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

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

School of Electrical Engineering,
Purdue University,
West Lafayette, IN 47907-1285
{frese, bouman, allebach}@ecn.purdue.edu

ABSTRACT

In this paper 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 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

The recent advances in computer technology have given rise to image databases containing tens of thousands of images. These large databases necessitate efficient image retrieval methods. Of particular interest are search algorithms which find images that are similar to a sketch or a query image provided by the user. The key problem in developing such algorithms is to define an underlying similarity metric which is consistent with the human perception of image similarity.

Previous approaches have defined image similarity metrics using classical image processing techniques. While the first approaches were based on a single discriminant such as the color-histogram,[1] many recent approaches use image compression techniques to obtain features which correspond to multiple aspects of image similarity such as color, texture and shape. While some of the published metrics directly compare the compression coefficients of, for example, the wavelet compressed[2] images, others explicitly extract features from the compressed representation.[3,4] MIT's photobook,[3] for example, extracts different features for appearance, texture and shape using methods such as the Karhuenen-Loeve transformation. While metrics like these perform well for the task of comparing images which contain distinct objects, it is unclear how well they relate to human perception of the similarity of complex images of natural scenes.

Although the objective of all image similarity metrics is to be consistent with human perception, little work has been done to systematically examine this consistency. For the specific case of shape similarity, Scasselati,

*Appeared in the *Proc. of SPIE/IS&T Conf. on Human Vision and Electronic Imaging II,* vol. 3016, pp. 472-483, San Jose CA, Feb. 10-13, 1997.

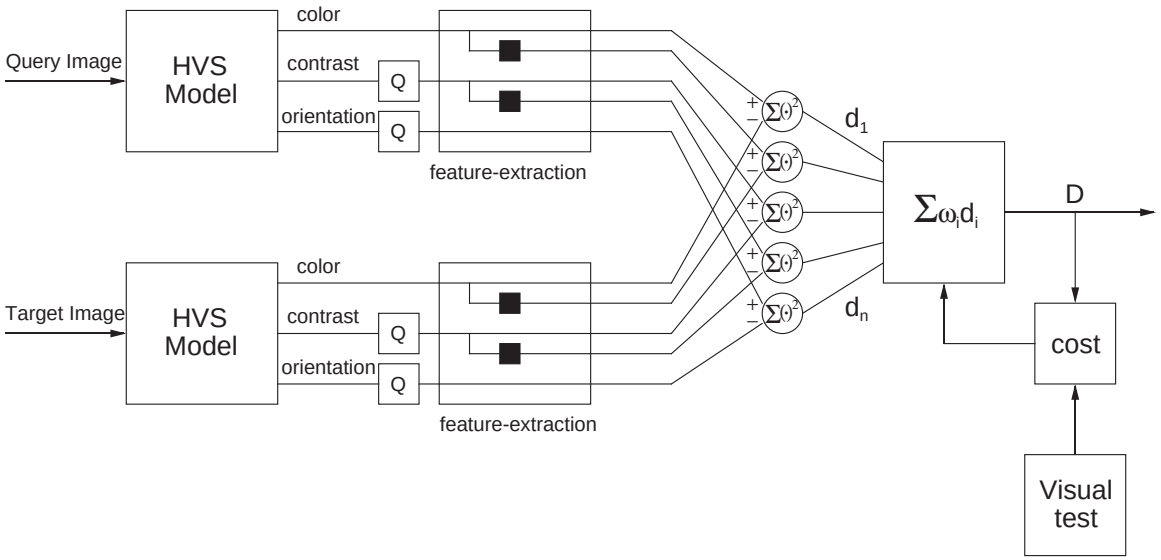


Figure 1: System overview: The query and the target images are processed by an HVS model to obtain a multiscale representation with separate channels for color, contrast and orientation. This is followed by a stage called feature-extraction which computes features for the actual image comparison. Query and target are then compared by taking the mean-square error separately for each feature, resulting in feature-distances d_i. These feature distances are combined into a global similarity measure using a linear classifier. The classifier weights are estimated using image matching from human subjects collected in a simple visual test.

Alexopoulos and Flickner[5] have used shape similarity judgments from human subjects to evaluate the performance of shape distance metrics. More commonly, however, the performance of similarity metrics is evaluated based on anecdotal accounts of good and poor matching results.

Recently, models of the early human visual system have been developed to design quality metrics for applications such as halftoning and perceptually lossless compression.[6–9] These channelized models measure the similarity between the original and a distorted version of an image. The models are typically based on a multiscale representation of oriented local contrast as first proposed by Peli.[10] Taking into account effects of early vision such as masking, the contrast representations of the original and the distorted image are subtracted and passed through a nonlinear psychometric function to compute the probability of detecting differences at threshold level.

The success of these HVS models in predicting the perceptibility of differences between images suggests that the use of visual system models may significantly improve similarity metric performance. However, the models for quality assessment are not directly applicable to the 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.

The objective of this work is to develop an image distance metric which is systematically optimized for maximum consistency with human perception of the similarity of images of natural scenes. The proposed metric is based on color as well as spatial attributes using features extracted from a multiscale HVS model. This channelized HVS model is derived from the models for image quality assessment. Due to the importance of color in image similarity perception, the model extends achromatic contrast concepts to color-contrast[11] and, in addition, retains color channels for direct color comparison. However, we replace conventional contrast with a power-law contrast model based on the CIE Lab color-space. This contrast model is well suited to the image similarity problem since it inherits the uniform perceptual properties of the Lab system. By subtracting the HVS model representations of the query and target images, we form feature distances. These distances are then combined into a single measure of image similarity using a linear classifier. The classifier weights form the parameters of our model and are estimated to fit image matching data from human subjects to obtain maximum consistency between metric performance and human perception.

Figure 1 shows an overview of the complete system. The input images are first processed by the HVS model which computes separate channels for color, contrast and orientation. In order to reduce background noise, the contrast and orientation channels are quantized. This is followed by a stage called feature-extraction which computes features

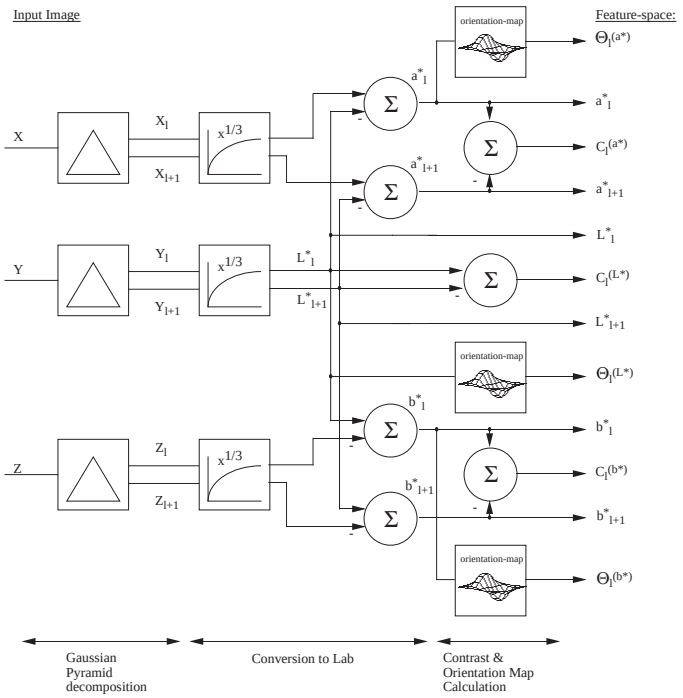


Figure 2: HVS Channel Model. The input image is decomposed into a Gaussian pyramid in XYZ. This pyramid is then converted to CIELab color-space, where contrast and color-contrast channels are obtained as channel differences. The orientation channels are computed using derivatives of the Lab pyramid. For reasons of simplicity, the diagram shows only two pyramid levels.

for the actual image comparison. In the current model, these features are simply the pixel-values themselves as well as block-variances which account for texture behavior. The query and target images are then compared by taking the mean-square error separately for each feature, resulting in feature-distances d_i. These feature-distances are combined in a linear classifier to compute a single measure of similarity. In order to estimate the classifier weights ω_i, we use experimental image matching data from human subjects collected in a simple visual test. In this visual test, subjects are presented with a query image and 209 target images. The subject's task is to find the two images most similar to the query and rate their similarity on a scale from 0 to 10. Using the similarity metric on the image set from the visual test, we define a cost function for the consistency between the metric's ranking of the target images and the subject's choices in the visual test. This cost function is used to select a small number of the most important features and to estimate their weights in the linear classifier.

2 HVS MODEL

The basic structure of the proposed model is illustrated in Fig. 2. We first apply a Gaussian pyramid decomposition[12] to each of the color channels of the input image in CIE XYZ color-space. The pyramid decomposition is computed by successively lowpass filtering the original image with a Gaussian kernel and decimating by two. The result is a multiscale representation of the image where each pyramid level l has the resolution of the original image divided by 2^l. For our experiments with image sizes of approximately 185×280 pixels, we chose the number of pyramid levels to be $L = 5$, resulting in a size of 11×17 pixels at the lowest level. Using images containing radially symmetric sine-waves, we experimentally determined the Gaussian filter kernel to have a sample spacing of $\alpha = 1/\sigma = 0.5$ and a kernel size of 15×15. The pyramid decomposition is followed by conversion of each pyramid level to Lab color-space. 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 contrast and color-contrast representations.

2.1 Contrast Representation

Weber's contrast is defined as

$$C_W = \frac{Y - Y_B}{Y_B} \approx \log Y - \log Y_B \quad \text{for} \quad |C_W| << 1 \tag{1}$$

where Y is the luminance of a single foreground stimulus and Y_B is the uniform background luminance. In addition, the contrast sensitivity using Weber's contrast can be defined as

$$S_W = \frac{1}{C_{WM}} \tag{2}$$

where C_{WM} is the minimum contrast detectable by human subjects. Measurements indicate that the human contrast sensitivity using Weber's contrast decreases at low background luminance levels. This is undesirable, since contrast should ideally be invariant to background luminance. Furthermore, Weber's contrast has the disadvantage, that C_W goes to infinity as Y_B approaches zero.

In order to avoid the disadvantages of Weber's contrast, we employ a different contrast measure called power-law contrast. Power-law contrast is defined as

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

This contrast definition has several desirable properties. As Y_B approaches zero, C goes to $Y^{1/3}$ which is more consistent with human vision. Furthermore, if we approximate (3) by its Taylor series at $Y/Y_B = 1$, we obtain

$$C \approx \frac{1}{3} Y_B^{1/3} C_W \quad \text{for} \quad |C| << 1. \tag{4}$$

Consequently, the contrast sensitivity S_W of Weber's contrast can be expressed as

$$S_W \approx \frac{1}{3} Y_B^{1/3} S \tag{5}$$

where $S = 1/C_M$ is the contrast sensitivity using power-law contrast. This result qualitatively accounts for the dependence of S_W on the background luminance since S_W decreases with decreasing Y_B. Note furthermore, that the definition of the power-law contrast is approximately consistent with the transfer function of the photoreceptors in the non-saturated range,[13] followed by the subtraction performed in the retinal ganglion cells.

In the multiscale pyramid, power-law contrast can be computed as the nonlinear difference between lowpass channels of different resolutions. Let Y_l denote the luminance channel at pyramid level l, where $l = 0$ is the finest and $l = L - 1$ is the coarsest resolution. Due to the spatial averaging of the lowpass 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 $\hat{C}_{l,i}$ at level l as

$$\hat{C}_{l,i}^{(Y)} = Y_l^{1/3} - Y_{l+i}^{1/3} \tag{6}$$

where $1 \leq i < L - l$. Notice that larger values of i average the background luminance over relatively larger areas.

An important advantage of the power-law contrast computation in XYZ is its consistency with the CIEXYZ to CIELab color-space conversion. This consistency allows us to convert the Gaussian pyramid to Lab and compute contrast directly as the difference between channels in L^* as shown in Fig. 2. Since we are only interested in contrast differences, we can define the scaled *contrast-channels* $C_{l,i}^{(Y)}$ as

$$\begin{aligned} C_{l,i}^{(Y)} &= 116(Y_l^{1/3} - Y_{l+i}^{1/3}) \\ &= L_l - L_{l+i}. \end{aligned} \tag{7}$$

The major advantage of this strategy is that we obtain color as well as contrast channels which both have the perceptual uniformity of the Lab color-difference equation when used with an Euclidean metric.

2.2 Color contrast

In analogy to the power-law contrast definition (7), we define the opponent *color-contrast channels* $C_{l,i}^{(a^*)}$ and $C_{l,i}^{(b^*)}$ as

$$C_{l,i}^{(a^*)} = a_l^* - a_{l+i}^* \tag{8}$$

$$C_{l,i}^{(b^*)} = b_l^* - b_{l+i}^*. \tag{9}$$

In order to interpret this color-contrast definition, we can write $C_{l,i}^{(a^*)}$ as

$$\begin{aligned} C_{l,i}^{(a^*)} &= a_l^* - a_{l+i}^* \\ &= 500[(X_l^{1/3} - Y_l^{1/3}) - (X_{l+i}^{1/3} - Y_{l+i}^{1/3})] \\ &= 500[(X_l^{1/3} - X_{l+i}^{1/3}) - (Y_l^{1/3} - Y_{l+i}^{1/3})] \\ &= 500[\hat{C}_{l,i}^{(X)} - \hat{C}_{l,i}^{(Y)}] \end{aligned} \tag{10}$$

where $\hat{C}_{l,i}^{(X)}$ and $\hat{C}_{l,i}^{(Y)}$ are power-law contrasts in X and Y. Therefore, $C_{l,i}^{(a^*)}$ is essentially a difference between 'red' and 'green' contrasts. Similarly, $C_{l,i}^{(b^*)}$ can be written as

$$C_{l,i}^{(b^*)} = 200[\hat{C}_{l,i}^{(Y)} - \hat{C}_{l,i}^{(Z)}] \tag{11}$$

which can be interpreted as a difference between 'yellow' and 'blue' contrasts.

In terms of the human visual system, the above color-contrast definitions are meaningful assuming 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 Lateral Geniculate Nucleus.

2.3 Quantization of contrast features

Quantizing the amplitude of the contrast and color-contrast channels significantly improved the performance of features derived from these channels. The best performance was obtained using only the three quantization levels {-1, 0, 1}. If $T_{C_{l,i}^{(Y)}}$ denotes the quantization threshold for channel $C_{l,i}^{(Y)}$, the quantized contrast channel $\tilde{C}_{l,i}^{(Y)}$ is obtained by

$$\tilde{C}_{l,i}^{(Y)} = \begin{cases} 1 & : \quad C_{l,i}^{(Y)} \geq T_{C_{l,i}^{(Y)}} \\ 0 & : \quad |C_{l,i}^{(Y)}| < T_{C_{l,i}^{(Y)}} \\ -1 & : \quad C_{l,i}^{(Y)} \leq -T_{C_{l,i}^{(Y)}} \end{cases} . \tag{12}$$

In order to obtain the thresholds $T_{C_{l,i}^{(Y)}}$, we compute the channel variances $\sigma^2_{C_{l,i}^{(Y)}}$ over a set of 200 images. The thresholds are then calculated as

$$T_{C_{l,i}^{(Y)}} = 0.3\sqrt{\sigma^2_{C_{l,i}^{(Y)}}} \tag{13}$$

where the constant was determined experimentally to eliminate most of the background noise while preserving the important foreground contours. The color-contrast channels are quantized in a similar manner, using the same constant and the respective channel variances.

2.4 Orientation channels

The objective of orientation channels for the image search problem is to extract the dominant orientation of image contours. In our model, we compute angular maps consisting of edge-angle and edge-amplitude values at every image

location for every level of the Lab pyramid. The angular maps are obtained by convolving the input channel with the horizontal and vertical derivative of a 2D Gaussian and converting the result to polar coordinates. The filter kernels h_x and h_y to compute the horizontal and vertical derivatives are given by

$$\begin{aligned} h_x(m,n) &= \alpha m e^{-((\alpha m)^2+(\alpha n)^2)} \\ h_y(m,n) &= \alpha n e^{-((\alpha m)^2+(\alpha n)^2)} \end{aligned} \tag{14}$$

where α is the sample spacing. In order to obtain kernels which perform some spatial smoothing but are sufficiently small to be used on the lowest pyramid level, we chose $\alpha = 1$. The derivatives $Dx_l^{(L^*)}$ and $Dy_l^{(L^*)}$ of luminance channel L_l^* are then computed as

$$\begin{aligned} Dx_l^{(L^*)} &= h_x ** L_l^* \\ Dy_l^{(L^*)} &= h_y ** L_l^* \end{aligned} \tag{15}$$

where the $**$ operator denotes 2D-convolution. Transforming these derivatives into polar coordinates, we can compute the edge-angle $\vartheta_l^{(L^*)}$ and edge-amplitude $s_l^{(L^*)}$ as

$$\begin{aligned} \vartheta_l^{(L^*)} &= \arg(Dy_l^{(L^*)}, Dx_l^{(L^*)}) \\ s_l^{(L^*)} &= \sqrt{(Dx_l^{(L^*)})^2 + (Dy_l^{(L^*)})^2} \end{aligned} \tag{16}$$

where the arg computes the angle in the full range from $-\pi \leq \vartheta < \pi$.

This method for computing the orientation maps performed considerably better than computing oriented energy using a steerable quadrature filter pair.[14,15] We believe that the difference in performance is due to the loss of contour polarity information in the energy computation of the quadrature filter pair.

Similar to the contrast channels, a quantization of the edge-amplitude values resulted in improved similarity metric performance. Analogous to (12) we obtain the quantized edge-amplitude $\tilde{s}_l^{L^*}(m,n)$ as

$$\tilde{s}_l^{(L^*)} = \begin{cases} 1 & : \quad s_l^{(L^*)} \geq T_{s_l^{(L^*)}} \\ 0 & : \quad s_l^{(L^*)} < T_{s_l^{(L^*)}} \end{cases} . \tag{17}$$

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 as the variance in (13)

$$T_{s_l^{(L^*)}} = 0.7 \mu_{s_l^{(L^*)}}. \tag{18}$$

The quantized luminance orientation map $\tilde{\Theta}_l^{(L^*)}$ is then given by

$$\tilde{\Theta}_l^{(L^*)} = \begin{bmatrix} \vartheta_l^{(L^*)} \\ \tilde{s}_l^{(L^*)} \end{bmatrix}. \tag{19}$$

The equations for the quantized color orientation maps $\tilde{\Theta}_l^{(a^*)}$ and $\tilde{\Theta}_l^{(b^*)}$ are analogous. In the remainder we will refer to the quantized orientation maps $\tilde{\Theta}_l^{(\bullet)}$ as *orientation channels*.

In conclusion, we developed and implemented a multiscale channel model which includes color, contrast, color-contrast and orientation channels. In particular we proposed a power-law contrast computation based on the uniform Lab color-space. Finally, it appears to be important to retain edge polarity information in the orientation maps. Table 1 shows a list of the computed channels.

3 FEATURE EXTRACTION AND METRIC COMPUTATION

In order to compare the query and target image, we extract features from the HVS channel representation and compute feature distances by taking the mean-squared error. In the case of the color and contrast channels we use

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 as a function of type, level l and level difference i. The last column contains the number of channels computed for the respective type.

the channels themselves as a first feature. In particular, let $\mathcal{C}_{l,Q}(m,n)$ and $\mathcal{C}_{l,T}(m,n)$ denote any color or contrast channel of the query image Q and the target image T. We then compute the feature-distance d_μ corresponding to this channel as

$$d_\mu = \frac{1}{M_l N_l} \sum_{m=1}^{M_l} \sum_{n=1}^{N_l} (\mathcal{C}_{l,Q}(m,n) - \mathcal{C}_{l,T}(m,n))^2 \tag{20}$$

where M_l and N_l denote the channel size.

In addition, we divide the color and contrast channels into B_l rectangular blocks of size $16/2^l$ square pixels and compute block variances as a second feature to account for textural behavior. The corresponding feature-distances d_{σ^2} are obtained as the mean squared error of the block-variances $\sigma^2_{\mathcal{C}_l,Q}$ and $\sigma^2_{\mathcal{C}_l,T}$

$$d_{\sigma^2} = \sqrt{\frac{1}{B_l} \sum_{b=1}^{B_l} (\sigma^2_{\mathcal{C}_l,Q}(b) - \sigma^2_{\mathcal{C}_l,T}(b))^2}. \tag{21}$$

The features for the orientation channels are the channels themselves, however, the distance computation has to combine edge-angle and edge amplitude into a single distance. This is not straightforward, since it is unclear how the angular difference should be weighted when the edge-amplitude in one image is quantized to zero but not in the other. An asymmetric approach is motivated by Jacobs, Finkelstein and Salesin,[2] 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 a comparison by computing the distance $d_{\tilde{\Theta}}$ between the quantized orientation maps $\tilde{\Theta}_Q(m,n)$ and $\tilde{\Theta}_T(m,n)$ as

$$\begin{aligned}
\Delta\vartheta(m,n) &= (\vartheta_Q(m,n) - \vartheta_T(m,n)) \bmod 2\pi \\
\Delta^*\vartheta(m,n) &= \begin{cases} \Delta\vartheta(m,n) & : \ \tilde{s}_Q(m,n) = 1\,,\ \tilde{s}_T(m,n) = 1 \\ \pi & : \ \tilde{s}_Q(m,n) = 1\,,\ \tilde{s}_T(m,n) = 0 \\ 0 & : \ \tilde{s}_Q(m,n) = 0 \end{cases} \\
d_{\tilde{\Theta}} &= \frac{1}{M_l N_l} \sum_{m=1}^{M_l} \sum_{n=1}^{N_l} [\Delta^*\vartheta(m,n)]^2.
\end{aligned} \tag{22}$$

The computation of d_μ, d_{σ^2} and $d_{\tilde{\Theta}}$ for the respective channels at all resolutions yields a total of 102 feature-distances as listed in Table 2. The final image similarity metric D is calculated as a linear combination of these distances. If we refer to the enumerated distances as d_i then

$$D = \sum_{i=1}^{n} \omega_i d_i \tag{23}$$

Channels	levels	distances	no. distances
L_l^*, a_l^*, b_l^*	$l = \{0, 1, 2, 3\}$	d_μ, d_{σ^2}	2×12
L_l^*, a_l^*, b_l^*	$l = 4$	d_μ	3
$\tilde{C}_{l,i}^{(Y)}$, $\tilde{C}_{l,i}^{(a^*)}$, $\tilde{C}_{l,i}^{(b^*)}$	all (l, i)	d_μ, d_{σ^2}	2×30
$\tilde{\Theta}_l^{(L^*)}$, $\tilde{\Theta}_l^{(a^*)}$, $\tilde{\Theta}_l^{(b^*)}$	all l	$d_{\tilde{\Theta}}$	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 last column lists the total number of distances computed for the channels in the row.

where the weights ω_i are the parameters of our model. In the following we will concentrate on estimating the ω_i based on human perceptual judgments.

4 WEIGHT ESTIMATION BASED ON SUBJECT DATA

In order to estimate the classifier weights in a manner that maximizes the consistency of the metric distance with human perceptual judgments, we perform a visual test to collect experimental image matching data. We then relate the metric performance to the experimental matching data and optimize the classifier weights for maximum consistency.

4.1 Visual tests

The experimental matching data is obtained by presenting a subject with a single query image and 209 target images which are randomly selected from a database of 3000 images. 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 target images which are most similar to the query image. In addition, the subject rates the similarity between the selected images and the query on a scale from zero to ten. We will denote these similarity ratings by S_1 and S_2 for the first and second image selected. If none or only one image is considered to be similar to the query, the subject can leave the corresponding answer fields blank. We have performed 200 such tests on one of the authors (TF) and 80 tests on a subject (CCT) who is familiar with image processing but not involved in database search.

4.2 Weight estimation

In order to relate the metric performance to the experimental matching data, we define a cost function which accounts for the consistency between the metric results and the subject's choices. In particular, we compute the metric distance D between the query and each target image in the visual test. We then order the target images with respect to increasing D to obtain a ranking from most to least similar to the query. Specifically, let R_1 and R_2 denote the metric's rankings of the two images selected by the subject in the visual test. In order to maximize the consistency between the rankings of the metric and the subject, we wish to minimize R_1 and R_2. We therefore define the cost for the metric's ranking as

$$\text{cost} = S_1 \log(R_1) + S_2 \log(R_2) \tag{24}$$

where S_1 and S_2 are the subject's similarity ratings from the visual test. We choose the cost to be a logarithmic function of R since only metric ranks among the first few matches are valuable to a user searching a large database. In order to obtain a global cost function C which incorporates the results of a set of T visual tests, we sum over the

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 features. The table shows the features selected for the training set of 74 visual tests. The features are listed in the order of decreasing weights ω_i.

costs for each test t

$$C = \sum_{t=1}^{T} (S_1(t) \log R_1(t) + S_2(t) \log R_2(t)). \tag{25}$$

In the following this cost function is used to select a subset of best features and estimate their weights ω_i.

The high dimensionality of the feature space extracted from the HVS model necessitates the selection of a small set of best features. We form such a set by sequentially selecting features in the order that leads to the most rapid improvement in the cost function. After selecting the feature subset, we use simulated annealing[16] to optimize the feature weights. For a detailed discussion of the feature selection and weight optimization the reader is referred to [17] .

5 RESULTS

The results presented are based on a selection of 13 out of 102 features trained on a set of 80 visual tests performed by TF.

5.1 Selected features

Table 3 shows the selected features in the order of decreasing weights. Since we normalized the distances d to have unit variance, the values of the weights are consistent between channels. Out of the 13 selected features, there are 4 luminance/color, 4 contrast/color-contrast and 5 orientation channel features. 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.

For the color-channels, all selected features are at low resolution and of distance type d_μ. The importance of low resolution color features is consistent with vision science and re-emphasizes that the color space be uniform at low spatial frequencies. For the contrast channels, all selected features are block variances at high resolutions. This indicates that the contrast channels predominantly account for textural behavior. 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.

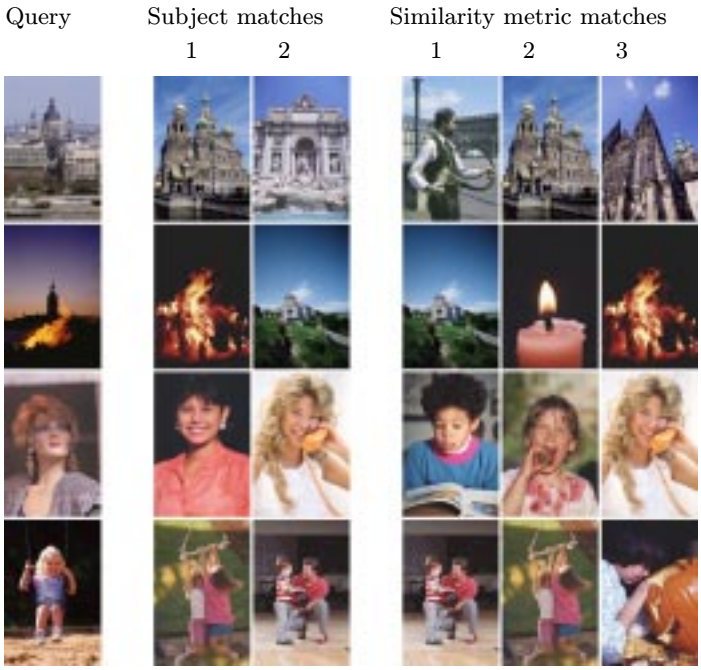


Figure 3: Matching results on the test set. Each row corresponds to a different visual test where the image in the first column is the query and the two images in columns 2 and 3 are the matches selected by the subject. The columns on the right show the images selected by the metric from rank 1 to 3.

5.2 Metric Performance

For each of 74 different query images that were not included in the training set, we used the metric to generate a ranked list of 209 matches. We then compared these matches to the matches obtained via visual tests performed with the same sets of query and target images by TF. Figure 3 shows a selection of the matching results. Each row

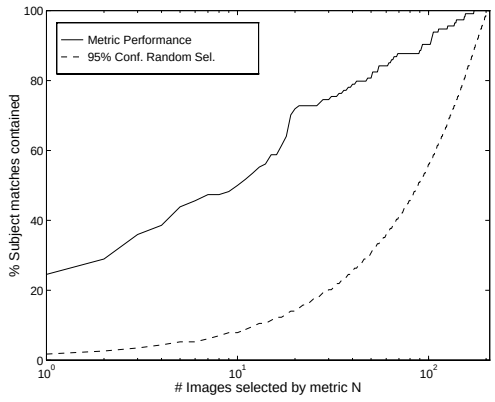


Figure 4: Metric performance on test set and 95% confidence interval for random selection of target images.

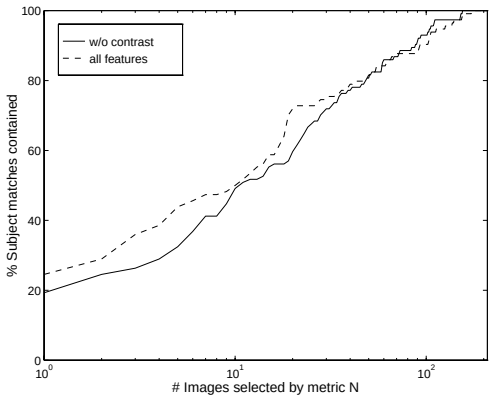


Figure 5: Performance analysis of contrast channels. The curves indicate that for rankings between 1 and 30 the contrast features resulted in a considerable improvement in accuracy.

of the figure contains a query image followed by the two images selected by the subject and the three 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 such as color or shape. In contrast to the color-histogram methods, the metric matches images containing similar shapes and textures but different colors.

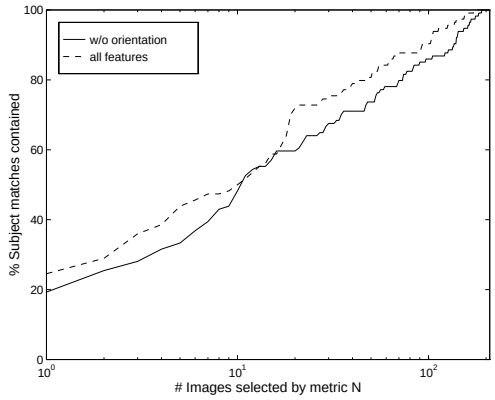


Figure 6: Performance analysis of orientation channels. By using the orientation channel distances, the metric's performance

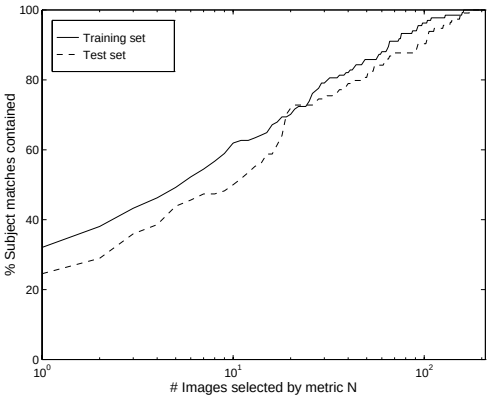


Figure 7: Comparison of training and test set performance. The graphs indicate, that the accuracies on the trained and the untrained sets differ by not more than 10-15% .

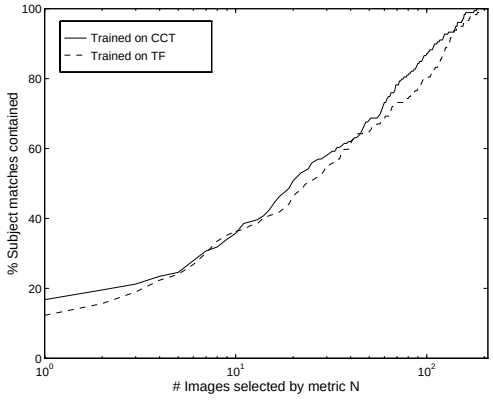


Figure 8: Crosstraining effects between subjects. Both curves show metric performance evaluated on a set of visual tests performed by subject CCT. The solid curve shows training performance for CCT. For the dashed curve, the metric was trained on the same set of visual tests performed by TF.

In order to perform a more systematic analysis, we plotted the classification accuracy of the metric as shown in Figs. 4 to 8. These figures show the probability that an image selected by the human subject is included among the top ranked N images chosen by the metric. The value of N is shown on the x-axis which is drawn in logarithmic scale since we are mainly interested in the classification accuracy obtained within the first few matches selected by the metric. Figure 4 compares the metric performance to an upper bound for the result of a random selection of target images (95% confidence interval).

In order to investigate the importance of the contrast channels, we trained the metric excluding these channels. Figure 5 shows a comparison to the performance obtained by allowing the selection of all channels. In the range of interest between rank 1 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 and color-contrast channels, these channels result in a significant increase in performance which cannot be obtained using only color and orientation channels.

A similar analysis for the orientation channels is shown in Fig. 6. 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.

Figure 7 shows the metric performance for the test and training set of images. The solid curve shows the performance on the matches from the training set of visual tests. The dashed curve evaluates the metric on the test set of untrained images as shown in the previous plots. Throughout the range of ranks, the classification accuracy differs by not more than 10 to 15% which indicates that the model is not overparameterized. However, more training data would be desirable in order to capture the similarity perception of different subjects and reduce the performance difference between training and test data.

Finally, we examine the consistency of the metric's training for different subjects performing the visual test. Figure 8 shows the metric performance evaluated on a set of visual tests performed by subject CCT. The solid curve shows the training performance for CCT whereas for the dashed curve, the metric was trained on the same set of

visual tests performed by TF. While it is generally harder to predict the subject matches selected by CCT, the crosstraining effect is small.

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 optimization strategy is independent of the underlying image representation and therefore well 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 comprehensively 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 similarity metrics which have substantial value in predicting image similarity as perceived by human subjects.

ACKNOWLEDGMENT

This work was supported by Apple Computer Inc.

7 REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] B. Scassellati, S. Alexopoulos, and M. Flickner, "Retrieving images by 2D shape: a comparison of computation methods with human perceptual judgments," *Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases*, vol. 2185, February 7-8 1994, San Jose, CA, pp. 2 – 14.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.*

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

[13] R. M. Boynton, *Human Color Vision.* Optical Society of America, 1992.

[14] 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.

[15] 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.

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

[17] T. Frese, C. A. Bouman, and J. P. Allebach, "A methodology for designing image similarity metrics based on human visual system models," Tech. Rep. TR-ECE 97-2, Purdue University, West Lafayette, IN, 1997.