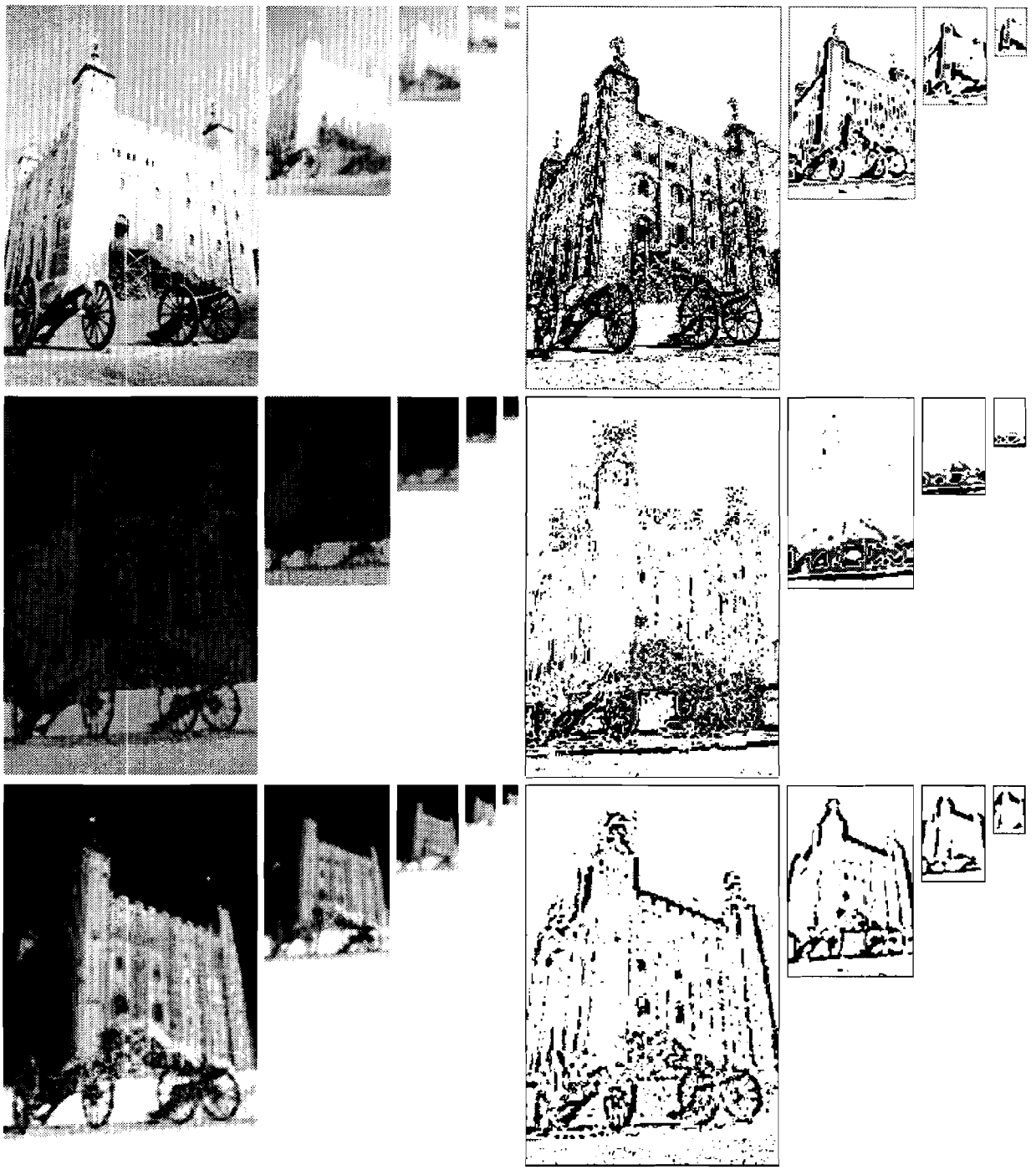

Figure 17: Contrast representation of example image **1.** The pyramids on the right show the quantized contrast and color-contrast pyramid representations derived from the L*, *a** and b* pyramids on. the left.

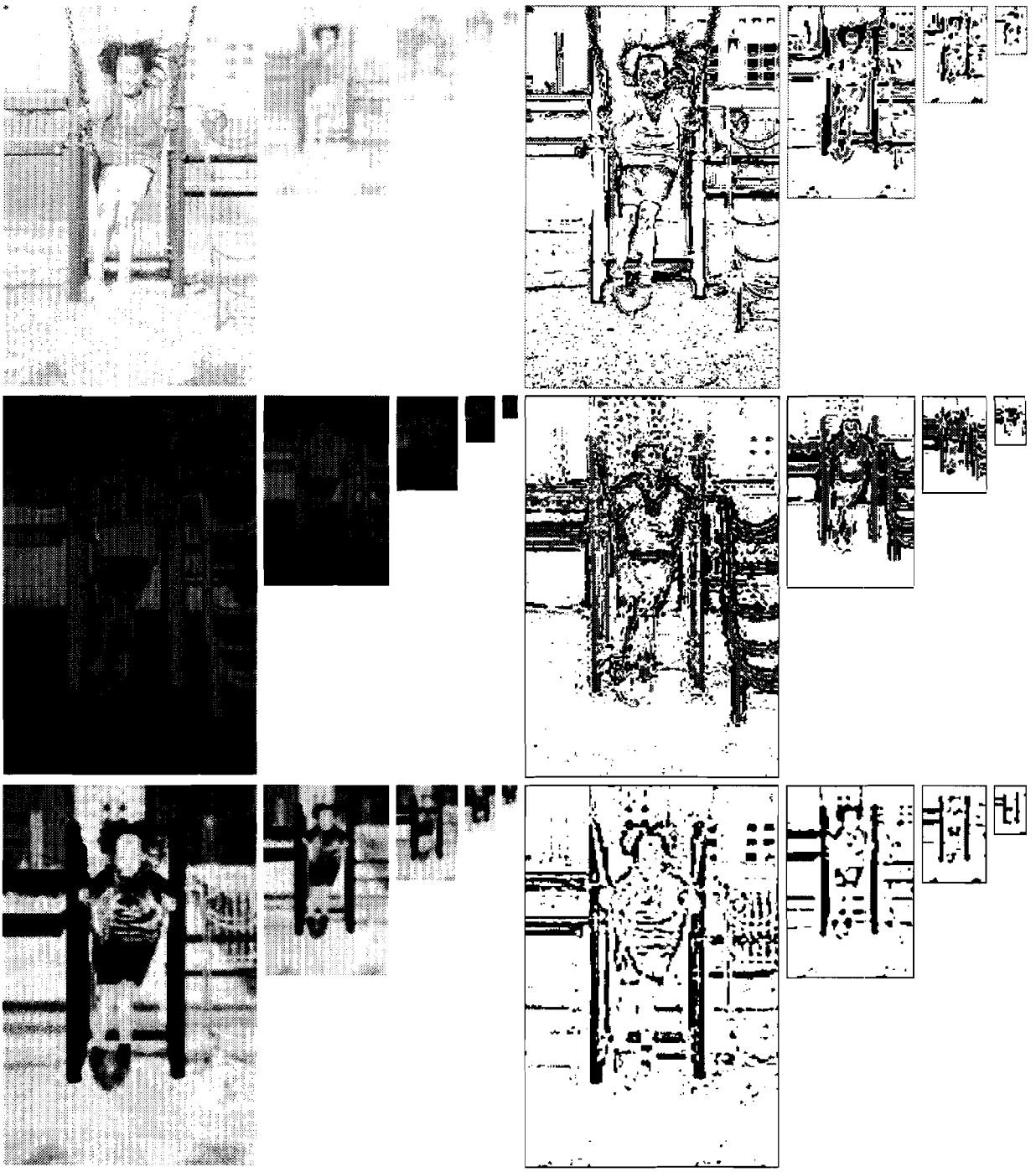

Figure 18: Contrast representation of example image **2.** The pyramids on the right show the quantized contrast and color-contrast pyramid representations derived from the L*, *a** and *b** pyramids on the left.

to effectively compare these orientations between images, it is desirable to represent the complete information about the entire angular range in one channel. This suggests the use of angular rnaps to model orientation selective perception. In our model, these maps consist of edge-angle and edge-amplitude values at each image location.

A popular method for computing orientation maps is to use a quadrature filter pair as proposed in [30, 31]. The idea is to filter the image with an oriented even symmetric filter and its odd Hilbert transform to obtain signals corresponding to the oriented real and imaginary part of the Fourier transform. Let both filters be oriented in direction ϑ_n and let $G^{\vartheta_n}(m, \mathrm{n})$ and $H^{\vartheta_n}(\mathrm{m.,n})$ denote the output of the even symmetric filter and its Hilbert transform respectively. The square sum of these outputs then results in the oriented energy $E^{\vartheta_n}(\mathrm{m,n})$.

$$E^{\vartheta_n}(\mathrm{m,n}) = [G^{\vartheta_n}(\mathrm{m},n)]^2 + [H^{\vartheta_n}(m,n)]^2 \tag{31}$$

The oriented energy is computed for a sufficient number N of equally spaced. orientations in order to satisfy the sampling theorem in polar coordinates. Consequently, the filter output for arbitrary orientations 6 can then be obtained as a linear combination of the N filter outputs. This linear combination can be expressed as a Fourier series of the form

$$E^{\vartheta}(m, \mathrm{n}) = C_1(\mathrm{m,n}) + C_2(m,n)\cos(2\vartheta) + C_3(m,n)\sin(2\vartheta) + \ldots \tag{32}$$

The dominant orientation angle ϑ_d and its orientation amplitude s_d can be approximated by

$$\vartheta_d(m,n) = \frac{1}{2}\arg(C_2(m,n), C_3(m,n)) \tag{33}$$

$$s_d(m,\mathrm{n}) = \sqrt{C_2^2(\mathrm{m,n}) + C_3^2(m,n)}. \tag{34}$$

We implemented the oriented filters for $E^{\vartheta_n}(\mathrm{m,n})$ using a second derivative-of-Gaussian filter kernel and its approximated Hilbert transform as given in [31]. Comparing images by their angular orientation maps obtained from (34) in the luminance ancl color channels resulted in significantly improved similarity metric performance over directly comparing the N energy outputs in separate channels. However, a disadvantage of the oriented energy computation is its inability to preserve contour polarity information. In particular, the orientation maps for images with the same contours but inverted contrast are identical. The map comparison of such images produces the same error as the comparison of images with identical contrast polarities.

In order to investigate the importance of edge polarity in orientation maps, we compared the quadrature filter method to an approach which preserves the polarity information. This method computes orientation as the argument of the horizontal and the vertical derivative of the input image. The derivative in each of the two directions is obtained by convolving the image with a first derivative of a two-dimensional Gaussian. The filter kernels $h_x(m, \mathrm{n})$ and $h_y(m,n)$ with sample spacing α are given by

$$\begin{aligned} h_x(m,n) &= \alpha m e^{-((\alpha m)^2+(\alpha n)^2)} \\ h_y(m,n) &= \alpha n e^{-((\alpha m)^2+(\alpha n)^2)}. \end{aligned} \tag{35}$$

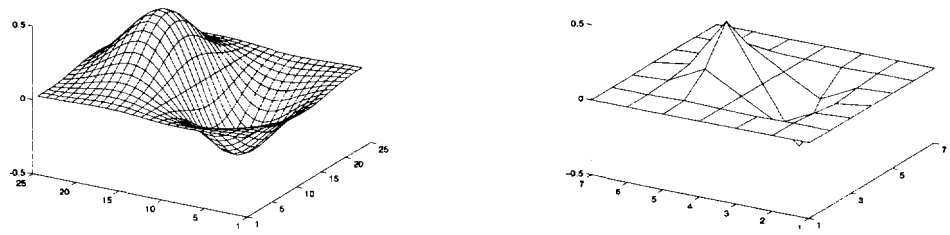

Figure 19: Orientation selective filter kernels. The two meshplots show the first derivative of Gaussian filters for the sample spacings 0.15 and 1.

Note that the leading constants have been ignored since we will only be interested in ratios and relative amplitude comparisons of the filter outputs. Figure 19 shows the filter kernel in one direction for two different sample spacings. The derivatives $Dx_l^{(L^*)}(\mathrm{m},n)$ and $Dy_l^{(L^*)}(m,n)$ of luminance channel $L_l^*(m,\mathrm{n})$ are then computed by

$$\begin{aligned} Dx_l^{(L^*)}(m,n) &= \sum_{i=1}^{M_l}\sum_{j=1}^{N_l} h_x(m-i,n-j)L_l^*(i,j) \\ Dy_l^{(L^*)}(m,\mathrm{n}) &= \sum_{i=1}^{M_l}\sum_{j=1}^{N_l} h_y(m-i,\mathrm{n}-j)L_l^*(i,\mathrm{j}). \end{aligned} \tag{36}$$

Transforming these derivatives into polar coordinates, we can compute the edge-orientation $\vartheta_l^{(L^*)}(\mathrm{m},\mathrm{n})$ and edge-amplitude $s_l^{(L^*)}(\mathrm{m},\mathrm{n})$ as

$$\begin{aligned} \vartheta_l^{(L^*)}(m,\mathrm{n}) &= \arg(Dy_l^{(L^*)}(m,\mathrm{n}),\, Dx_l^{(L^*)}(m,\mathrm{n})) \\ s_l^{(L^*)}(m,n) &= \sqrt{(Dx_l^{(L^*)}(m,n))^2 + (Dy_l^{(L^*)}(m,n))^2} \end{aligned} \tag{37}$$

where the arg computes the angle in full the range from $-\pi \leq 6 < \pi$. We implemented this method computing the luminance orientation maps

$$\Theta_l^{(L)}(m,n) = \begin{bmatrix} \vartheta_l^{(L)}(m,n) \\ s_l^{(L)}(m,n) \end{bmatrix} \tag{38}$$

as well as the color orientation maps $\Theta_l^{(a^*)}(m,\mathrm{n})$ and $\Theta_l^{(b^*)}(m,\mathrm{n})$ for all pyramid levels l. We found that angular image comparisons based on this method improved the performance of our image similarity metric considerably in comparison to using the quadrature filter method. We believe that this difference in performance is due to the loss of contour polarity information in the energy computation of the quadrature filter pair.

Experimenting with different sample spacings we found that a is not a critical parameter. In order to obtain a kernel which is sufficiently small to be used on the lowest pyramid level and to still obtain some spatial smoothing, we chose $a = 1$. However, even replacing the first derivative of Gaussian filters with simple derivative kernels of the form h = [−1, 1]

resulted only in a small loss in similarity metric performance. This is due to the fact, that the generation of high frequency noise by such small kernels is limited since the pyramid levels are already low-pass filtered.

Similarly to the results for the contrast channels, a quantization of the edge-amplitude values resulted in improved similarity metric performance. Analogous to (28) we obtain the quantized edge-amplitude $\tilde{s}_l^{L^*}(\mathrm{m,n})]$ **as**

$$\tilde{s}_l^{(L^*)}(m,n) = \begin{cases} 1 & : \quad s_l^{(L^*)}(m,\mathrm{n}) \geq T_{s_l^{(L^*)}} \\ 0 & : \quad \text{otherwise} \end{cases} \tag{39}$$

The quantization thresholds $T_{s_l^{(L^*)}}$ are computed using the edge-amplitude mean $\mu_{s_l^{(L^*)}}$ calculated over the same set of 200 images q as the variance in (30).

$$\begin{aligned} T_{s_l^{(L^*)}} &= A^* \mu_{s_l^{(L^*)}} \\ \mu_{s_l^{(L^*)}} &= \frac{1}{200 M_l N_l} \sum_q \sum_{m=1}^{M_l} \sum_{n=1}^{N_l} s_l^{(L^*)}(m,n) \end{aligned} \tag{40}$$

The constant **A*** was experimentally determined to be $\mathbf{A^*} = 0.7$. The quantized luminance orientation map $\tilde{\Theta}_l^{(L^*)}(\mathrm{m,n})$ is then given by

$$\tilde{\Theta}_l^{(L^*)}(m,n) = \begin{bmatrix} \vartheta_l^{(L^*)}(m,n) \\ \tilde{s}_l^{(L^*)}(m,n) \end{bmatrix} \tag{41}$$

Again, the equations for the quantized color orientation maps $\tilde{\Theta}_l^{(a^*)}(m,\mathrm{n})$ and $\Theta_l^{(b^*)}(\mathrm{m,n})$ are analogous. In the remainder we will refer to these quantized orientation maps **as** orientation channels.

Figures 20 and 21 show visualizations of the quantized luminance orientation-maps for our two example images. The visualizations represent each angular map entry by a short line, oriented at 0, 45, 90 or 135 degrees. The colors have been introduced in order to display the full angular range $-\pi \leq 6 < \pi$. In the positive direction of x from left to right, red lines correspond to dark to light transitions, whereas black lines represent transitions from light to dark.

In conclusion, we developed and implemented a multiscale channel model which includes color, contrast, color-contrast and orientation-selective channels. In particular we proposed a new contrast computation based on the uniform Lab color-space. Furthermore, we found that for general image similarity assessment, angular orientation maps are more efficient than separate channels for different orientations. Finally, it appears to be important to retain edge polarity information in the orientation maps. Table 1 shows a list of the computed channels.

4.3 Feature extraction and distance computation

The HVS model provides us with pyramid representations of color, contrast;, color-contrast and orientation maps. From these channels we need to extract features for image comparison. The choice of features extracted is closely linked to the desired feature invariance.

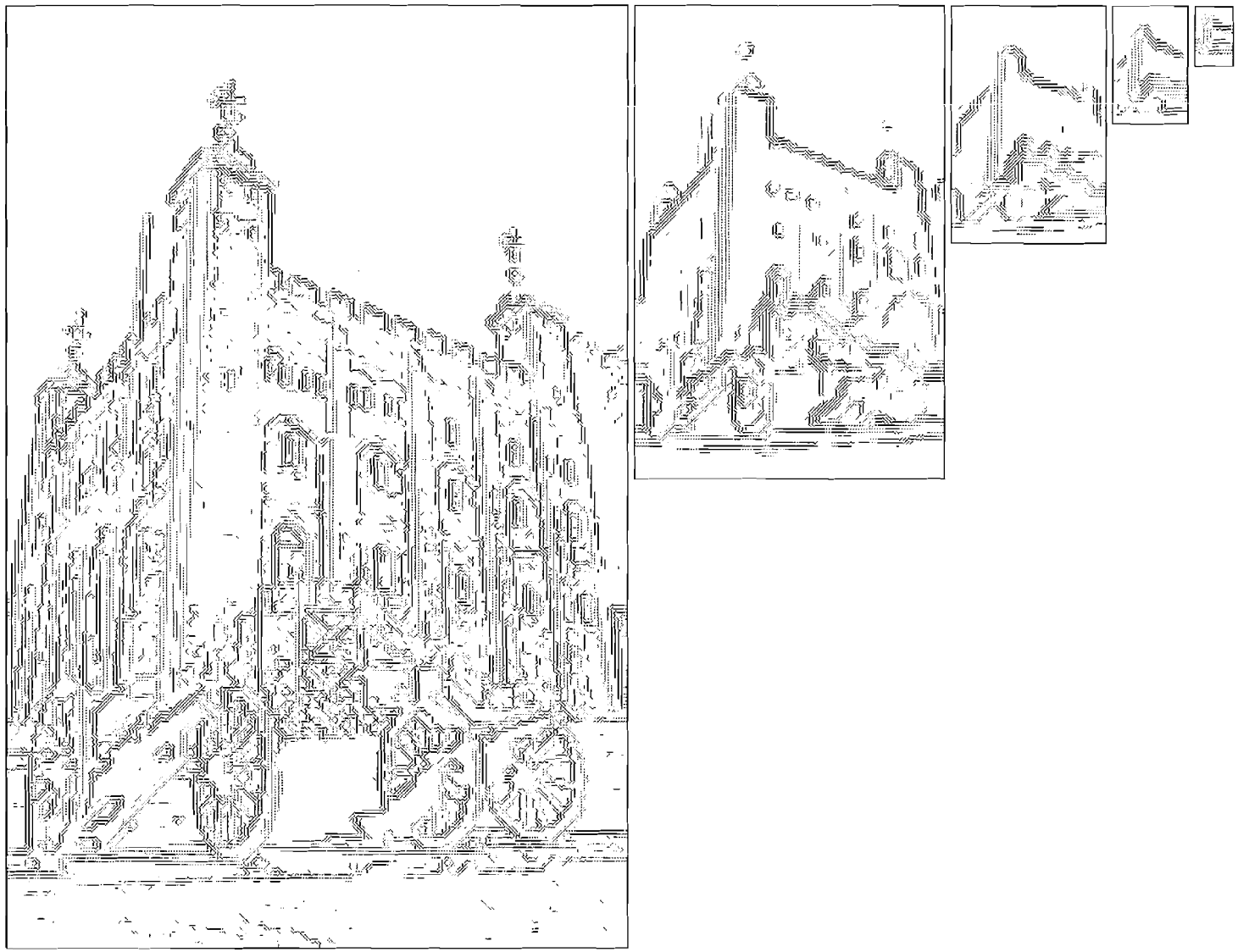

Figure 20: Visualization of the luminance orientation-map representation for example image 1. Each angular value is visualized as a short line oriented in the directions 0, 45, 90 or 135 degrees. In the positive x direction from left to right, red lines correspond to luminance transitions from dark to light and black lines to transitions from light to dark.

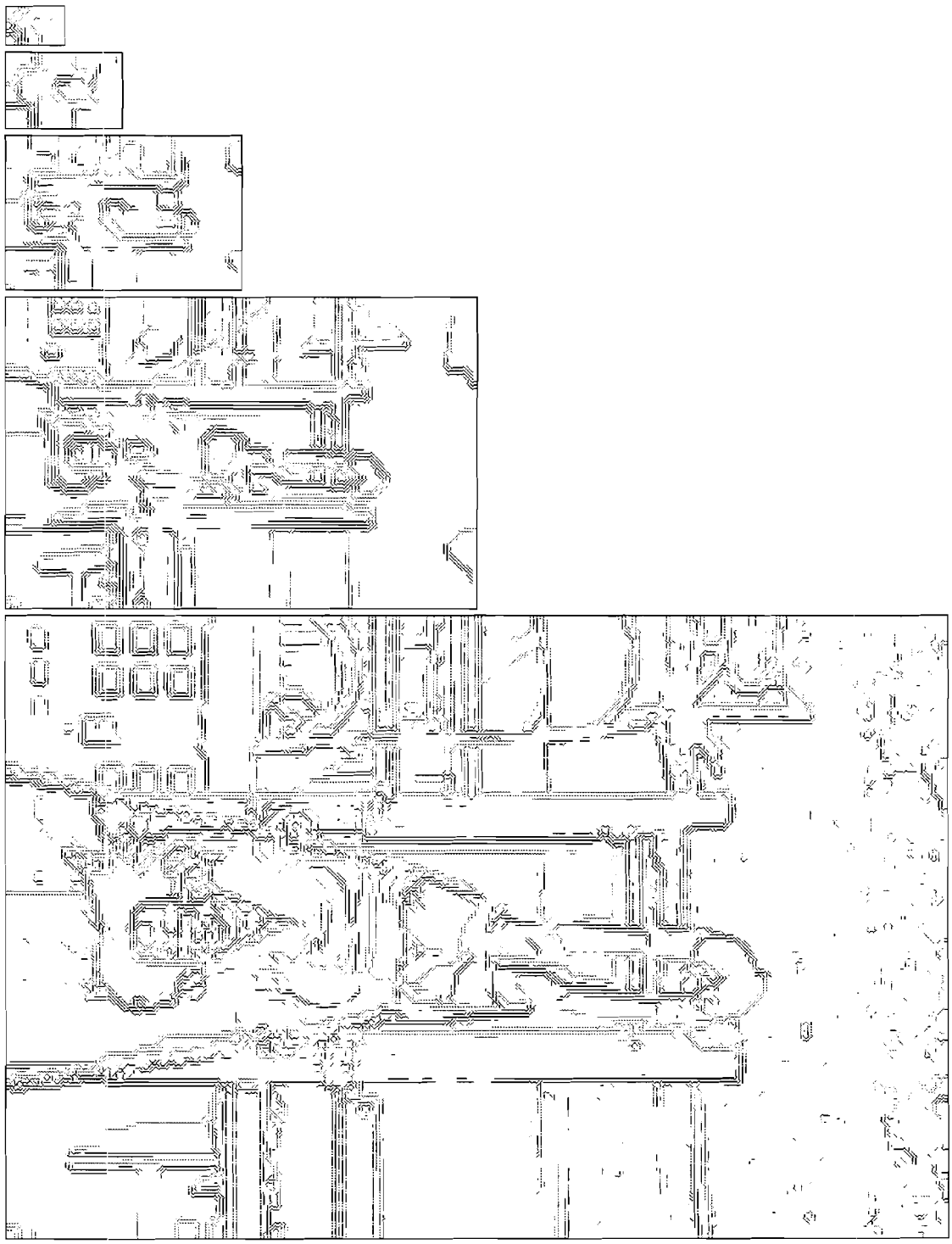

Figure 21: Visualization of the luminance orientation-map representation for example image 2. Each angular value is visualized as a short line oriented in the directions 0, 45, 90 or 135 degrees. In the positive x direction from left to right, red lines correspond to luminance transitions from dark to light and black lines to transitions from light to dark.

Type		levels l and level differences i	NO. channels
luminance	L_l^*	$l = \{0,1,2,3,4\}$	5
color	a_l^*		5
	b_l^*		5
contrast	$\tilde{C}_{l,i}^{(Y)}$	$(l,i) = \{(0,1),(0,2),(0,3),(0,4),(1,1),$ $(1,2),(1,3),(2,1),(2,2),(3,1)\}$	10
color-contrast	$\tilde{C}_{l,i}^{(a^*)}$		10
	$\tilde{C}_{l,i}^{(b*)}$		10
orientation	$\tilde{\Theta}_l^{(L^*)}$	$l = \{0,1,2,3,4\}$	5
	$\tilde{\Theta}_l^{(a^*)}$		5
	$\tilde{\Theta}_l^{(b^*)}$		5
		Total:	60

Table 1: Channels computed by the HVS model. The table gives an overview of the channels obtained from the HVS model as a function of type, level and level difference. The last column contains the number of channels computed for the respective type.

As mentioned in the introduction, histograms of the entire image are too invariant to be consistent with human perception. A tempting concept is to use a clustering algorithm to perform image segmentation. Theoretically, the clustering algorithm segments the image into objects which can then be compared and matched at different locations in the image. In practice, however, existing algorithms are unable to segment natural images into meaningful objects. Typically, similar images are segmented very differently which makes comparison of meaningful objects and regions impossible.

To avoid segmenting similar images differently, the approach taken in this work is based on the identical segmentation of all images. In particular, all channels except the orientation maps are partitioned into a fixed set of rectangular blocks. If the channel size at level l is M_l x N_l, the channel is divided into an array of U_l x V_l blocks of size $M_{B_l} \times M_{B_l}$. If $\lfloor\rfloor$ denotes the floor operation, the size of the block array is given by

$$\begin{aligned} U_l &= \left\lfloor \frac{M_l}{M_{B_l}} \right\rfloor \\ V_l &= \left\lfloor \frac{N_l}{M_{B_l}} \right\rfloor . \end{aligned} \tag{42}$$

Let $\mathcal{C}_l(\mathrm{m,n})$ denote an arbitrary channel such as $L_l^*(\mathrm{m,n})$. The block $B_{l,u,v}(m,\mathrm{n})$ at position (u,v) in the array can then be extracted from $\mathcal{C}_l$ as

$$B_{l,u,v}(m,n) = \mathcal{C}_l(m + uM_{B_l}, n + vM_{B_l}) \quad \text{for} \quad \mathrm{m,n} = 1 \ldots M_{B_l} \tag{43}$$

where $0 \leq \mathrm{u} < U_l$ and $0 \leq v < V_l$. The underlying strategy of this partitioning is to dynamically match the blocks of two different images. If we compare features of blocks in different locations, we can identify regions of similar feature behavior and match the blocks accordingly.. Currently, we have not yet implemented the dynamic matching and therefore,

are limited. to comparing images of similar size. The concept, however, serves as the basic motivation to perform a block-wise comparison of the query and target images.

The selection of features computed for each block depends considerably on how the blocks will be compared. Common methods include the mean-squared error (MSE), histogram matching and statistical modeling. Most similarity metrics prefer histograms and statistical modeling to the mean-squared error because the MSE is not well suited to compare images directly. In a pixel-wise comparison of high-resolution images, small spatial differences and uniform color shifts result in a high MSE, whereas color-histograms and statistical models remain almost unaffected. In the multiscale framework, however, we can calculate the MSE of spatial averages so that small spatial deviations will affect only the error in the high-resolution channels as desired. Furthermore, we distinguish between color and contrast channels so that a uniform color-shift will affect only the distance of the color-channels but not of the contrast and color-contrast channels.

For these reasons, we think that the mean-squared error can serve as a first approximation to a meaningful comparison for the color, contrast and color-contrast channels of our model. In order to compare the channels of query and target image, we calculate two different MSE distances. The first is the pixel-wise MSE which can be considered to be the MSE of the block-means setting the block size equal to one. A linear averaging over blocks is not necessary, since the pyramid levels contain spatial averages already. Consequently, the channel features for this distance are simply the pixel values of the pyramid. representation. If $\mathcal{C}_l(\mathrm{Q},\mathrm{m},\mathrm{n})$ and $\mathcal{C}_l(T,\mathrm{m},\mathrm{n})$ denote the same channel of the query image Q and the target image T, the distance d_μ is given by

$$d_\mu = \frac{1}{M_l N_l} \sum_{m=1}^{M_l} \sum_{n=1}^{N_l} (\mathcal{C}_l(Q,m,n) - \mathcal{C}_l(T,m,n))^2. \tag{44}$$

The second distance computed is the mean squared error of the block variances. This concept is 'based on the fact, that considerable information about the appearance of images is contained in the statistical behavior of image regions. More specifically, humans perceive very fine contrast patterns not as single contours but as averaged textures. Since the local contrast of uniform patterns over larger image regions averages out to zero, the HVS is likely to perform an energy calculation over such regions. In our model, we first calculate the block variances $\sigma^2_{B_{l,u,v}}$ for each channel as

$$\sigma^2_{B_{l,u,v}} = \frac{1}{M^2_{B_l}} \sum_{m=1}^{M_{B_l}} \sum_{n=1}^{M_{B_l}} (B_{l,u,v}(m,n) - \mu_{B_{l,u,v}})^2 \tag{45}$$

where

$$\mu_{B_{l,u,v}} = \frac{1}{M^2_{B_l}} \sum_{m=1}^{M_{B_l}} \sum_{n=1}^{M_{B_l}} B_{l,u,v}(m,n). \tag{46}$$

The distance d_{σ^2} between the query image Q and the target image T is then obtained by calculating the mean square error of the block variances

$$d_{\sigma^2} = \sqrt{\sum_{u=0}^{U_l-1} \sum_{v=0}^{V_l-1} (\sigma^2_{B_{l,u,v}}(Q) - \sigma^2_{B_{l,u,v}}(T))^2} \tag{47}$$

The block-sizes M_{B_l} for the variances were selected under the consideration that the blocks should be smaller than any objects of interest in the image. The current configuration uses a constant number of blocks at each pyramid level, which implies that the block-size at pyramid level l is half of that of level $l-1$

$$M_{B_l} = \frac{1}{2} M_{B_{l-1}} \tag{48}$$

where $M_{B_0} = 16$. Since this yields a block size of $M_{B_4} = 1$ for $l = 4$, we only compute block variances for channels with $l \leq 3$. Note, however, that the constant number of blocks in each level implies that the variance computation is always performed at the same resolution. Future work will investigate the effects of different block-sizes for the variance computation.

The comparison of the orientation channels is similar to the MSE comparison in the color and contrast channels. An important difference, however, is that the orientation maps contain two entries at each pixel location - edge-orientation and edge-amplitude. In order to compute a distance measure which combines both entries, we can calculate the mean square error of the angular differences weighted by a function of the edge-amplitudes. An intuitive way of weighting the angular difference is to use the average of the edge-amplitudes of the query and the target image. For quantized orientation maps, however, this implies that if the edge-amplitude is zero in one of the images but equal to one in the other image, the edge-angles will still be compared and still be considerably weighted.. This effect is generally not desirable since it results in an angular comparison of important contours in one image to background texture orientation in the other. A different approitch is motivated by Jacobs, Finkelstein and Salesin in [13], where wavelet coefficients of a query and a target image are only compared if the coefficient in the query image is not quantized to zero. We implemented such an unsymmetric comparison by computing the distance $d_{\tilde{\Theta}}$ between the quantized orientation maps $\tilde{\Theta}_q(\mathrm{m},n)$ and $\tilde{\Theta}_t(\mathrm{m},n)$ as

$$d_{\tilde{\Theta}} = \sum_{m=1}^{M} \sum_{n=1}^{N} [\Delta^* \vartheta(m,n)]^2 \tag{49}$$

$$\Delta^* \vartheta(m,n) = \begin{cases} \Delta\vartheta(m,n) & : \ \tilde{s}_q(m,n) = 1 \,,\ \tilde{s}_t(m,n) = 1 \\ \pi & : \ \tilde{s}_q(m,n) = 1 \,,\ \tilde{s}_t(m,n) = 0 \\ 0 & : \ \tilde{s}_q(m,n) = 0 \end{cases}$$

$$\Delta\vartheta(m,n) = \min_k (\vartheta_q(m,n) - \vartheta_t(m,n) + 2\pi k). \tag{50}$$

This comparison can be interpreted as a search for the important contours in the query image. In other words, the comparison only assigns positive distances if the target image is different at location:; where the query has important contours. The result is a better matching of target images which do not only contain objects similar to those in the query but also additional objects or strong background contours. At the same time, the unsymmetric comparison introduces some unreasonable matches if the query has only few contours which happen to line up with edges of different objects in the target. However, the overall similarity metric performance using the unsymmetric comparison on the orientation channels was slightly improved.

Channels	levels	distances	equation(s)	no. distances
L_l^*. a_l^*, b_l^*	$l = \{0, 1, 2, 3\}$	d_μ, d_{σ^2}	(44),(47)	2×12
L_l^*. a_l^*, b_l^*	$1 = 4$	d_μ	(44)	3
$\tilde{C}_{l,i}^{(Y)}$, $\tilde{C}_{l,i}^{(a^*)}$, $\tilde{C}_{l,i}^{(b^*)}$	all (I, i)	d_μ, d_{σ^2}	(44),(47)	2×30
$\tilde{\Theta}_l^{(L^*)}$, $\tilde{\Theta}_l^{(a^*)}$, $\tilde{\Theta}_l^{(b^*)}$	all l	$d_{\tilde{\Theta}}$	(50)	15
			Total:	102

Table 2: List of the computed feature distances. The table gives an overview of distances d for the different channel types, levels l and level differences i. The equation:; corresponding to the distance computations are indicated in the 4th column. The last column lists the total number of distances computed for the channels in the :row.

In summary, we compute feature distances corresponding to the color, contrast and color-contrast channels by calculating the pixel-wise mean-square error as well as the MSE of the block-variances between images. Feature distances for the orientation channels are obtained by calculating an unsymmetric MSE of angular differences. Note that this difference computation irnplies a feature representation containing the original channels as well as the block-variances. A list of the calculated feature differences is given in Table 2.

4.4 Distance selection and optimization

The comparison of each channel obtained from the HVS model provides us with 102 channel distances as shown in Table **2.** A space of this dimensionality is impractical to be used in any similarity metric because it requires intractable amounts of training data for classifier estimation. Furthermore, the space complexity is prohibitory for precomputing and storing the features for each image in the database. It is therefore particularly important to select a small set of the best features to be used in the final metric.

4.4.1 Visual tests

In order to select and optimize a subset of 'good' features, we need a cost function to evaluate the metric's performance. For the task of image comparison, this is problematic, since performance measures for image similarity metrics have not been defined. In fact, most of the existing algorithms are evaluated by looking at the result and stating whether the results look better or worse. To devise a systematic method of optimization and performance evaluation, we developed a visual test to collect experimental image matclning data. The matching data is obtained by presenting a subject with a single query image and 209 target images randomly selected from a database of 10000 images. As shown in Fig. 22, the query image and thumbnails of all target images are simultaneously displayed on the screen. The subject can click on the thumbnail images to bring up potential matches at their original size and compare them in different positions to the query image. The subject's task is to find the two images which are most similar to the query image and rate their similarity on

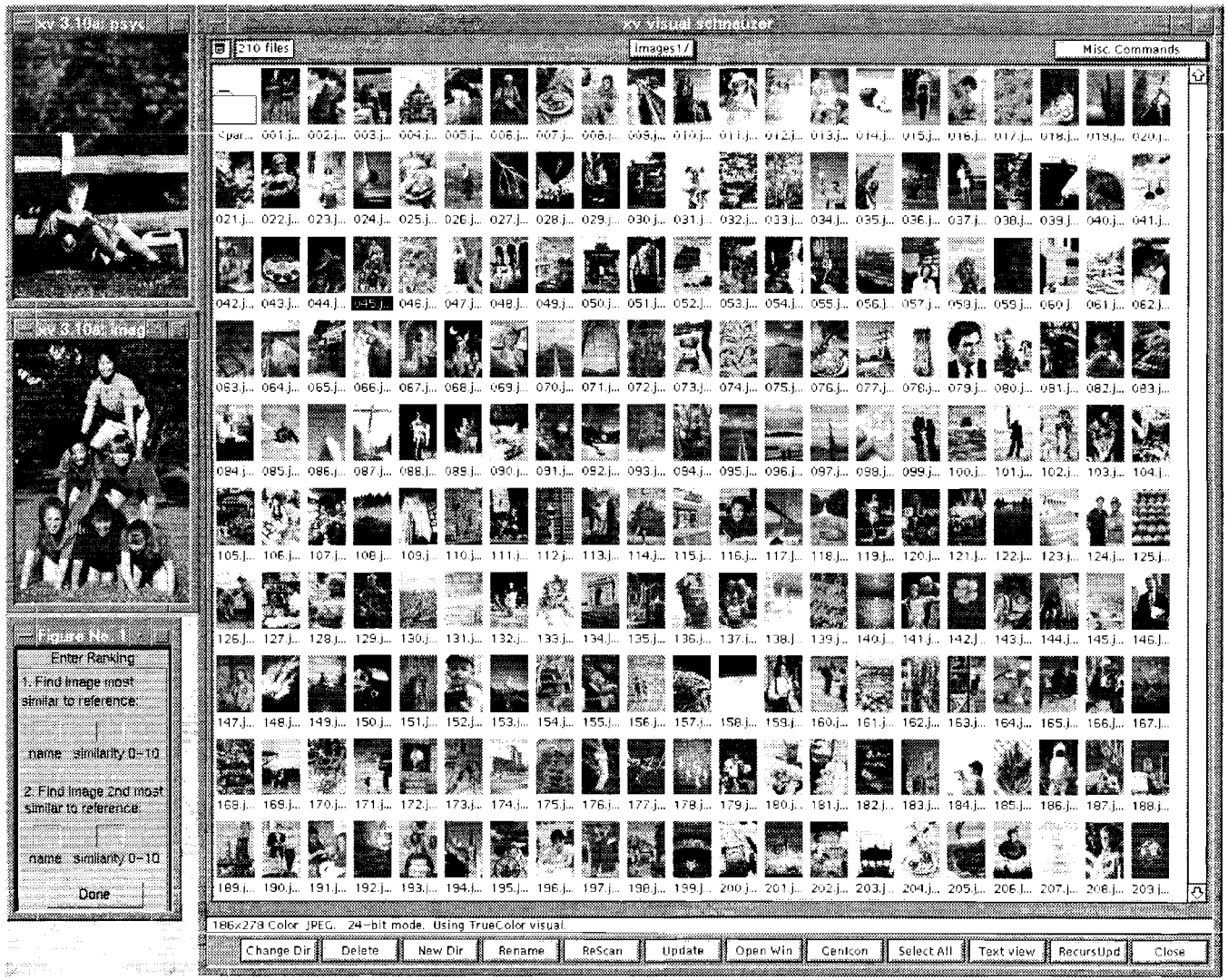

Figure 22: Visual test for experimental image matching. The query image is displayed in the upper left corner. The window on the right contains thumbnail images of the 209 target images. The subject can click on each of the thumbnail images to bring up the target image at its original size in the middle window on the left. In the lower left window, the subject enters the indices and similarity ratings of the two images most similar to the query.

a scale from zero to ten. If none or only one image is considered to be similar to the query, the subject, can leave the corresponding answer fields blank.

The number of target images was chosen to be 209 because this was the maximum number of thumbnails we could fit on the screen. Smaller sets have the advantage that the test is faster to perform, however, such sets frequently do not contain any images which are similar to the query. Furthermore, we found that it was important for the subject to view potential matches in different spatial arrangements relative to the query. This is a somewhat interesting observation since it indicates that the human perception of image similarity is less invariant to spatial perturbations than often assumed in computer visiori models. These models typically pursue a translation and rotation invariant data representation.

I have performed the test on myself and collected the data of 200 image matches. Using this experimental data, we define a cost function which accounts for the consistency in similarity rankings between the metric and the subject's choices. Let $I_1(t)$ and $I_2(t)$ be the two target images selected by a subject in visual test t. Furthermore, let S_1(t) and S_2(t) denote the subject's similarity ratings associated with I_1(t) and I_2(t). If the similarity metric is then used to order the set of target images in t from highest to lowest similarity to the query, we can define the metric's rankings of the images selected by the subject as $R_1(t)$ and $R_2(t)$. The function $c(t)$ then computes the cost for the visual test as a function of the metric's rankings and the similarities rated by the subject

$$c(t) = \tilde{c}\,[R_1(t), S_1(t)] + \tilde{c}\,[R_{2(\mathrm{t})}, S_{2(\mathrm{t})}] \tag{51}$$

where the cost function for the individual matches $\tilde{c}(R,\mathrm{S})$ is a monotonically increasing function of R and S. The main requirement for the function $\tilde{c}(R,\mathrm{S})$ is that it be consistent with the application of a user searching a large image database. A linear function of the rank R, for example, would imply that the factor of cost increase from ranking an image 5th to ranking it 10th would be the same as from ranking it 100th to 200th. Especially if we extend our problems to larger databases of many thousands of images, going through a number of images on the order of the size of the database becomes intractable and the image must be considered lost. One possibility to account for this behavior is to use a clipped linear function of R which does not assign any additional cost once the rank exceeds a certain limit R_L. This yields

$$\tilde{c}(r,S) = f(S) \times \begin{cases} R & : \quad R < R_L \\ R_L & : \quad R \geq R_L \end{cases} \tag{52}$$

where f(S) denotes a monotonically increasing function of S. We have experimented with such a function as shown in Fig. 23. A disadvantage of the clipped linear function is that it is not well suited for optimization of the similarity metric. Since changes in ranking above the clipping limit are not detected, the cost becomes constant for a wide range of model parameters. Furthermore, if we consider a more sophisticated method of database browsing such as sequentially eliminating sets of mismatches and restarting the automated search on the reduced set, the order of magnitude of large rankings becomes important. For these reasons we decided to employ an individual cost function which grows logarithmically with the rank R

$$\tilde{c}(R,\mathrm{S}) = \mathrm{f}\,(\mathrm{S})\log(R). \tag{53}$$

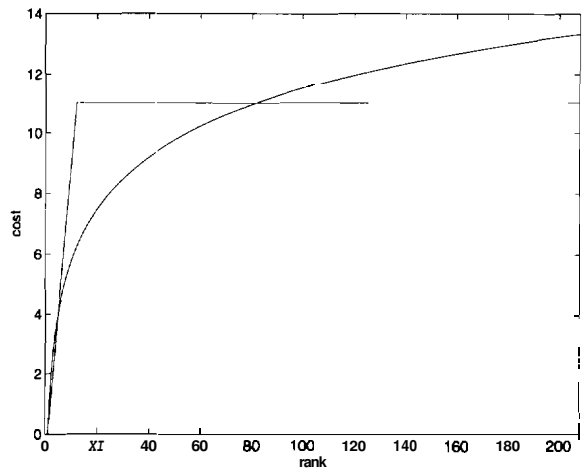

Figure **23:** Cost functions for individual matches $\tilde{c}(R, S)$. The plot shows a comparison of the clipped linear function and the logarithmic function for $f(\mathrm{S})$==1 and $R_L = 11$. Compared to the clipped linear function, the logarithmic function is more consistent with a search strategy of iteratively eliminating false matches and searching the remaining images in the target set.

Since the subjects are asked to rate similarity on a uniform scale from zero to ten, S would ideally be linearly related to perceived similarity. In practice, however, subjects might not use the full scale but only a sub-scale, for example from **3** to 8. Consequently, S must be normalized when visual tests from different subjects are considered at the same time. However, since we currently only use visual tests performed by one subject;, we use a cost function $\tilde{c}(R, \mathrm{S})$ which depends linearly on S

$$\tilde{c}(R, S) = S \log(R). \tag{54}$$

In order to obtain a global cost function C which incorporates the results of a set of T visual tests, we sum over the functions $c(t)$ for each test. This yields

$$\begin{aligned} C &= \sum_{t=1}^{T} c(t) \\ &= \sum_{t=1}^{T} (\tilde{c}[R_1(t), S_1(t)] + \tilde{c}[R_2(t), S_2(t)]) \\ &= \sum^{T} (S_1 \log R_1(t) + S_2(\mathrm{t}) \log R_2(t)). \end{aligned} \tag{55}$$

This global cost function will be used to estimate the classifier parameters of our similarity metric as well as for the final performance evaluation.

4.4.2 Selecting a classifier

A general framework for combining the feature distances extracted from the channel model into a single image similarity metric is to employ classification theory from pattern recognition. In particular, we can model the task of image comparison as a two class problem - either a target image is a match or it is not. After estimating the statistic:; for the classes 'match' and 'noise', we can compute the distance of each input to both classes. In order to

rank the similarity of a set of target images, we do not have to decide between classes but only to sort the classifier distances.

A simple method is to use a linear classifier, which computes a global distance by calculating a weighted sum of the feature distances. This seems reasonable since we know that the inputs are non-negative distances so that a result of zero corresponds to a perfect match whereas a high value indicates large differences between images.

Since valuable information might also be contained in the covariances between feature distances, we compared using a linear classifier to using a quadratic Bayes rule classifier. The results, however, indicated that in terms of similarity metric performance, the linear classifier is superior to the quadratic classifier. We therefore decided, to implement a linear classifier.

Linear classifier The linear classifier calculates the global distance D as the weighted linear combination of the feature-distances listed in Table 2. Let d_i denote the n = 102 feature-distances and ω_i the corresponding classifier weights. If furthermore H_1 denotes the class 'match' and H_0 the class 'noise', then the equation for the linear classifier is given by

$$D(d_i) = \sum_{i=1}^{n} \omega_i d_i \underset{H_0}{\overset{H_1}{<>}} T. \tag{56}$$

As mentioned above, we can avoid using the threshold T since we are only interested in ranking a set of target images. In order to use vector notation, we define d as the vector of the feature distances d_i and Ω as the vector of the weights ω_i. We can then write the global distance computation as

$$D(\mathbf{d}) = \Omega^T \mathbf{d}. \tag{57}$$

In order to estimate the classifier weights R, we first implemented a method developed for statistical signal detection [32]. The weighted sum of the feature distances in *(56)* represents an FIR filter with impulse response Ω which is applied to the input signal d. The signal detection method then estimates the Ω vector by maximizing the signal to noise ratio of the filter output. In our case, however, the filter inputs are feature distances which are small for matches and large for noise. Consequently, our goal must be to maxirnize the inverse signal to noise ratio. If E denotes the expectation operator and $\mathbf{d}_{H_1}$ and $\mathbf{d}_{H_0}$ are inputs corresponding to the classes H_1 and H_0, we can express the inverse signal to noise ratio as

$$SNR^{-1} = \frac{E\{|D(\mathbf{d}_{H_0})|^2\}}{E\{|D(\mathbf{d}_{H_1})|^2\}} \tag{58}$$

It can be shown that this leads to the generalized eigenvalue problem

$$R_{H_0}\Omega = \lambda R_{H_1}\Omega \tag{59}$$

where $R_{H_0} = E(\mathbf{d}_{H_0}^T \mathbf{d}_{H_0})$ and $R_{H_1} = E(\mathbf{d}_{H_1}^T \mathbf{d}_{H_1})$ are the covariance matrices corresponding to the inputs and λ is a generalized eigenvalue. The Ω which maximizes *(58)* is; the generalized eigenvector which corresponds to the largest eigenvalue λ.

We implemented this method by estimating the covariance matrices for preselected subsets of features and solving the generalized eigenvalue problem in Matlab. The results, however, were not very promising.

The low performance obtained using this method as well as using the quadratic classifier suggest, that methods based on the covariance between feature distances do not perform well. In order to compare the class separabilities due to mean and covariance differences, we computed the Bhattacharya distance of the two classes. The Bhattacharya distance is a well known measure of similarity between distributions [33]. We found that the class separability due the covariance between feature distances is small.

As a consequence of the limited value of the covariance based methods, we decided to pursue a different strategy which directly estimates the classifier weights without assuming a specific class distribution. This approach is based on sequential feature-distance selection followed by a stochastic search technique to optimize the weights of the selected set. An important advantage of such an approach is that the optimization technique becomes independent of the method of feature computation. This allows us to evaluate modifications of the HVS model and feature extraction. If for example we were to decide to use histogram distances instead of the MSE, we could still use the same feature selection and classification.

In order to make a random search of the feature-space feasible, we first must reduce the dimensionality of the space. To achieve this, we sequentially select feature distances in the order that leads to the most rapid improvement in the cost function derived from the visual test. More specifically, starting with a set size of zero, we sequentially add remaining distances to the selected set until the desired set size of k is reached. In order to find the distance which should be selected at a given step, we have to test adding each of the remaining distances with a fixed set of weights. However, since the different feature distances d_i have very different numerical values, we have to normalize the distances first to be able to limit the range of weights ω_i. If $\sigma_{d_i}^2$ denotes the variance of d_i over all comparisons between query and target images in the training set, the normalized feature distances $d_{i,n}$ are given by

$$d_{i,n} = \frac{d_i}{\sigma_{d_i}} \quad \forall i = 1 \ldots 102\,. \tag{60}$$

Let $\mathcal{S}_j$ denote the set containing a number of j distances $d_{i,n}$ with associated classifier weights ω_i. At a given set size j we can then write the selection of the distance $d_{i,n}^*$ with weight ω_i^* as

$$\mathcal{S}'_{j+1} = \mathcal{S}_j \cup \{(d_{i,n}^*, \omega_i^*)\} \quad d_{i,n}^* \notin \mathcal{S}_j. \tag{61}$$

In order to select the distance to be added to $\mathcal{S}_j$, we test adding each of the remaining distances with a set of different weights $\mathcal{W}_j$. The selected combination $(d_{i,n}^*, \omega_i^*)$ is the one which minimizes the cost function C over the training set of visual tests

$$(d_{i,n}^*, \omega_i^*) = \arg \min_{\substack{d_{i,n} \notin \mathcal{S}_j \\ \omega \in \mathcal{W}_j}} C(S_j \cup (d_{i,n}, \omega)). \tag{62}$$

Note that the argument of C means that the cost function from (55) is evaluated using the linear classifier (56) with the distance weight combinations in the argument set. All other classifier weights ω_i are set to zero. The initial set of weights, $\mathcal{W}_0$, is selected to contain

weights from 0.1 to 1, spaced 0.1 apart. After each iteration, all weights ω_i' of the set $\mathcal{S}_{j+1}'$ are normalized such that the smallest weight is equal to one.

$$\mathcal{S}_{j+1} = \mathcal{S}_{j+1}' \;:\; \omega_i = q\omega_i^{\mathrm{I}} \quad \text{such that} \quad \min_i(\omega_i) = 1. \tag{63}$$

For the selection of the next distance, the set of weights $\mathcal{W}_{j+1}$ is extended to contain all weights from 0.1 to twice the maximum weight of the set $\mathcal{S}_{j+1}$.

$$\mathcal{W}_{j+1} = \{0.1, 0.2, \ldots, \max_{\omega_i \in \mathcal{S}_{j+1}} 2\omega_i\} \tag{64}$$

This normalization strategy is based on the observation, that the algorithm usually selects the distances in order of decreasing weights. Consequently, the weight sets $\mathcal{W}_{j+k}$ of subsequential steps should contain a good selection of weights smaller than those contained in $\mathcal{S}_j$. The selection of distances is terminated when $\mathbf{j} = k$, i.e. the desired set size is reached. We finally normalize all weights ω_i in $\mathcal{S}_k$ to the range $0 \leq \omega_i \leq 1$. In the remainder, we will denote these normalized selected weights by the vector $\Omega_s = [\omega_1, \ldots, \omega_k]^T$ and the selected normalized distances by $d_{s,j}$ where $1 \leq \mathbf{j} \leq k$.

After selecting the feature subset, we use simulated annealing to optimize the feature weights. Simulated annealing is a stochastic search technique based on Markov chains and thermodynamical principles [34]. The method relies on computing acceptance probabilities for randomly perturbed weights. In order to optimize Ω_s, we randomly perturb one of the weights ω_j resulting in the perturbed weight vector Ω_s'. Let $C(\Omega_s)$ and $C(\Omega_s')$ denote the values of the cost function C from (55) using the linear classifier with the distances $d_{s,j}$ and weights $C(\Omega_s)$ and $C(\Omega_s')$. Energies corresponding to the cost of these sets are then obtained by

$$\begin{aligned} E(\Omega_s) &= \frac{C(\Omega_s)}{T} \\ E(\Omega_s') &= \frac{C(\Omega_s')}{T} \end{aligned} \tag{65}$$

where the temperature T is a parameter. We then calculate an acceptance :probability α as the ratio of the Gibbs measures of the perturbed and unperturbed sets.

$$\begin{aligned} \alpha &= \min\left(\frac{e^{E(\Omega_s)}}{e^{E(\Omega_s')}}, 1\right) \\ &= \min\left(e^{\frac{1}{T}(C(\Omega_s) - C(\Omega_s'))}, 1\right) \end{aligned} \tag{66}$$

We see that if the perturbed set performs better than the unperturbed, i.e. $C' \leq C$, it is accepted with probability one. If, however, the perturbed set performs worse, i.e. $C' > \mathrm{C}$, then there is still a probability of $\exp((C - C')/T)$ to accept it. Clearly, the parameter T determines how conservative the acceptance probability for $\mathrm{C}' > C$ will be.

In order to calculate the complete transition probability from Ω_s to Ω_s', we must first choose $p(j)$, the probability of perturbing the j'th element of the vector Ω_s. We will let

$p(j) = 1/k$, so it is uniformely distributed. Then, the total transition probability $p(\Omega'_s|\Omega_s)$ is given by

$$p(\Omega'_s|\Omega_s) = \prod_{j=1}^{l} \alpha f(\omega'_j|\omega_j) \tag{67}$$

where $f(\omega'_j|\omega_j)$ is the probability distribution for changing ω_j to ω'_j. It can be shown, that if the transition probabilities are symmetric, i.e. $p(\Omega'_s|\Omega_s) = p(\Omega_s|\Omega'_s)$, the sequence of weight vectors $\Omega_s(n)$ in the algorithm satisfies the properties of a regular Markov chain.

In order to obtain symmetric transition probabilities, we require that the $f(\omega'_j|\omega_j)$ be symmetric. This can be obtained by perturbing ω_j with an additive random variable symmetrically distributed around zero. In that case, $f(\Delta\omega_j)$ becomes a function of the absolute difference between ω_j and ω'_j only

$$f(\omega'_j|\omega_j) = f(|\omega'_j - \omega_j|). \tag{68}$$

However, since we use the ω_j to weight feature distances, we would like to restrict them to non-negative values. We can achieve this without loss of symmetry by limiting the ω_i to the range $0 \leq \omega_i \leq 1$ and perform a wrap-around whenever the perturbation violates these limits. The final perturbation of the selected weight ω_j using the random variable ξ can then be expressed as

$$\omega'_i = \begin{cases} \omega_i + \xi & : \quad 0 \leq \omega_i + \xi \leq 1 \\ (\omega_i + \xi) + 1 & : \quad \omega_i + \xi < 0 \\ (\omega_i + \xi) - 1 & : \quad \omega_i + \xi > 1 \end{cases} \tag{69}$$

where ξ is uniformly distributed from $-\phi \ldots \phi$

$$\xi = \mathcal{U}[-\phi \ldots \phi]. \tag{70}$$

The resulting regular Markov chain with states Ω_s and transition probabilities $p(\Omega'_s|\Omega_s)$ can be shown to converge to a local minimum for sufficiently small T. For a specific schedule of iterations using different T's, the convergence becomes global. The global convergence, however, is very slow so that in any practical application the algorithm is stopped after a fixed number of iterations resulting in an approximate solution.

We implemented this technique as specified above limiting the the range of the ω_j from zero to one and using a $\phi = 0.02$. Reasonable convergence behavior was obtained by starting the iteration with a $T = 20$ and successively decreasing it to $T = 0.5$. After approximately 10^5 iterations, no further convergence was noticeable. Note that these values are based on a set size of $k = 13$ and cost function values C in the range from 500 to 1500.

The results obtained by sequentially selecting features and optimizing their weights using simulated annealing were far superior to those of the covariance based methods. Note also, that the direct method estimates only the k weights in comparison to the $2(k^2+k)$ parameters needed by the covariance methods. This is an important advantage, since it implies that the direct methods requires only a fraction of the amount of training data, needed for the covariance based methods.

Weight ω_i	Type	Channel	Distance-Type	Level l and difference i
1.000	color	b^*_4	d_μ	$l = 4$
0.972	orientation	$\tilde{\Theta}^{(a^*)}_4$	$d_{\tilde{\Theta}}$	$l = 4$
0.962	color	a^*_4	d_μ	$l = 4$
0.599	luminance	L^*_3	d_μ	$l = 3$
0.422	orientation	$\tilde{\Theta}^{(b^*)}_4$	$d_{\tilde{\Theta}}$	$l = 4$
0.289	contrast	$\tilde{C}^{(Y^*)}_{0,1}$	d_{σ^2}	$(l, i) = (0, 1)$
0.286	orientation	$\tilde{\Theta}^{(L^*)}_3$	$d_{\tilde{\Theta}}$	$l = 3$
0.256	color-contrast	$\tilde{C}^{(a^*)}_{2,1}$	d_{σ^2}	$(l, i) = (2, 1)$
0.230	contrast	$\tilde{C}^{(Y^*)}_{1,1}$	d_{σ^2}	$(l, i) = (1, 1)$
0.223	color-contrast	$\tilde{C}^{(a^*)}_{0,1}$	d_{σ^2}	$(l, i) = (0, 1)$
0.194	orientation	$\tilde{\Theta}^{(b^*)}_3$	$d_{\tilde{\Theta}}$	$l = 3$
0.177	color	b^*_3	d_μ	$l = 3$
0.148	orientation	$\tilde{\Theta}^{(L^*)}_2$	$d_{\tilde{\Theta}}$	$l = 2$

Table **3:** Selected feature distances. The table shows the feature distances selected for the training set of 80 visual tests. The distances are listed in the order of decreasing weights ω_i.

5 Results

The results presented are based on a selection of **13** out of **102** features distances computed on a training set of 80 visual tests.

5.1 Selected features

Table **3** shows the selected distances in the order of decreasing weights. Since we normalized the distances d to have zero mean and unit variance, the values of the weights are consistent between channels. Out of the **13** selected distances, there are 4 luminance/color, 4 contrast/color-contrast and **5** orientation channel distances. This is remarkable, since it indicates that all types of channels including color-contrast and orientation channels contribute to the classification. However, the distribution of weights implies a ranking of the types where color and luminance are of highest importance followed by orientation and contrast.

The examination of the selected distances shows, that for the contrast channels only variance distances d_{σ^2} were selected. This is due to the fact, that the contours of different images in the contrast representation do not line up very well. The variance computation, however, is a measure for the statistical behavior of a block region and can therefore successfully be compared between images even if single contours do not line up. Consequently, the selected contrast distances correspond to texture comparisons between images.

Looking: at the selected pyramid levels, we see that most of the color and orientation map comparisons are made at low resolutions. For the color channels, this is consistent with

vision science and re-emphasizes that the color space be uniform at low spatial frequencies. The selected contrast levels and level differences seem to suggest, that contrast comparisons at high pyramid levels l with low level differences i are meaningful. However, this result must be further investigated since the block-sizes for the variances are currently chosen to yield a constant resolution. In conclusion, the selection of features is very promising. It suggests, that all types of representations computed by the HVS model might be important for image similarity assessment.

5.2 Metric Performance

In order to evaluate the metric's performance, we tested it on a set of **80** trained and **80** untrained visual tests. Figures **24** and **25** show a selection of the matching results for the trained and untrained case. Each row of the figure contains a query image followed by the images selected by the subject and the six best matches found by the metric. The results indicate that the metric has considerable potential in matching images consistently with human observers. In particular, the metric is capable of finding matches which only share single aspects of similarity. In contrast to the color-histogram methods, the metric matches images containing similar shapes and textures but different colors. Given that we use a fairly small set of target images, we can not expect to see five or more good matches on a single visual test since most sets contain only **1** to **3** good matches. Instead, we focus on the ranking: of the images selected in the visual test which we consider very promising. Over the entire untrained set we observed that approximately **50%** of the images selected by the subjects were among the **10** best matches predicted by the metric. Note that this compares to a probability of **4.9%** for random selection.

In order to perform a more systematic analysis, we plotted the classification accuracy of the metric as shown in Figs. 26 to **30.** The values on the y-axis denote the percentage of subject matches found within the first number of metric matches drawn on the x-axis. The x-axis is drawn in logarithmic scale since we are mainly interested in the classification accuracy obtained within the first few matches selected by the metric. In the following, we will refer to results for the untrained set simply as results.

Figure 26 compares our results to the performance obtained using a method developed by Bouman and Chen [35]. This method uses block-wise color-histogram. matching and orientation selective features which are similar to our orientation maps. While the histogram method obtains higher accuracies for the first two matches, our method outperforms it for the rest of the range. The superiority of our method at higher ranks is remarkable since we have to consider at least **5** to **10** metric matches to obtain reasonable classification accuracies over **40%.** The third curve in this diagram shows the **95%** confidence interval of the performance that could be expected from performing a random selection of target images. Clearly, both methods perform considerably better than random selection.

In order to investigate the importance of the contrast channels, we trained the metric excluding these channels. Figure **27** shows a comparison to the performance obtained by allowing the selection of all channels. In the range of interest between rank **5** and **30,** a considerable increase in performance is obtained by using the contrast features. Although the distance selection and optimization process tend to assign lower weights to the contrast

Figure **24:** .Matching results on the training set. Each row corresponds to a different visual test where the image in the first column is the query and the two images in columns **3** and 4 are the matches selected by the subject. In this case, the metric is trained on these matches. The columns on the right show the images selected by the metric from rank 1 to 6.

Figure 25: Matching results on the untrained set. The images are in the same spatial arrangement as in Fig. 24. In this case, the metric is not trained on the query images shown in columns 2 and 3. The results indicate, that; the metric has high potential in finding images which are consistent with human similarity perception.

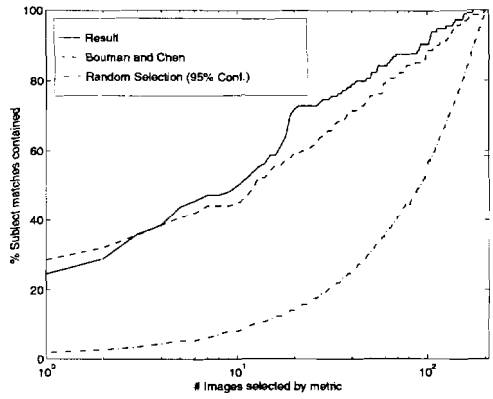

Figure 26: Performance comparison to the method developed by Bouman and Chen. In the range of interest between rank **5** and **20,** our model achieved higher classification accuracies than the method by Bouman and Chen. The third curve in this diagram shows the **95%** confidence interval of the performance that could be expected from a random selection of target images. Clearly, both methods perform considerably better than random selection.

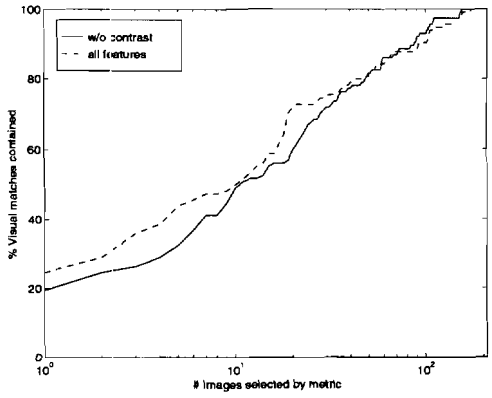

Figure **27:** Performance analysis of contrast channels. The curves indicate that for rankings between **1** and **30** the contrast features resulted in a considerable improvement in accuracy.

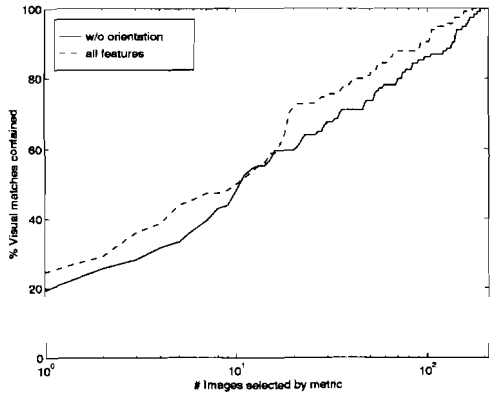

Figure 28: Performance analysis of orientation channels. By using the orientation channel distances, the metric's performance is considerably increased.

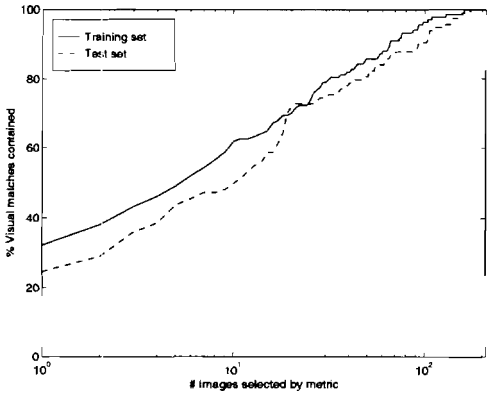

Figure 29: Comparison of training and test set performance. The graphs indicate, that the accuracies on the trained and the untrained sets differ by not mort: than 10-15%.

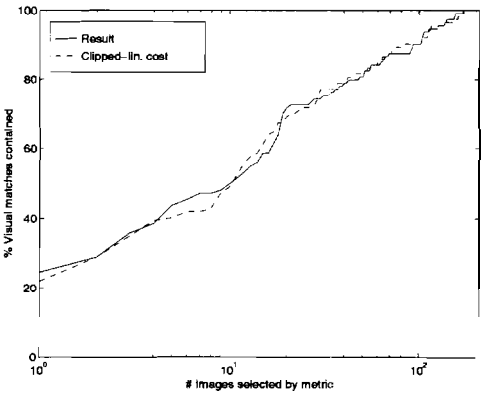

Figure 30: Comparison of the logarithmic and the clipped linear cost function. The accuracy using the clipped linear function drops before reaching the clipping limit of rank 11.

and color-contrast channels, these channels result in a significant increase in performance which cannot be obtained using the color and orientation channels only.

A similar analysis for the orientation channels is shown in Figure 28. The figure shows the classification result obtained by excluding the orientation channels compared to the result obtained by allowing all channels. Clearly, the orientation channels increase the metric's performance throughout the range of ranks.

An important question is whether the classification accuracies on the training set and the test set are comparable. A large difference between the cases would imply that the model is over-parameterized for the amount of training data available. As shown in Fig. 29 the classification accuracies obtained on the training and test set differ by not more than 10 to 15% which is in the common range for pattern recognition algorithms. We therefore conclude, that our model is not overparameterized. However, more training data would be desirable in order to capture the similarity perception of different subjects and reduce performance differences between trained and untrained data.

Finally, we compared training the metric using the clipped linear function from Fig. 23 instead of the logarithmic function from (53). The result is shown in Fig. 30. The curves indicate that the accuracy obtained using the different cost functions is almost identical. Note, however, that the performance of the clipped linear function decreases directly below the clipping limit of 11. The reason for this might be that the clipped function can not optimize the ranking of matches directly above the clipping matches. In contrast, differences in rankings in this range are still substantially weighted by the logarithmic function.

Overall, the performance of the metric is very promising. Although it captures different aspects of similarity using a fixed set of features and weights, the rate of unreasonable matches is acceptable.

6 Conclusions

In this work we presented the development of an image similarity metric based on features extracted from a simple model of the human visual system. Our emphasis is not so much on the specific model, but on the methodology of feature optimization and metric evaluation. The presented optimization strategy is independent of the underlying image representation and therefore suited to systematically optimize and compare different kinds of image similarity metrics.

The visual test that we propose is only a first approximation to a more comprehensivly designed and psychologically relevant measurement. However, we believe that the method is an important step toward a more standardized evaluation methodology. In particular, this new methodology seems to be much better than conventional methods of evaluation based on anecdotal accounts of good and poor matching results.

In addition, we have demonstrated that features such as contrast and color-contrast might be of considerable value for image similarity assessment. The performance of our model suggests that the proposed methodology can lead to derived similarity metrics which have substantial value in predicting image similarity as perceived by human subjects.

6.1 Future work

The methodology we describe can be improved in many ways. In particular, the feature extraction and distance comparison of the contrast channels must be further investigated. In the first stage, this will include experimenting with different block-sizes for the variance, replacing the binary threshold with a sigmoid-shaped nonlinearity and implementing the dynamic block matching. In addition, we will seek to improve the distance selection and optimization. In particular, simulated annealing iterations should be performed directly after the selection of each feature distance. A further improvement might be obtained by selecting more than the desired number k of feature distances and then performing a Branch and Bound search to optimally reduce the space.

The proposed visual test methodology is only a first approximation to a psychological meaningful measurement. The current task to 'find similar images' leaves the subject with a great amount of interpretation which introduces noise to the measurements. More accurate measurements can be obtained by decoupling the different aspects of similarity into separate tests. Such a method is based on the psychological concept that the quantities intrinsic to the human visual system should be invariant to changes of other aspects. Under this assumption, we will seek to identify aspects of similarity which are invariant to each other. These aspects can then be measured separately to yield a better defined task for the subjects.

References

[1] R. Jain and W. Niblack, eds., NSF workshop on visual information management, February 1992.

[2] W. Niblack and R. Jain, eds., Proc. Storage and Retrieval for Image and Video Databases **I, 11** and **111,** vol. 1908; 2185 and 2420, Bellingham, Wash, SPIE, 1993, 1994 and 1995.

[3] R. Mattheus, "European integrated archiving and communication systems, CEC/AIM," Computer *Methods* and Programs in Biomedicine, vol. 45, no. 1-2, pp. 65 – 69, 1994.

[4] S. K. Bhatia, V. Lakshminarayanan, A. Samal, and G. Welland, "Human face perception in degraded images," *Journal* of Visual Communication and Image Representation, vol. 6, no. 3, pp. 280 – 295, September 1995.

[5] C. C. Chang and S. Y. Lee, "Retrieval of similar pictures on pictorial databases," Pattern Recognition, vol. 24, no. 7, pp. 675 – 680, 1991.

[6] C. C. Chang and T. C. Wu, "Retrieving the most similar symbolic pictures from pictorial databases," Information Processing and Management, vol. 28, no. 5, pp. 581 – 588, 1992.

[7] S. Y. Lee and F. J. Hsu, "2D C-string: A new spatial knowledge representation for image database systems," Pattern Recognition, vol. 23, no. 10, pp. 1077 – 1087, 1990.

[8] M. Swain and D. Ballard, "Color indexing," International Journal of *Computer* Vision, vol. 7, no. 1, pp. 11 – 32, November 1991.

[9] M. J. Swain, "Interactive indexing into image databases," Proc. of *SPIE/IS&T* Conf. on Storage and Retrieval for Image and Video Databases, vol. 1908, February 2-3 1993, San Jose, CA, pp. 95 – 103.

[10] Z. Chen and S. Y. Ho, "Computer vision for robust 3D aircraft recognition with fast library search,'" Pattern Recognition, vol. 24, no. 5, pp. 375 – 390, 1991.

[11] H. Jagadish, "A retrieval technique for similar shapes," Proc. of ACM *SIGMOD Int'l* Conf. on Management of Data, May 29-31 1991, Denver, CO, pp. 208 – 217.

[12] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," International Journal of Computer Vision, vol. 18, no. 3, pp. 233 – 254, June 1996.

[13] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast multiresolution image querying," Proc. of ACM *SIGGRAPH* Conf. on Computer Graphics, August 9-11 1995, Los Angeles, CA, pp. 277 – 286.

[14] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," Computer, vol. 28, no. 9, pp. 23 – 32, September 1995.

[15] B. A. Wandell, Foundations of Vision. Suderland, Massachusetts: Sinauer Press, 1995.

[16] C. Zetzsche and G. Hauske, "Multiple channel model for the prediction of subjective image quality," Proc. of SPIE Conf. on Human Vision, Visual Processing and Digital Display **11,** vol. 1077, January 18-20 1989, Los Angeles, CA, pp. 209 – 216.

[17] C. J. C. Lloyd and R. J. Beaton, "Design of a spatial-chromatic human vision model for evaluating full-color display systems," Proc. of SPIE: Human Vision and Electronic Imaging: Models, Methods and Applications, vol. 1249, February 12-14 1990, Santa Clara, CA, pp. 23 – 37.

[18] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 179 – 205, Cambridge, MA: MIT Press, 1993.

[19] J. Lubin, "The use of psychophysical data and models," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 171 – 178, Cambridge, MA: MIT Press, 1993.

[20] S. J. P. Westen, R. L. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel model," *Proc. of IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.,* vol. 4, May 9-12 1995, Detroit, MI, pp. 2351 – 2354.

[21] T. V. Papathomas, R. S. Kashi, and A. Gorea, "A human vision based computational model for chromatic texture segregation," *(submitted to) IEEE Trans. on Systems, Man and Cybernetics.*

[22] E. M. Granger, "Uniform color space as a function of spatial frequency," *Proc. of SPIE Conf. on Human Vision, Visual Processing, and Digital Display IV,* vol. 1913, February 1-4 1993, San Jose, CA, pp. 449 – 461.

[23] S. L. Guth, "Model for color vision and light adaptation," *J. Opt. Soc. Am. A,* vol. 8, no. 6, pp. 976 – 993, June 1991.

[24] D. B. Judd and G. Wyszecki, *Color in Business, Science and Industry.* New York, NY: John Wiley and Sons, 3rd ed., 1975.

[25] M. D. Fairchild and R. S. Berns, "Image color-appearance specification through extension of CIELAB," *Color research and application,* vol. 18, no. 3, pp. 178 – 190, 1993.

[26] K. T. Mullen, "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings," *Journal of Physiology,* vol. 359, pp. 381 – 399, 1985.

[27] A. B. I'oirson and B. A. Wandell, "The appearance of colored patterns: pattern-color separability," *J. Opt. Soc. Am.,* vol. 10, no. 12, pp. 2458 – 2470, 1993.

[28] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Tmns. on Communications,* vol. 31, no. 4, pp. 532 – 540, April 1983.

[29] E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A,* vol. 7, no. 10, pp. 2032 – 2040, October 1990.

[30] H. Knutsson and G. H. Granlund, "Texture analysis using two-dimensional quadrature filters," *IEEE Comput. Soc. Workshop Computer Architecture for Pattern Analysis and Image Database Management,* October 12-14 1983, Pasadena, CA, pp. 206 – 213.

[31] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 13, no. 9, pp. 891 – 906, September 1991.

[32] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1992.

[33] K. Fukanaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 2nd ed., 1990.

[34] R. Kintlermann and J. L. Snell, *Markov Random Fields and their Applications.* Providence, Rhode Island: American Mathematical Society, 1980.

[35] J. Y. Chen, C. A. Bouman, and J. P. Allebach, "Multiscale branch and bound algorithm for image database search," *(submitted to) SPIE Storage and Retrieval for Image and Video Databases V,* February 9 - 15 1997, San Jose, CA.