PERCEPTUAL CONSISTENCY FOR IMAGE RETRIEVAL

WEE KHENG LEOW
Department of Computer Science, National University of Singapore,
3 Science Drive 2, Singapore 117543, Singapore.
leowwk@comp.nus.edu.sg

An ideal image retrieval system should retrieve images that satisfy the user's need, and should, therefore, measure image similarity in a manner consistent with human's perception. Unfortunately, perceptual consistency is very difficult to achieve, even for simple features such as color and texture. This paper summarizes current results of perceptual consistency and suggests possible future work in this direction. Striving for perceptual consistency should be a goal of the next-generation multimedia retrieval systems.

1 Introduction

An ideal image retrieval system should retrieve images that satisfy the user's need. It should, therefore, measure image similarity in a manner consistent with human's perception. Unfortunately, this goal turns out to be very difficult to achieve. This problem leads to retrieval results that do not always meet the users' expectations.[38] Existing systems often make use of relevance feedback techniques to improve the quality of the retrieved results.[26,32,40] However, very few users are willing to go through endless iterations of feedback in hope of retrieving the best results. Moreover, previous feedback results are typically not retained in the system and each new query always begins in an unrefined state. A user has to go through the feedback process even if the same feedback information has been given in the past.

Striving for perceptual consistency should be the goal of a good image retrieval system. At present, progress has been made only for simple features such as color and texture. This article summarizes current results of perceptual consistency and suggests possible future work in this direction.

2 Overview of Perceptual Consistency

There are many ways of defining perceptual consistency. This section discusses some common definitions. Let p_{ij} denote the perceptual distance between samples i and j, and d_{ij} denote the corresponding measured or computational distance. A simple notion of perceptual consistency is that d_{ij} is *proportional* to p_{ij}. That is, there exists a linear function f such that

$$p_{ij} = f(d_{ij}), \quad \forall i, j \ . \tag{1}$$

Then, perceptual consistency can be measured in terms of the mean squared error (MSE) e of linear regression:

$$e = \frac{1}{N} \sum_{i,j} \left(p_{ij} - f(d_{ij})\right)^2 \tag{2}$$

where N is the number of sample pairs. The smaller the MSE, the better is the consistency. A perfect consistency has an MSE of 0.

A less stringent notion of perceptual consistency is to require that f be a monotonic function which can be nonlinear. The problem with this definition is that it is difficult to determine the best nonlinear function to use in practice.

An alternative definition is to require that d_{ij} be *statistically correlated* to p_{ij}. In this case, it is useful to transform the populations $\{d_{ij}\}$ and $\{p_{ij}\}$ to equivalent zero-mean unit-variance populations $\{d'_{ij}\}$ and $\{p'_{ij}\}$:

$$p'_{ij} = \frac{p_{ij} - \bar{p}}{\sigma_p}, \quad d'_{ij} = \frac{d_{ij} - \bar{d}}{\sigma_d} \tag{3}$$

where $\bar{p}$ and $\bar{d}$ are the means and σ_p and σ_d are the standard deviations of the populations. Then, perceptual consistency can be measured in terms of the correlation r:

$$r = \sum_{i,j} p'_{ij}\, d'_{ij} \ . \tag{4}$$

Substituting Eq. 3 into Eq. 4 yields the Pearson's correlation coefficient:

$$r = \frac{\sum_{i,j} (p_{ij} - \bar{p})(d_{ij} - \bar{d})}{\left[\sum_{i,j} (p_{ij} - \bar{p})^2 \sum_{i,j} (d_{ij} - \bar{d})^2\right]^{1/2}} \ . \tag{5}$$

The coefficient r ranges from -1 to $+1$.

With perfect consistency ($e = 0$ or $r = 1$), we obtain the following condition:

$$d_{ij} \le d_{kl} \Rightarrow p_{ij} \le p_{kl} \quad \text{for any samples } i, j, k, l. \tag{6}$$

That is, if perfect consistency is achieved, computational similarity would imply perceptual similarity.

3 Color

3.1 Color Spaces and Color Differences

Various color spaces have been used in image retrieval. The more commonly used spaces include HSV, CIELUV, and CIELAB. The HSV space consists of hue, saturation, and value dimensions. It is used in VisualSeek [43] and PicHunter,[9] and by Vailaya et al.[50] CIELUV and CIELAB are color spaces developed by the International Commission on Illumination (Commission International de l'Éclairage, CIE). They consist of a luminance dimension L^* and two chromatic dimensions namely u^*, v^* and a^*, b^*. Among these three spaces, CIELUV and CIELAB are more perceptually uniform than HSV.[4] CIELUV is used in ImageRover [42] and by Mehtre et al.[30] while CIELAB is used in Quicklook.[8]

In recent years, there is also a move to standardize the conversion formula between RGB and various CIE spaces. This effort gives rise to the so-called *sRGB*, which is a proposed standard or default RGB color space for the internet.[1,46] It

captures the averaged characteristics of most computer monitors. With sRGB, there is now a unique formula for converting to and from CIE color values.

The difference between two colors is typically measured as the Euclidean distance in the target color space. Several improvements over the CIELAB Euclidean color difference equation have been proposed, including CIE94, CMC, and BFD.[4] Recent psychological tests show that these color difference equations are more perceptually uniform than Euclidean distance in the CIELAB and CIELUV spaces.[4,14,18,31,45] In particular, CIE94 has a simpler form, which is a weighted Euclidean distance:[4]

$$\Delta E^*_{94} = \left[\left(\frac{\Delta L^*}{k_L S_L} \right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C} \right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H} \right)^2 \right]^{1/2} \tag{7}$$

where ΔL^*, ΔC^*_{ab}, and ΔH^*_{ab} are the differences in lightness, chroma, and hue, $S_L = 1$, $S_C = 1 + 0.045\,\bar{C}^*_{ab}$, $S_H = 1 + 0.015\,\bar{C}^*_{ab}$, and $k_L = k_C = k_H = 1$ for reference conditions. The variable $\bar{C}^*_{ab}$ is the geometric mean between the chroma values of the two colors, i.e., $\bar{C}^*_{ab} = \sqrt{C^*_{ab,1} C^*_{ab,2}}$.

In addition to these colors spaces, the modified Munsell HVC space, which consists of hue, value, and chromaticity dimensions, and is used in QBIC [33] and by Gong et al.[13] It is perceptually quite uniform, but is less commonly used than CIELAB. Gong et al. uses the Godlove equation [12] to measure color difference. It was derived by Godlove to improve the perceptual uniformity of color difference measured in the Munsell space. Recent psychological studies show that CIE94 is more accurate in measuring human color perception than the modified Judd and Adams-Nickerson formulae,[18] which are similar to the Godlove equation.

3.2 Color Histograms and Dissimilarity

An image or image region typically contains more than one color. Therefore, color histograms are used to represent the distributions of colors in images. There are two general approaches to generating color histograms from images: *fixed binning* and *adaptive binning*. The fixed binning approach induces histogram bins by partitioning the color space into fixed color bins. Once the bins are derived, they are fixed and the same binning is applied to all images. On the other hand, adaptive binning adapts the bins to the actual distributions of the images. As a result, different binnings are induced for different images.

There are two types of fixed binning schemes: *regular partitioning* and *clustering*. The first method simply partitions the axes of a target color space into regular intervals, thus producing rectangular bins.[9,42,43] The second method partitions a color space into a large number of rectangular cells, which are then clustered by a clustering algorithm, such as k-means, into a smaller number of bins.[8,15,50]

Adaptive binning is similar to color space clustering in that k-means clustering or its variants are used to induce the bins.[20,37] However, the clustering algorithm is applied to the colors in an image instead of the colors in an entire color space. Therefore, adaptive binning produces different binning schemes for different images.

Experimental results show that adaptive-binning histograms can represent color distributions more accurately than can fixed-binning histograms and yet use fewer

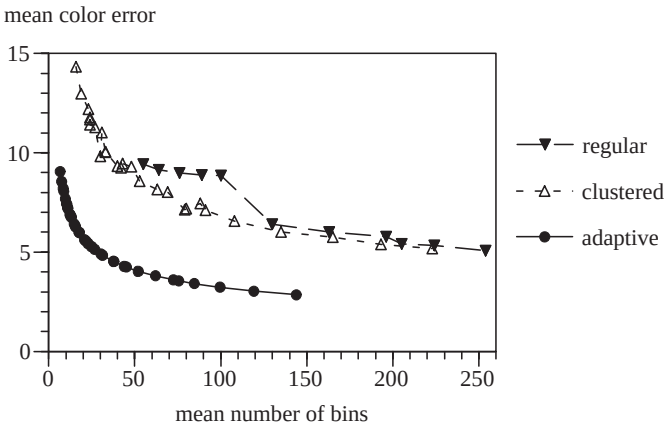

Figure 1. Comparison of mean color errors of regular, clustered, and adaptive histograms.

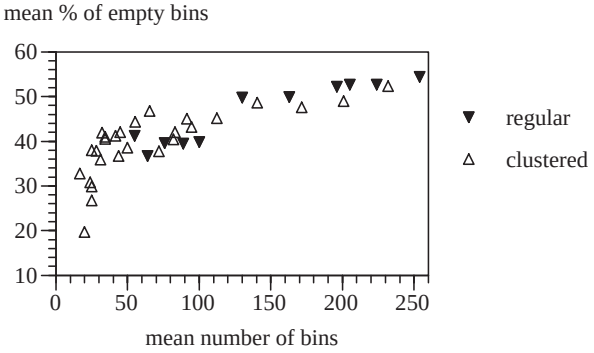

Figure 2. Average percentage of empty bins in regular and clustered histograms. Adaptive histograms have no empty bins.

bins and no empty bins [20] (Fig. 1, 2). In particular, adaptive histograms can achieve a mean color error below the human color acceptability threshold of 4.5,[45] which is a threshold below which two colors are regarded as practically identical. Note that the acceptability threshold is slightly higher than the perceptibility threshold of 2.2,[45] which is the threshold below which two colors are perceptually indistinguishable.

Although color difference measured using CIE94 in CIELAB color space is perceptually consistent, the difference between color histograms measured by various dissimilarity measures have not been shown to be perceptually consistent. Empirical tests performed by Puzicha et al.[34] and Leow and Li [20] confirmed that the Euclidean distance between color histograms is not as reliable as other measures are in computing dissimilarity. In particular, the results of Puzicha et al. show that dissimilarities such as χ^2, Kullback-Leibler divergence, and Jessen difference

divergence[a] (JD) performed better than other measures do for large sample size (i.e., number of pixels sampled in an image), while Earth Mover's Distance (EMD), Kolmogorov-Smirnov, and Cramer/von Mises performed better for small sample size.

The study of Leow and Li show that JD is most reliable for image retrieval. JD measures the difference between two histograms G and H, with bin counts g_i and h_i, as follows:

$$d(G,H) = \sum_i \left(g_i \log \frac{g_i}{m_i} + h_i \log \frac{h_i}{m_i} \right) \tag{8}$$

where $m_i = (g_i + h_i)/2$. Although JD is reliable, it can be applied only on fixed-binning histograms. On the other hand, the weighted correlation dissimilarity [20] (WC) can be applied to adaptive histograms.

An adaptive histogram $H = (n, \mathcal{C}, \mathcal{H})$ is a 3-tuple consisting of a set $\mathcal{C}$ of n bins $\mathbf{c}_i$, $i = 1, \ldots, n$, and a set $\mathcal{H}$ of corresponding bin counts $h_i \geq 0$. The similarity $w(\mathbf{b}, \mathbf{c})$ between bins $\mathbf{b}$ and $\mathbf{c}$ is given by a monotonic function inversely related to the distance $d(\mathbf{b}, \mathbf{c})$ between them. For color histograms, the weight $w(\mathbf{b}, \mathbf{c})$ can be defined in terms of the volume of intersection V_s between the bins:

$$w(\mathbf{b}, \mathbf{c}) = w(\alpha) = \frac{V_s}{V} = \begin{cases} 1 - \frac{3}{4}\alpha + \frac{1}{16}\alpha^3 \text{ if } 0 \leq \alpha \leq 2 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{9}$$

where αR is the distance between the bins and R is the radius of a bin.

The weighted correlation $G \cdot H$ between histograms $G = (m, \{\mathbf{b}_i\}, \{g_i\})$ and $H = (n, \{\mathbf{c}_i\}, \{h_i\})$ is defined as follows:

$$G \cdot H = \sum_{i=1}^{m} \sum_{j=1}^{n} w(\mathbf{b}_i, \mathbf{c}_j)\, g_i h_j \ . \tag{10}$$

For a histogram H, its norm $\|H\| = \sqrt{H \cdot H}$, and its normalized form $\overline{H} = H/\|H\|$. The similarity $s(G,H)$ between histograms G and H is $s(G,H) = \overline{G} \cdot \overline{H}$, and the dissimilarity $d(G,H) = 1 - s(G,H)$.

The retrieval performance of WC dissimilarity is comparable to that of JD (Fig. 3). Unlike EMD, which is also applicable to adaptive histograms, WC does not require an optimization process. It is, thus, more efficient to compute than EMD.

4 Texture

4.1 Texture Features and Dissimilarity

Commonly used texture features can be divided into two main categories: statistical and spectral. Statistical features characterize textures in terms of local statistical measures (such as coarseness, directionality, contrast [47]), simultaneous

[a] The formula that Puzicha et al.[34] called "Jeffreys divergence" is more commonly known as "Jessen difference divergence" in Information Theory literature.[6,7,48]

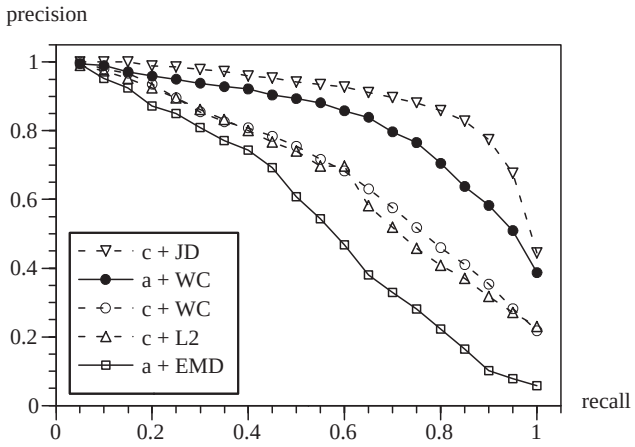

Figure 3. Precision-recall curves of various combinations of binning methods (c: clustered, dashed line; a: adaptive, solid line) and dissimilarities (JD: Jessen difference divergence, WC: weighted correlation, L2: Euclidean, EMD: Earth Mover's Distance).

autoregressive model[29] (MRSAR), or Markov random field[10] (MRF). In general, these features are good at modeling random patterns such as sand and pebbles, but not suitable for modeling structured patterns such as bricks and roof tile.[21] Among them, the statistical features of Tamura et al.[47] are used in QBIC,[33] and MRSAR is used in PhotoBook.[21]

The spectral approach is based on the response of a set of band-pass filters, typically 2D Gabor and wavelet filters.[5] Each filter responds most strongly to the patterns at a specific spatial-frequency and orientation band. These features have been used in NeTra,[27,28] VisualSeek,[43] etc. In addition, features derived directly from Discrete Fourier Transform (DFT) has also been used, for instance, in PhotoBook.[21]

Various dissimilarity measures have been defined for computational texture features, including Euclidean and scaled Euclidean distance,[41], Mahalanobis distance,[21] and weighted mean-variance,[27,28] most of which are variations of the weighted Euclidean distance. As expected, these dissimilarity measures are not perceptually consistent (see next Section for details).

An interesting exception is Santini and Jain's *Fuzzy Features Contrast* model (FFC).[41] FFC is based on Tversky feature contrast model[49] which can account for various peculiarities of human's perceptual similarity. Santini and Jain applied FFC to measure similarity of Gabor texture features, and obtained encouraging results.

4.2 Perceptual Texture Models

The earliest study of human's perception of texture similarity was conducted by Tamura et al.[47] In their experiments, 48 human subjects were asked to judge the similarity of texture pairs according to six visual properties, namely, coarseness, contrast, directionality, line-likeness, regularity, and roughness. Similarity judg-

ments were measured and each texture was assigned a perceptual rating value along each of the six visual scales. Due to the combinatorial nature of the task, only 16 textures were used. Amadasun and King[2] and Benke et al.[3] conducted similar ranking experiments to measure similarity judgments according to various visual properties including some of the features of Tamura et al. as well as busyness, complexity, bloblikeness, and texture strength.

The major difficulty with these studies is that the subjects were asked to judge texture similarity according to subjective visual properties. Unfortunately, the subjects' interpretations of the meaning of these visual properties are expected to vary from one person to the next. Therefore, it is uncertain whether the individual ranking results can be combined into group ranking results that represent the perception of a typical person. The second difficulty is that the ranking results were measured according to individual visual properties. But, the relative scale between two visual properties is unknown. For example, one unit difference in coarseness may not be perceptually equal to one unit difference in regularity. So, the different visual dimensions cannot be easily combined to form a perceptual texture space.

To avoid these difficulties, Rao and Lohse[35] performed an experiment in which 20 subjects were asked to sort 30 textures into as many groups as the subjects wished such that the textures in each group were perceptually similar. The textures were sorted based on the subjects' perception of overall texture similarity without using subjective visual properties. A co-occurrence matrix of the sorting results was computed and multidimensional scaling[16] was performed to derive a 3D perceptual space. The experiment was repeated in another study using 56 textures.[36] Rao and Lohse concluded that the 3 dimensions of the space strongly correlate with the visual properties of repetitiveness, orientation, and complexity.

Heaps and Handel[17] conducted further studies using the same methodology. However, they arrived at different conclusions than those of Rao and Lohse. They concluded that it is not possible to reliably associate a visual property to each dimension of the texture space. In addition, perception of texture similarity depends on the context in which the similarity is judged. That is, how similar two textures appear to humans depends not only on the two textures being judged, but also on the whole set of textures with which pairwise judgments are made.

Long and Leow[24] applied a similar approach to develop a perceptual texture space. However, they do no attempt to assign visual properties to the dimensions of the space. In addition, the influence of the context problem is reduced by normalizing the intensity, contrast, scale, and orientation of the textures used in the psychological experiment. In measuring perceptual distance, both the co-occurrence matrix and the information measurement of Donderi[11] were used.

A comparison of the above perceptual texture spaces show that they are very consistent with each other (Table 1). Heaps and Handel reported a good correlation ($r = 0.790$) with Rao and Lohse's data.[17] The perceptual space of Long and Leow constructed using co-occurrence has a better correlation with Rao and Lohse's space compared to that using Donderi's information measurement. This is expected because Rao and Lohse's space was developed using co-occurrence as well. Table 1 shows that the spaces are mutually consistent, thus establishing the perceptual texture space as a reliable measurement of human's perception of texture similarity.

Table 1. Comparison of various perceptual texture spaces with that of Rao & Lohse. Pearson's correlation coefficients show that the spaces are consistent with each others.

perceptual space	3D	4D	5D
Heaps & Handel	0.790	–	–
Long & Leow (co-occurrence)	0.722	0.732	0.713
Long & Leow (info. measure)	0.726	0.694	0.695

Table 2. Assessment of computational texture dissimilarity measures. r = Pearson's correlation coefficient; e = mean squared error.

feature	distance	r	e
Tamura	Euclidean	0.251	0.132
Gabor	Euclidean	0.273	0.131
Gabor	scaled Euclidean	0.282	0.121
Gabor	FFC	0.430	0.098
MRSAR	Euclidean	0.144	0.139
MRSAR	Mahalanobis	0.061	0.152

4.3 Mapping Computational Features

Perceptual consistency of computational dissimilarity measures can be assessed by comparing them with the distances measured in the perceptual space. The following features are considered: Tamura's features, Gabor, and MRSAR. For all the features, Euclidean distance is used to provide baseline results. In addition, Gabor is also paired with FFC (following Santini and Jain [41]) and MRSAR is also paired with Mahalanobis distance (following Liu and Picard [21]).

Table 2 summarizes the results of comparing the computational distances to the distances measured in the 4-D perceptual texture space of Long and Leow.[24] Gabor feature and Gabor with FFC are most consistent with the perceptual space. In particular, measuring Gabor similarity with FFC does improve Gabor feature's perceptual consistency. Measuring MRSAR similarity with Euclidean distance is perceptually more consistent than measuring with Mahalanobis distance. The degrees of consistency of computational features ($r \leq 0.43$) are, however, not very high compared to those between various perceptual spaces (Table 1, $r \approx 0.7$). Therefore, it can be concluded that these computational features and similarity measures are not consistent with human's perception.

The deficiency of computational dissimilarity measures can be mitigated by mapping texture features into the perceptual texture space and then measuring texture dissimilarity in the perceptual space. Long and Leow explored the application of neural networks and support vector machines (SVMs) for the mapping task,[22,24,25] and five test cases were examined:

1. I_e: test with new instances not in the training set, canonical scale and orientation

Table 3. Mean squared errors of texture mapping tests under various conditions. The first three rows are the results of mapping various features by SVM. The last three rows are the results of mapping Gabor features. The last two columns show the perceptual consistency of mapping texture features by SVM under the T_c condition: r = Pearson's correlation coefficient; e = mean squared error.

	mapping tests					consistency	
	I_c	T_c	I_v	T_v	R	e	r
Tamura	0.144	0.256	—	—	—	0.020	0.857
MRSAR	0.096	0.244	—	—	—	0.019	0.859
SVM	0.0052	0.216	0.244	0.245	0.240	0.019	0.859
NN	0.0099	0.238	0.0074	0.148	0.016	—	—
NN+SVM	0.0065	0.226	0.0061	0.143	0.012	—	—

2. T_c: test with new texture types not in the training set, canonical scale and orientation

3. I_v: test with new instances not in the training set, variable scale and orientation

4. T_v: test with new texture types not in the training set, variable scale and orientation

5. R: test with randomly selected samples not in the training set

Table 3 summarizes the testing results. Tamura features and MRSAR were tested only for the cases of canonical scale and orientation because it is unknown how to perform scale- and orientation-invariant mapping of these features. As expected, for all the features, testing errors for new instances are smaller than those for new texture types. Moreover, being most consistent with the perceptual space (Table 2), Gabor features can be mapped to the perceptual space more accurately than other features.

For the cases of canonical scale and orientation (I_c, T_c), SVM can map Gabor features more accurately than other texture features to the perceptual texture space. The hybrid system (NN+SVM) is composed of a convolutional neural network, for performing invariant mapping, and four SVMs, for performing perceptual mapping to the four dimensions of the perceptual space.[25] The hybrid system performs better than pure neural network but marginally poorer than SVM. This result is expected since pure SVM regression takes the original Gabor features as the inputs. On the other hand, the SVMs of the hybrid system take the outputs of the convolutional network as the inputs, and inevitably, some information is loss by network processing.

For the cases of variable scale and orientation (I_v, T_v, R), the hybrid system performs much better than pure SVM because the hybrid system performs invariant mapping whereas pure SVM does not. It's performance is also better than that of pure neural network. As a whole, the integration of the convolutional neural network and SVM produces better overall mapping accuracy than individual neural network and individual SVM.

After mapping computational features to perceptual space, one would expect the mapped coordinates to be more perceptually consistent. An evaluation of the computational features mapped by SVM is performed for the test case of new texture types under canonical scale and orientation T_c. The distance correlation results are shown in the last two columns of Table 3. Comparing Table 3 with Table 2 shows that mapping computational features to perceptual space does improve the perceptual consistency of the features. In summary, it can be concluded that accurate mapping to the perceptual space can be achieved, at least for Gabor features.

4.4 Incremental Perceptual Space

To improve retrieval performance, relevance feedback technique is often used to tune computational similarity measures.[9,26,32,39,40,44] Typically, each new query resets the similarity measure back to its initial state, which is not perceptually consistent. Subsequent feedback for the query is used to adjust the weighting factors of the similarity measure to improve retrieval performance.

The main difficulty with this method is that very few users are willing to go through endless iterations of feedback in hope of retrieving the best results. A successful relevance feedback process must yield positive results within three or four iterations.[19] So, feedback methods that require many iterations to improve retrieval performance are not practically useful. Another shortcoming of this method is that previous feedback results are typically not retained in the system. Each new query starts with a similarity measure that is not perceptually consistent. The users have to go through the relevance feedback process even if the same feedback information has been given in the past. This problem is partially alleviated with user profiling.

A direct method of improving perceptual consistency is to construct a perceptual space of images using psychological experiments (such as the methods discussed in Section 4.2). The Euclidean distances measured in this perceptual space would be consistent with human's judgments. Then, images can be mapped to the perceptual space and retrieval performed in the perceptual space would yield results that are consistent with human's judgments.

This direct approach is appropriate if the construction of the perceptual space involves a small data set, such as the 100 or so images in the Brodatz album. For general image retrieval applications, it is not feasible to construct a perceptual space using thousands of images because it is practically impossible to conduct psychological experiments involving such a large number of images.

Long and Leow presented a method of incrementally measuring perceptual distances and constructing perceptual space based on relevance feedback.[23] Only a small number of relevant judgments is required in each feedback iteration. Feedback results from multiple queries are accumulated and incrementally update the measurements of perceptual distances between images. If the feedback results are provided by the same user, then the perceptual distances measured would be consistent with a single user's perception. Otherwise, the measurements would reflect the average perception of typical users. In the case of a single user, the measurements would eventually stabilize if the user's relevant judgment remains consistent over time. Otherwise, the measurements would adapt to the changes in the user's

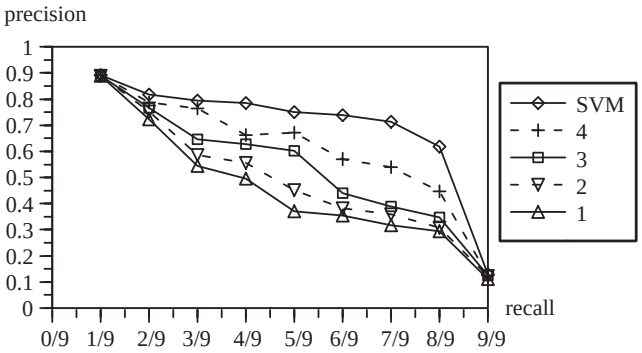

Figure 4. Precision-recall curves of the incremental space at stages 1 to 4. As information coverage increases, the precision-recall curve of the space shifts towards the upper bound achieved by direct mapping to PTS using SVM.

relevant judgment.

Figure 4 plots the results of testing the incremental space construction method at various incremental stages. The precision-recall curve of the SVM model corresponds to the case that all the texture images in the database have been mapped to the perceptual space. This is the condition of 100% information coverage and marks the best performance achievable by the incremental method. The results show that the incremental method indeed improves retrieval performance over time.

Figure 5 plots the perceptual consistency of the constructed space at various percentage of information coverage. In addition, the operating points of four computational texture models are marked in the figure according to their perceptual consistency. It can be seen that, below 5% information coverage, the incremental space is a Euclidean space. At 20% coverage, the space shifts to an FFC space. It becomes a highly perceptually consistent space at 80% coverage. In between about 30% and 70% coverage, the space behaves as a mixture of computational and perceptual spaces. Therefore, the incremental space (marked as squares in Fig. 5) undergoes phase shifts from computational towards perceptual as more and more computational distances are replaced by true perceptual distances. This phase shifting property offers another advantage in addition to making the constructed space perceptually consistent. As a user's relevance judgment changes over time, the space can also change accordingly, thus adapting to the user's changing need.

5 Beyond Single Feature

Several conclusions can be drawn from the above discussion:

- Euclidean distance is an unreliable and inaccurate measure of feature and image dissimilarity.
- Computational dissimilarity measures are not perceptually consistent, though some of them perform better than others in image retrieval.

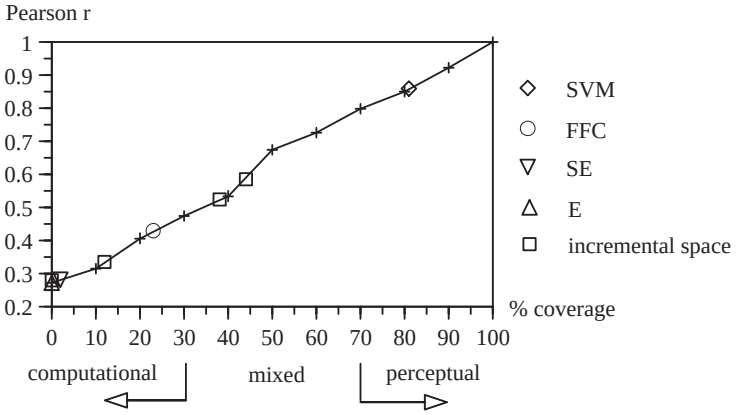

Figure 5. Phase shifts of incremental space. The line indicates the degree of perceptual consistency at various percentage of information coverage. The labels along the line mark the positions of the four texture models (E: Euclidean, SE: scaled Euclidean, FFC, and SVM mapping) according to their perceptual consistency. At low coverage, the incrementally constructed space behaves as a computational space. It shifts towards a mixed space at moderate coverage, and a perceptual space at high coverage.

- The components of a feature cannot be considered as forming the orthogonal dimensions of a multidimensional feature space that is consistent with human's perception. For instance, the various bins of a color histograms are not mutually independent. Likewise, the various texture measurements do not form a perceptually consistent texture space.

Knowing the above, it is not surprising that combining different features to form a linear vector space cannot support perceptually consistent retrieval. Unfortunately, most existing systems adopt this method of combining different features due to its mathematical simplicity.

The problem of perceptually consistent retrieval is further complicated by the fact that many interesting images contain more than one region or object of interest. For example, a beach scene image contains regions of sky, sea, sand, and often human and other objects. Moreover, the same image can be interpreted differently by different users in different application context.

To deal with these complications in an unbiased manner, the Bayesian approach seems to be a natural choice.

1. **Training Stage**
   An image I contains a set of features $f_1, \ldots, f_n$. Given a set of training images, which are categorized into various perceptually meaningful classes c_i (also called *semantic classes*), estimate the probability $P(c_i|f_1, \ldots, f_n)$ that a set of features f_j reliably characterizes a class c_i. Since different feature types are

independent of each other, we have:

$$P(c_i|f_1,\ldots,f_n) = \frac{P(c_i,f_1,\ldots,f_n)}{\prod_j P(f_j)} = \frac{\left|c_i \cap \bigcap_j f_j\right|}{\prod_j |f_j|} \tag{11}$$

The set $c_i \cap \bigcap_j f_j$ can be computed recursively from the sets $c_i \cap f_j$. That is, the various feature types can be decoupled and the sets $c_i \cap f_j$ can be estimated according to each individual feature type. This method overcomes the problem of arbitrarily combining feature types to form a vector space. After training, each image I_j can be associated with a semantic class c_i by the probability $P(c_i|I_j)$.

2. **Retrieval by Category**
   Given a query Q which is a single semantic class, the images I_j can be retrieved by ordering them in decreasing order of $P(Q|I_j)$.

3. **Retrieval by Example**
   Given a query Q which contains sample features f_i, estimate for each semantic class c_i the probability $P(c_i|Q) = P(c_i|f_1,\ldots,f_n)$. Next, compare the probabilities $P(c_i|Q)$ of the query Q with the probabilities $P(c_i|I_j)$ of the images I_j using an appropriate dissimilarity measure, for instance, JD (Equation 8). Finally, the images can be retrieved by ordering them in increasing order of dissimilarity.

The estimation of $P(c_i|f_1,\ldots,f_n)$ is certainly a non-trivial task. At the every least, efficient algorithms will be needed because brute force methods will be computationally too expensive. Nevertheless, the above approach is viable as it can successfully combine various features without resorting to an unreliable combined feature space and can relate low-level features to semantically meaningful classes.

6 Conclusion

Perceptual consistency is important for supporting good image retrieval performance but is very difficult to achieve. Currently, difference between individual color can be measured in a perceptually uniform color space, but the dissimilarity measure between color histograms have not been shown to be perceptually consistent. Nevertheless, empirical tests have shown that non-Euclidean measures are more reliable than Euclidean ones.

In the case of texture, known perceptual texture spaces have yielded consistent results. As for color histograms, computational dissimilarity measures of texture are not consistent with the distances measured in the perceptual space. Fortunately, it is possible to map computational features, particularly Gabor features, to a perceptual space accurately. In this way, texture difference can be measured in the perceptual space to yield perceptually consistent dissimilarity measurement.

It is observed that different feature types, even different components of a feature, cannot be regarded as forming the orthogonal dimensions of a multidimensional combined feature space. Instead, a Bayesian approach is proposed to combine the features in an unbiased manner, which can also relate low-level features to semantically meaningful classes. In general, an image can contain more than one interesting regions. It would be necessary to extend the method to matching a query with images containing multiple regions.

Acknowledgments

This research was supported by NUS ARF R-252-000-049-112, R-252-000-072-112 and, NSTB UPG/98/015.

References

1. IEC 61966-2.1. *Default RGB Colour Space - sRGB*. International Electrotechnical Commission, Geneva, Switzerland, 1999. see also www.srgb.com.
2. M. Amadasun and R. King. Textural features corresponding to textural properties. *IEEE Trans. SMC*, 19(5):1264–1274, 1989.
3. K. K. Benke, D. R. Skinner, and C. J. Woodruff. Convolution operators as a basis for objective correlates of texture perception. *IEEE Trans. SMC*, 18(1):158–163, 1988.
4. R. S. Berns. *Billmeyer and Saltzman's Principles of Color Technology*. John Wiley & Sons, 3rd edition, 2000.
5. A. C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filter. *IEEE Trans. PAMI*, 12(1):55–73, 1990.
6. J. Burbea and C. R. Rao. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multivariate Analysis*, 12:575–596, 1982.
7. J. Burbea and C. R. Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Information Theory*, 28(3):489–495, 1982.
8. G. Ciocca and R. Schettini. A relevance feedback mechanism for content-based image retrieval. *Infor. Proc. and Management*, 35:605–632, 1999.
9. I. J. Cox, M. L. Miller, S. O. Omohundro, and P. N. Yianilos. PicHunter: Bayesian relevance feedback for image retrieval. In *Proc. ICPR '96*, pages 361–369, 1996.
10. G. R. Cross and A. K. Jain. Markov random field texture models. *IEEE Trans. PAMI*, 5:25–39, 1983.
11. D. C. Donderi. Information measurement of distinctiveness and similarity. *Perception and Psychophysics*, 44(6):576–584, 1988.
12. I. H. Godlove. Improved color-difference formula, with applications to the perceptibility and acceptibility fadings. *J. Optical Society of America*, 41(11):760–772, 1951.
13. Y. Gong, G. Proietti, and C. Faloutsos. Image indexing and retrieval based on human perceptual color clustering. In *Proc. CVPR '98*, 1998.
14. S.-S. Guan and M. R. Luo. Investigation of parametric effects using small

colour differences. *Color Research and Application*, 24(5):331–343, 1999.
15. J. Hafner, H. S. Sawhney, W. Esquitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. PAMI*, 17:729–736, 1995.
16. Joseph. F. Hair, R. E. Anderson, and R. L. Tatham. *Multivariate Data Analysis*. Prentice Hall, 1998.
17. C. Heaps and S. Handel. Similarity and features of natural textures. *J. Expt. Psycho.: Human Perception and Performance*, 25(2):299–320, 1999.
18. T. Indow. Predictions based on munsell notation. I. perceptual color differences. *Color Research and Application*, 24(1):10–18, 1999.
19. R. R. Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, 1997.
20. W. K. Leow and R. Li. Adaptive binning and dissimilarity measure for image retrieval and classification. In *Proc. IEEE CVPR 2001*, 2001.
21. F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. PAMI*, 18(7):722–733, 1996.
22. H. Long and W. K. Leow. Invariant and perceptually consistent texture mapping for content-based image retrieval. In *Proc. ICIP*, 2001.
23. H. Long and W. K. Leow. Perceptual consistency improves image retrieval performance. In *Proc. SIGIR*, 2001.
24. H. Long and W. K. Leow. Perceptual texture space improves perceptual consistency of computational features. In *Proc. IJCAI*, pages 1391–1396, 2001.
25. H. Long and W. K. Leow. A hybrid model for invariant and perceptual texture mapping. In *Proc. ICPR (submitted)*, 2002.
26. W. Y. Ma and B. S. Manjunath. Texture features and learning similarity. In *Proc. IEEE CVPR '96*, pages 425–430, 1996.
27. W. Y. Ma and B. S. Manjunath. NeTra: A toolbox for nagivating large image databases. In *Proc. ICIP '97*, pages 568–571, 1997.
28. B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI*, 8(18):837–842, 1996.
29. J. C. Mao and A. K. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25:173–188, 1992.
30. B. M. Mehtre, M. S. Kankanhalli, A. Desai, and G. C. Man. Color matching for image retrieval. *Pattern Recognition Letters*, 16:325–331, 1995.
31. M. Melgosa. Testing CIELAB-based color-difference formulas. *Color Research and Application*, 25(1):49–55, 2000.
32. T. P. Minka and R. W. Picard. Interactive learning using a "society of models". In *Proc. IEEE CVPR '96*, pages 447–452, 1996.
33. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content using color, texture, and shape. In *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases*, volume 1908, pages 173–181, 1993.
34. J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity for color and texture. In *Proc. ICCV '99*, pages 1165–1172,

1999.
35. A. R. Rao and G. L. Lohse. Identifying high level features of texture perception. *CVGIP: Graphical Models and Image Processing*, 55(3):218–233, 1993.
36. A. R. Rao and G. L. Lohse. Towards a texture naming system: Identifying relevant dimensions of texture. In *Proc. IEEE Conf. Visualization*, pages 220–227, 1993.
37. Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proc. ICCV '98*, pages 59–66, 1998.
38. Y. Rui and T. Huang. Optimizing learning image retrieval. In *Proc. IEEE CVPR*, 2000.
39. Y. Rui and T. S. Huang. A novel relevance feedback technique in image retrieval. In *Proc. ACM MM '99*, pages 67–70, 1999.
40. Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in MARS. In *Proc. ICIP '97*, 1997.
41. S. Santini and R. Jain. Similarity measures. *IEEE Trans. PAMI*, 21(9):871–883, 1999.
42. S. Sclaroff, L. Taycher, and M. La Cascia. Image-Rover: A content-based image browser for the world wide web. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1997.
43. J. R. Smith and S.-F. Chang. Single color extraction and image query. In *Proc. ICIP '95*, 1995.
44. J. R. Smith and S.-F. Chang. Multi-stage classification of images from features and related text. In *Proc. 4th Europe EDLOS Workshop*, 1997.
45. T. Song and R. Luo. Testing color-difference formulae on complex images using a CRT monitor. In *Proc. of 8th Color Imaging Conference*, 2000.
46. M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta. A standard default color space for the internet - sRGB. www.color.org/srgb.html, November 1996.
47. H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. SMC*, 8(6):460–47, 1978.
48. I. J. Taneja. New developments in generalized information measures. In P. W. Hawkes, editor, *Advances in Imaging and Electron Physics*, volume 91. Academic Press, 1995.
49. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
50. A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921–1935, 1998.