Computer Vision and Image Understanding **84,** 25–43 (2001)
doi:10.1006/cviu.2001.0934, available online at http://www.idealibrary.com on IDEAL®

Empirical Evaluation of Dissimilarity Measures for Color and Texture

Yossi Rubner,* Jan Puzicha,† Carlo Tomasi,* and Joachim M. Buhmann†

**Robotics Laboratory, Department of Computer Science, Stanford University, Stanford, California 94305; and †Institut für Informatik III, Römerstraße 164, Rheinische Friedrich-Wilhelms-Universität, D-53117 Bonn, Germany*
E-mail: rubner@cs.stanford.edu

Received September 15, 1999; accepted August 8, 2001

This paper empirically compares nine families of image dissimilarity measures that are based on distributions of color and texture features summarizing over 1000 CPU hours of computational experiments. Ground truth is collected via a novel random sampling scheme for color, and by an image partitioning method for texture. Quantitative performance evaluations are given for classification, image retrieval, and segmentation tasks, and for a wide variety of dissimilarity measure parameters. It is demonstrated how the selection of a measure, based on large scale evaluation, substantially improves the quality of classification, retrieval, and unsupervised segmentation of color and texture images.

1. INTRODUCTION

Measuring the dissimilarity between images and parts of images is of central importance for low-level computer vision. The following vision tasks directly rely on some notion of image dissimilarity:

- In *classification* [13, 16, 19, 22], a new image sample has to be assigned to the most similar of a given number of classes. A set of labeled training examples is available. *Supervised segmentation*, i.e., the assignment of image regions to predefined classes is also a classification task.
- In *image retrieval* [1, 9, 10, 18, 23, 24, 27, 30, 33] the user may search a large collection of images for pictures that are similar to a query image. The search is based on perceptual similarities of the attributes color, texture, shape, and composition. *Image annotation* is a special case, where a prototypical image region is interactively specified and all parts of the identical texture class in the image must be labeled. Image annotation techniques are of central importance in applications like map generation from SAR images where completely unsupervised methods often fail.

1077-3142/01 $35.00

• In *unsupervised segmentation* [11, 14, 15, 17, 29] an input image is divided into parts that are homogeneous according to some perceptual attribute. No predefined attribute classes are available in this case. In this context, edge detection can be considered as a special segmentation technique aiming at precise boundary localization [2, 31].

In recent years, dissimilarity measures, based on empirical estimates of the *distribution* of feature, have been developed for classification [22], image retrieval [9, 27, 30, 33], unsupervised segmentation [11, 14], and edge detection [31]. Preliminary benchmark studies have confirmed that distribution–based dissimilarity measures exhibit excellent performance in image retrieval [18, 27], in unsupervised texture segmentation [14], and in conjunction with a *k*–nearest-neighbor classifier, in color- or texture-based object recognition [22, 33]. However, most of these empirical evaluations provide only incomplete and partial information. They either pit one favorite dissimilarity measure against a small number of others, or they provide merely anecdotal evidence, or they only expose a small portion of the space of the parameters that the various dissimilarity measures depend on. Some benchmark studies [18, 27] are more systematic but apply to generic measures and do not elucidate strengths and weaknesses of the various dissimilarity measures for the specific tasks of classification, retrieval, or unsupervised segmentation.

Classification fundamentally differs from the other two applications since with the restriction to a set of known classes the similarity relationship between objects and classes can be inferred from the training data. On the other hand, unsupervised segmentation and image retrieval are concerned with the weaker notion of *image proximity*, based on a *generic* (not class-specific) *similarity measure*. The definition of a similarity measure is based on three design decisions.

1. A feature space representation has to be chosen. For color images, the color vector provides a simple feature vector. For texture, it should characterize the texture content in a neighborhood of a pixel position.
2. The (local) distribution of feature values is estimated. Here, a histogram representation is chosen as a suitable nonparametric estimate of the feature distribution.
3. A measure to compare histograms must be selected, which properly assesses the difference in the represented distributions. Several possible choices are examined in detail.

In this paper, we report on the results of a systematic comparison of nine different families of dissimilarity measures for color and texture. The plots summarize over 1000 hours of CPU time, spent in an exhaustive exploration of a rather large space of parameter settings.

First, in Sections 2 and 3 we review and categorize distribution-based dissimilarity measures, discussing strengths and limitations of each with respect to the different vision tasks mentioned above.

Second, in Section 4 we propose a methodology for the quantitative comparison of color and texture dissimilarity measures. A major contribution here is a statistically sound procedure for the establishment of ground truth, against which the various dissimilarity measures can be compared. This section also explains the principles we adhered to in order to enforce fairness in our comparisons.

Finally, Section 5 provides quantitative comparison results as a function of several parameters such as number of histogram bins, query detail, size of the response to a query, and dimensionality of the feature space. Comparisons are tailored to the specific requirements

of classification, retrieval, and segmentation. The results are interpreted in order to explain which measure works best for which task. As summarized in the concluding Section 6, there are no winners or losers, but rather different tools for different tasks.

2. IMAGE REPRESENTATION

In this section, we describe the color and texture feature spaces that we use in this paper and the representations that we use for the resulting distributions in these spaces.

2.1. Color

Human color perception is based on the incidence of visible light (with wavelengths in the 400 to 700 nm range) upon the retina. Since there are three types of color photo-receptor cone cells in the retina, each with a different spectral response curve, all colors can be completely described by three numbers, corresponding to the outputs of the cone cells. In 1931, the Commission Internationale de l'Éclairage (CIE) adopted standard curves for the color photo-receptor cone cells of a hypothetical standard observer, and defined the CIE XYZ tristimulus values, where all visible colors can be represented using only positive values of X, Y and Z.

Besides the CIE XYZ, other color spaces are used to specify, create and visualize color information (see [35] for more information about color spaces). For instance, the popular RGB color space, as used by television displays, can be visualized as a cube with red, green and blue axes. Different applications have different needs which can be handled better using different color spaces. For our needs it is important to be able to measure differences between colors in a way that matches perceptual similarity as good as possible. This task is simplified by the use of *perceptually uniform* color spaces. A color space is perceptually uniform if a small change of a color will produce the same change in perception anywhere in the color space. In this paper we use the $L^*a^*b^*$ (CIE Lab) color space which was designed such that the perceived differences between single, nearby colors correspond to the Euclidean distance of the color coordinates.

The (nonlinear) conversions from RGB to CIE Lab are given by:[1]

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

$$L^* = \begin{cases} 116\,(Y/Y_n)^{1/3} - 16 & \text{if } Y/Y_n > 0.008856 \\ 903.3(Y/Y_n) & \text{otherwise} \end{cases},$$

$$a^* = 500(f(X/X_n) - f(Y/Y_n)),$$

$$b^* = 200(f(Y/Y_n) - f(Z/Z_n)),$$

where

$$f(t) = \begin{cases} t^{1/3} & \text{if } Y/Y_n > 0.008856 \\ 7.787t + 16/116 & \text{otherwise} \end{cases}.$$

[1] Following ITU-R Recommendation BT.709, we use D_{65} as the reference white point so that $[X_n\ Y_n\ Z_n] = [0.95045\ 1\ 1.088754]$ (see [26]).

2.2. Texture

Over the past decades numerous approaches for the representation of textured images have been proposed ranging from the means and variances of a filter bank output [8, 15], wavelet coefficients [25], wave-packets [16], fractal dimension [3], or parameters of an explicit Markov random field model [5, 19]. Recent comparative studies on this subject can be found in [7, 21, 22, 25]. It should be noted though that in most approaches it can be well distinguished between image representation, i.e., the extraction of a pixel-wise descriptor representing the local texture content and the definition of a similarity measure. As a consequence, most textural features can be incorporated into the proposed distribution-based scheme to define image similarity.

While color is a purely point-wise property of images, texture involves some notion of spatial extent: a single point has no texture. If texture is defined in the frequency domain, the texture information of a point in the image is carried by the frequency content in a local neighborhood. *Gabor filters* are often used for texture analysis and have been shown to exhibit excellent discrimination properties over a broad range of textures [14, 15, 18]. Let $\mathbf{u} \in \mathbb{R}^2$ denote the position in the image or, if discretized, a pixel in the image. Gabor filters are defined by

$$G_{\sigma,\mathbf{f}}(\mathbf{u}) = \frac{1}{2\pi\sigma^2} \exp(-\mathbf{u}^t\mathbf{u}/2\sigma^2) \exp(i\mathbf{f}^t\mathbf{u}), \quad \mathbf{u}, \mathbf{f} \in \mathbb{R}^2, \tag{1}$$

where σ is a localization parameter typically chosen proportionally to the wavelength or *scale* $\frac{1}{\|\mathbf{f}\|}$ of the filter. In keeping with most of the literature on texture, the phase information is ignored by taking only the magnitude of the Gabor responses obtained by convolution with the input image X

$$x_{\mathbf{u}}^r = \left| \left(X * G_{\sigma_r,\mathbf{f}_r} \right)(\mathbf{u}) \right|, \tag{2}$$

where $\mathbf{f}_r$ encodes the scale and the orientation of the r-th filter. Applying a dictionary of Gabor filters results in a vector $\mathbf{x}_{\mathbf{u}} = (x_{\mathbf{u}}^r)$ of responses or *feature channels* for every site in the image, where the number of entries equals the number of scales times the number of orientations that are used.

In this paper we used the family of Gabor filter derived in [18]. Dictionaries with 4, 6, and 8 different orientations over 3, 4, and 5 different scales, respectively, are employed, leading to filter banks of 12, 24, and 40 filters.

2.3. Distribution of Features

Color and texture descriptors vary substantially over an image or image part,[2] both because of inherent variations in surface appearance and as a result of changes in illumination, shading, shadowing, foreshortening, etc. Thus, the appearance of a region is best described by the *distribution of features*, rather than by individual feature vectors. Histograms can be used as nonparametric estimators of empirical feature distributions. However, for high-dimensional feature spaces regular binning often results in poor performance: coarse binning

[2] In the following, we restrict the notation to complete images $X = (\mathbf{x}_{\mathbf{u}})$ for convenience. However, the adaptation to image regions and small image patches as needed for supervised and unsupervised segmentation as well as annotation and edge detection is straightforward.

dulls resolving power, while fine binning leads to large fluctuations due to statistically insignificant sample sizes for most bins. A partial solution is offered by *adaptive binning*, whereby the histogram bins are adapted to the distribution [22, 28, 30]. In the texture segmentation context adaptive bins have recently been popularized as a computational operationalization of Julesz' textons [17]. The binning is induced by a set of *prototypes* $\{\mathbf{c}_i\}$ and is given by the corresponding Voronoi tessellation. Adaptive histograms are formally defined by

$$f(i;X) = \left|\left\{\mathbf{u} : i = \arg\min_j \left\|\mathbf{x}_\mathbf{u} - \mathbf{c}_j\right\|\right\}\right|. \tag{3}$$

Here $\mathbf{x}_\mathbf{u}$ denotes the feature vector at image site $\mathbf{u}$ and $|\cdot|$ denotes the size of a set. The histogram entry $f(i;X)$ corresponds to the number of image pixels in bin i. A suitable set of prototypes can be determined by a vector quantization procedure, e.g., the K-means algorithm used in this paper. Usually, a common set of prototypes is used for a collection of images where the prototypes are computed from the combined distribution over all images. However, sometimes it is useful to compute the adaptive binning separately for each image. We refer to this case as *individual binning*.

For small sample sizes it may be better to estimate only the *marginal histograms*. While information about the *joint occurrence* of feature coefficients in different channels is lost, bin contents in the marginals may be significant where those in the full distribution may be too sparse. Formally, the marginal histograms of the coefficients in feature channel r are given by

$$f^r(i;X) = \left|\left\{\mathbf{u} : t^r_{i-1} < x^r_\mathbf{u} \le t^r_i\right\}\right|. \tag{4}$$

Here, bin i is defined as the feature interval $(t^r_{i-1}, t^r_i]$ of channel r where the boundaries of the interval can be regular or can be determined by adaptive binning. The *cumulative histogram* for marginal histograms is defined as

$$F^r(i;X) = \left|\left\{\mathbf{u} : x^r_\mathbf{u} \le t^r_i\right\}\right|. \tag{5}$$

3. DISSIMILARITY MEASURES

In the following, $D(X,Y)$ denotes a dissimilarity measure between the images X and Y. A superscript $D^r(X,Y)$ indicates that the respective measure is applicable only to the marginal distributions along dimension (channel) r. The dissimilarity values obtained for single feature channels must be combined into a joint overall dissimilarity value. In [27] the Minkowski norms $D(X,Y) = \sum_r (D^r(X,Y))^p$ were investigated, including the limiting case of the maximum norm ($p = \infty$) utilized in [11]. Based on their results $p = 1$ is used in the sequel.

3.1. Distance Measures for Histograms

We distinguish the following four categories of dissimilarity measures:

Heuristic histogram distances. A variety of heuristic histogram distances has been proposed mostly in the context of image retrieval:

- The *Minkowski-form distance* $\mathcal{L}_p$ is defined by

$$D(X, Y) = \left(\sum_i | f(i; X) - f(i; Y)|^p \right)^{1/p}. \tag{6}$$

Here, we chose $1 \leq p \leq \infty$ to ensure metric properties. For example, the $\mathcal{L}_1$ distance has been proposed for computing the dissimilarity scores between color images [33], and the $\mathcal{L}_\infty$ norm was used to measure texture dissimilarity [34]. $\mathcal{L}_1$ computes the sum of absolute distances while $\mathcal{L}_\infty$ measures the maximal difference. Other values of p compromise between these two extremes. *Histogram Intersection* (HI) as proposed in [33] provides a generalization of $\mathcal{L}_1$ to deal with partial matches. Basically, whenever the histograms are of equal size, HI and $\mathcal{L}_1$ are identical, so we treat them as the same category in the sequel.

- The *Weighted-Mean-Variance* (WMV) has been proposed in [18]. For empirical means $\mu_r(X)$, $\mu_r(Y)$ and standard deviations $\sigma_r(X)$, $\sigma_r(Y)$ this distance is defined by

$$D^r(X, Y) = \frac{|\mu_r(X) - \mu_r(Y)|}{|\sigma(\mu_r)|} + \frac{|\sigma_r(X) - \sigma_r(Y)|}{|\sigma(\sigma_r)|}, \tag{7}$$

where $\sigma(\cdot)$ denotes an estimate of the standard deviation of the respective entity. In [18] it is shown that for texture-based image retrieval, this measure based on a Gabor filter image representation has outperformed several parametric models.

Nonparametric test statistics. Nonparametric test statistics provide a sound probabilistic procedure for testing the hypothesis that two empirical distributions have been generated from the same underlying true distribution.

- The *Kolmogorov–Smirnov distance* (KS) was originally proposed in [11] for image segmentation. It is defined as the maximal discrepancy between the cumulative distributions

$$D^r(X, Y) = \max_i |F^r(i; X) - F^r(i; Y)| \tag{8}$$

and has the desirable property to be invariant to arbitrary monotonic feature transformations. However, it is defined only for one dimension.

- A *statistic of the Cramer/von Mises type* (CvM) is defined as the squared Euclidean distance between the distributions,

$$D^r(X, Y) = \sum_i (F^r(i; X) - F^r(i; Y))^2. \tag{9}$$

Similarly to KS, it is defined only for one dimension.

- The χ^2*-statistic* proposed in [27] for segmentation and image retrieval is given by

$$D(X, Y) = \sum_i \frac{(f(i; X) - \hat{f}(i))^2}{\hat{f}(i)}, \tag{10}$$

where $\hat{f}(i) = [f(i; X) + f(i; Y)]/2$ is the mean histogram.

Information-theory divergences. Information-theoretically motivated divergences provide an interesting alternative. Here we examine two special cases:

- The *Kullback–Leibler divergence* (KL) suggested in [22] as an image dissimilarity measure is defined by

$$D(X, Y) = \sum_i f(i; X) \log \frac{f(i; X)}{f(i; Y)}, \tag{11}$$

and measures how inefficient on average it would be to code one histogram using the other as the true distribution for coding. The KL-divergence becomes infinite if $f(i; Y)$ does not dominate $f(i; X)$.

- The *Jeffrey-divergence* (JD) is defined by

$$D(X, Y) = \sum_i f(i; X) \log \frac{f(i; X)}{\hat{f}(i)} + f(i; Y) \log \frac{f(i; Y)}{\hat{f}(i)}.$$

In contrast to the KL-divergence, JD is symmetric and numerically stable when comparing two empirical distributions. JD is sometimes referred to as Jensen–Shannon divergence.

Ground distance measures. A *ground distance* is defined as a distance between individual feature vectors in the underlying feature space. Incorporating this additional information has the potential for dissimilarity measures with improved performance. More formally, the ground distance g_{ij} between bins i and j is defined in terms of the distance between the associated centroids c_i and c_j in the underlying feature space. To some extent, the notion of ground distance is used by measures like the Kolmogorov–Smirnov distance and the statistic of the Cramer/von Mises type, which are based on the cumulative histograms and therefore take into account the relative location of the feature vectors. However, these measures are defined only in one dimension and cannot exploit the ground distance in the full feature space.

- The *Quadratic Form (QF)* [12] provides a heuristic approach,

$$D(X, Y) = \sqrt{(\mathbf{f}_X - \mathbf{f}_Y)^T \, \mathbf{A}(\mathbf{f}_X - \mathbf{f}_Y)}, \tag{12}$$

where $\mathbf{f}_X$ and $\mathbf{f}_Y$ are vectors that list all the entries in $f(i; X)$ and $f(i; Y)$ respectively. We refer to [20] for more details including an efficient implementation. Cross-bin information is incorporated via a similarity matrix $\mathbf{A} = (a_{ij})$ where a_{ij} denotes the similarity between bins $\mathbf{c}_i$ and $\mathbf{c}_j$.

- The *Earth Movers Distance (EMD)* [30] is based on the solution of a transportation problem which is a linear optimization problem. If the cost for moving a single feature unit in the feature space is defined based on the ground distance, then the distance between two distributions is given as the minimal cost to transform one distribution to the other, where the total cost is the sum of the costs needed to move the individual features.

$$D(X, Y) = \frac{\sum_{i,j} g_{ij} d_{ij}}{\sum_{i,j} g_{ij}}, \tag{13}$$

where d_{ij} denotes the dissimilarity between bins i and j, and $g_{ij} \geq 0$ is the optimal flow between the two distributions such that the total cost $\sum_{i,j} g_{ij} d_{ij}$ is minimized, subject to the following constraints:

$$\sum_i g_{ij} \leq f(j;Y), \quad \sum_j g_{ij} \leq f(i;X)$$
$$\sum_{i,j} g_{ij} = \min(f(j;Y),\ f(i;X)), \tag{14}$$

for all i and j. The denominator in (13) is a normalization factor that allows matching parts of distributions that have different total mass. A key advantage of the EMD is that each image may be represented by an individual (possibly with different number of bins) binning that is adapted to its specific distribution.

3.2. Properties of Histogram Distances

Table 1 compares the properties of the different measures. WMV is a parametric measure relying on the means and variances of the marginal distributions. KS and CvM are defined only for cumulative distributions and therefore can be used only with the marginal distributions, while the others are applicable to multidimensional histograms. The EMD has the additional advantage to be applicable to histograms with individual binning.

The validity of the triangle inequality is another important property which distinguishes different measures. The triangle inequality entails lower bounds which can be often exploited to substantially alleviate the computational burden in retrieval tasks [4]. For χ^2, KL and JD the triangle inequality does not hold, for the QF it only holds for specific families of its ground distance, and for the EMD it only holds if the ground distance is metric. All the evaluated measures are symmetric except the KL divergence.

A useful feature for image retrieval is the ability to obtain *partial matches*, i.e., to compute the dissimilarity score only with respect to the most similar image part. Only the HI and the EMD directly enable partial matches. The ability of partial matching is of minor importance for the other applications.

Computational complexity is an important consideration in many applications. For classification and retrieval applications, it is important to differentiate between online and offline complexity. Especially for the WMV the standard deviations can be computed in advance

TABLE 1
Characteristics and Advantages of the Different Distribution-Based Similarity Measures

	$\mathcal{L}_p$	WMV	KS/CvM	χ^2	KL	JD	QF	EMD
Symmetrical	+	+	+	+	−	+	+	+
Triangle inequality	+	+	+	−	−	−	+	+
Comp. complexity	medium	low	medium	medium	medium	medium	high	high
Ground distance	−	−	+	−	−	−	+	+
Multivariate	+	−	−	+	+	+	+	+
Individual binning	−	+	−	−	−	−	−	+
Partial matches	−	−	−	−	−	−	−	+
Nonparametric	+	−	+	+	+	+	+	+

and the similarity scores for a new query can be evaluated efficiently. In contrast to all other advantages, the computational complexity of the EMD is the highest among the evaluated measures since for each dissimilarity score it is necessary to solve a combinatorial optimization problem. However, while using the EMD on large histograms is prohibitive for certain applications, its ability to represent different images by a different binning often enable good results even with a small number of bins, and consequently less computation. In the experiments, the number of bins used for the EMD has been limited to 32 bins, while for the other dissimilarity measures up to 256 bins have been used. Still, the computational complexity of the EMD has turned out to be prohibitive for texture segmentation.

4. BENCHMARK METHODOLOGY

Any systematic comparison of dissimilarity measures should be conform at least to the following guidelines:

1. A meaningful *quality measure* must be defined. Different tasks usually entail different quality measures. The subdivision into classification, retrieval, and segmentation makes it possible to define general-purpose quality criteria for each task. The presented results may thus serve as a useful guide in many practical situations.

2. Performance comparisons should account for the *variety of parameters* that can affect the behavior of each measure. These parameters include the size of the images, queries and statistical samples; the number of neighbors in a k-nearest-neighbor classifier and the number of bins in a histogram; the shape of the bins and their detailed definition; and, for texture, the dimensionality of feature space. A fair comparison in the face of this variability can be achieved by giving every measure the best possible chance to perform well. However, it has to be emphasized that multiple free algorithmic parameters have to be considered as a deficit of a method, since each additional parameter value must be estimated from experimental data, e.g., by cross-validation techniques which may cause tremendous experimental effort. Thus, the degrees of freedom is an important factor for choosing an appropriate method.

3. Processing steps that affect performance independently ought to be *evaluated separately* in order to both sharpen insight and reduce complexity. For instance the effect of different image representations can be understood separately from those of different dissimilarity measures. Also, for segmentation, the grouping procedure can be evaluated separately [14].

4. *Ground truth* should be available. This is a set of data for which the correct solution for a particular problem is known. Collecting ground truth is arguably the hardest problem in benchmarking, because the data should represent a broad range of possible applications, the "correct solution" ought to be uncontroversial, and the ground-truth data set should be large enough to allow a statistically significant performance evaluation. In the following, we summarize our choice of ground truth for color and texture.

4.1. Color

Defining ground truth to measure color similarity over a set of color images is difficult. Our approach was to create disjoint sets of randomly sampled pixels from an image and to consider these sets as belonging to the same class. While for large sets of pixels within a class the color distributions of their pixels will be very similar, for small sets the variations

are larger, mimicking the situation in image retrieval where images of *moderate* similarity have to be identified. From a data base of 20,000 color images comprising the Corel Stock Photo Library, we randomly chose 94 images. This is the same number of images as in the texture case, so that we can compare the results. We defined set sizes of 4, 8, 16, 32, 64 pixels, and for each image we obtained 16 disjoint sets of random samples in all sample sizes, resulting in a ground truth data set of 1504 samples with 94 different classes, one class per image. For the QF and the EMD that employ a ground distance, we use

$$a_{ij} = \exp(-\alpha \|\mathbf{c}_i - \mathbf{c}_j\|) \quad \text{and} \quad d_{ij} = 1 - a_{ij} \tag{15}$$

as the measure of similarity and dissimilarity of bins i and j, where $\|\mathbf{c}_i - \mathbf{c}_j\|$ is the $\mathcal{L}_2$ distance between the bin centers in the CIE $L^*a^*b^*$ color space (see Section 2). The exponential map limits the effect of large distances that otherwise dominate the result. Here we set α to half the standard deviation of all the feature values in the data base. This ground distance makes closeness a relative notion, agrees with results from psychophysics [32], and was found empirically to give good results.

4.2. Texture

In the benchmark study we concentrated on textured images from the Brodatz album as they are widely accepted within the texture research community and provide a joint database which is commonly available. To define ground truth each image is considered as a single, separate class. This is questionable in a few cases, which are circumvented by a preselection of images. We *a priori* selected 94 Brodatz textures by visual inspection. We excluded the textures d25, d30–d31, d39–d45, d48, d59, d61, d88–d89, d91, d94, d97 due to missing micropattern properties. That is, those textures are excluded where the texture property is lost when considering small image blocks. From each of the Brodatz images we extracted sets of 16 random, nonoverlapping blocks sizes $8^2, 16^2, \ldots, 256^2$ pixels.[3] For each sample size this resulted in a ground truth data set of 1504 samples with 94 different classes, just as for color. For the QF and the EMD we again employ (15), with the only difference that $\|\mathbf{c_i} - \mathbf{c}_j\|$ is defined as the $\mathcal{L}_1$ distance between the Gabor responses. Unlike with color where the $\mathcal{L}_2$ distance has a solid psychophysical justification, for texture it is not clear how to relate the different (normalized) channels, so we simply sum them.

4.3. Performance Evaluation

Next, we describe the quality measures which we used for classification, retrieval, and segmentation. For *classification*, a k-NN classifier is used, with different values for k. As a performance measure we use the average misclassification rate in percent estimated by a leave-one-out estimation procedure.

For *image retrieval*, performance is usually measured by *precision*, which is the number of relevant images retrieved relative to the total number of relevant images in the data base, and *recall*, which is the number of relevant images retrieved, relative to the total number of retrieved images. Since our goal is to compare the different methods and not measure performance of a retrieval system, we only plot the precision vs the number of retrieved images.

[3] For a sample size of 256^2 we only extracted 4 samples per class due to the limited size of the original image.

For *unsupervised texture segmentation* we followed the approach of [14] and used a database of random mixtures (512^2 pixels each) containing 100 entities of five Brodatz textures each (such as depicted in Fig. 8). Segmentations are computed on a regular subgrid of size 128^2 by assigning each site to one out of K segments. For each site, a local histogram is extracted to estimate the local feature distribution. We compute marginal histograms which are proportional to the Gabor filter wavelength [15]. For the multivariate histograms, the binning has been adapted to the specific image. Each local histogram is then compared with 80 randomly selected image sites using the dissimilarity measure. To compute an optimal segmentation we implemented the approach of [14] which groups image sites with a high average similarity to obtain a segmentation. As a performance measure we then report the median classification error evaluated over 100 images, where each site is labeled according to the majority rule of corresponding pixels. In addition, we report the percentage of images with more than 20% errors. We consider these as structural segmentation errors, in which typically entire textures are misclassified.

5. RESULTS AND INTERPRETATION

5.1. Classification

The classification performance has been estimated in a leave-one-out procedure for all combinations of the parameter $k \in \{1, 3, 5, 7\}$ and the number of bins $\in \{4, 8, 16, 32, 64, 128, 256\}$.[4] Only odd values are used for k to reduce the chances of ties. It can be shown that the odd values dominate the even values [6, Section 5.6], i.e., a k-NN classifier with odd k performs at least as good as the even $(k + 1)$-NN classifier in the limit of infinite training data. In the texture case, we tried three different filter banks with 12, 24, and 40 filters. The experiments resulted in an enormous amount of information, computed in over 1000 CPU hours. Due to limitations in space, we only plot a few informative cuts from the high-dimensional parameter space. The classification results are summarized in Fig. 1 for the color case and Fig. 2 for the texture case. In the figures, we plot the classification error of the dissimilarity measures as a function of the sample size both for the full distribution and for the marginals cases. The results are further separated into two cases: small histograms (using 8 bins), and large histograms (using 256 bins). An exception to these histogram sizes is the EMD which uses individual adapted histograms that contain more information than the regular global histograms (see Section 3). For the EMD we use 4 bins (instead of 8) for the small histogram case, and 32 bins (instead of 256) for the large histogram.

The following main conclusions can be drawn from the results:

1. Two regimes can be distinguished based on the sample size.

For small sample sizes, the measures which are based on cumulative distributions (KS and CvM) and which thus incorporate ground distance information perform well using marginal distributions. The EMD performed exceptionally well with full distributions, even for the hard case of small histograms where other measures performed poorly. This is explained by the local binning that provides additional information, not available to the other measures. For very small sample sizes the WMV measure performs best in the texture case. This is explained by the fact that WMV only estimates the means and variances of the marginal

[4] For the EMD, because of computational reasons and the additional information carried by the individual binning, we limited the number of bins to a maximum of 32.

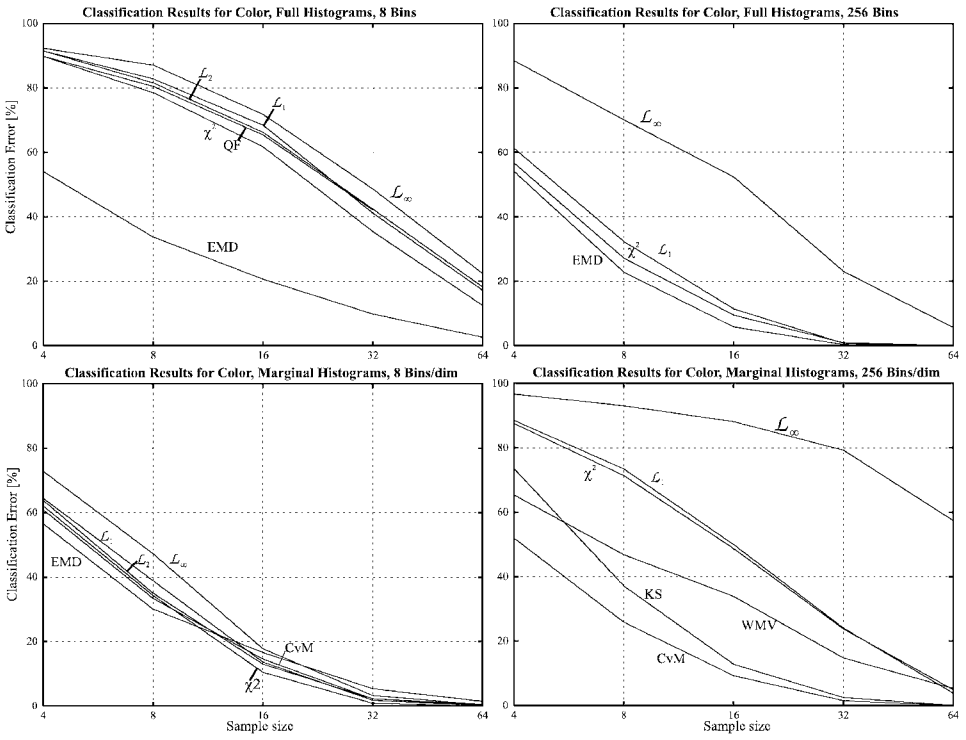

FIG. 1. Classification results for the color data base for different sample sizes and different binning. For each result, an optimal value $k \in \{1, 3, 5, 7\}$ for the k–nearest neighbor classifier has chosen. To assess the statistical significance of the results one should note that the standard deviation can be estimated by $\sigma = \sqrt{e \cdot (1-e)/1504}$, where e denotes the error probability. This yields standard deviations of 2.29, 0.77, 0.56, and 0.36% for an error rate of 50, 10, 5, and 2%, respectively. The corresponding error bars have been omitted from the plot for increased readability.

distributions. These aggregate measurements are less sensitive to sampling noise. The WMV competes less satisfactorily on color since histograms can be more reliably estimated in this case.

For large sample sizes ($\geq 32^2$), the classical χ^2 test statistic and the divergence measures perform best. Jeffrey's divergence behaves more stably than the KL-divergence, as expected. The χ^2-statistic and JD yield nearly identical results in all experiments. We discarded JD from some plots since the curves become visually indistinguishable. The $\mathcal{L}_1$ does best from the class of heuristic measures. $\mathcal{L}_2$ and L_∞ turned out to be consistently inferior in all experiments and should thus not be considered as competitive measures.

2. For texture classification, marginal distributions do better than the multidimensional distributions except for very large sample sizes (256^2). This is explained by the fact that the binning is not well adapted to the data, since it has to be fixed for all sample images over all 94 texture classes. The EMD with its local adaptation does much better in this case. For color, due to the lower dimensionality multivariate adaptive histograms perform better than marginals with the EMD performing best, since histograms can be more reliably estimated even for small sample sizes. We conclude that marginal distributions or well-adapted measures should be used for large feature spaces.

3. More bins help in the multivariate case. The maximally allowed number of bins performs best for multidimensional histograms as shown in Fig. 3. Spending even more

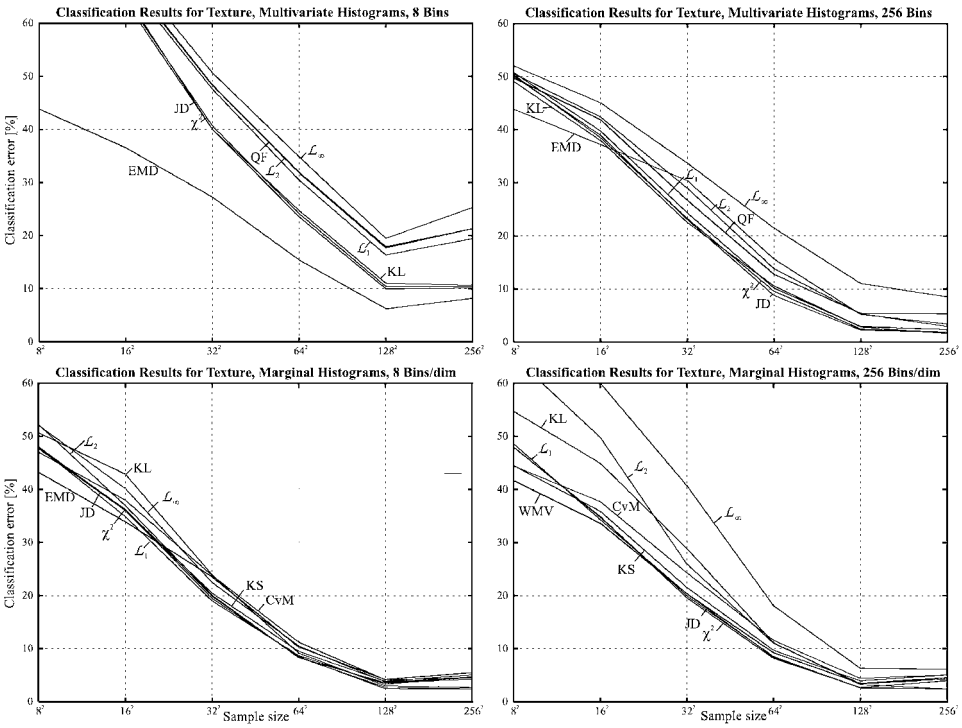

FIG. 2. Classification results for texture classification for different sample sizes and different binning. In each case, the best possible k and the best number of filters has been chosen. The slight deterioration in quality for the sample size of 256^2 is explained by the fact that only 4 samples instead of 16 have been available for each class. Concerning the statistical significance of the results, see the discussion in the caption of Fig. 1.

bins might result in an increased performance, up to a point where close features fall in separate bins, but also results in a prohibitive run-time behavior. Only for the EMD, the local adaptation allows to represent the distribution with a small number of bins which is an advantage if storage complexity is an issue.

For marginal histograms, the binning details often play a negligible role as seen from Fig. 3. However, for small sample sizes as in the color experiments in Fig. 3 too many bins may result in a severe degradation of classification performance. Thus in the following experiments we used 16 bins for marginal histograms and 256 bins for full histograms. It is interesting to note that cumulative histograms with many bins do not suffer from overfitting for small sample sizes, see again Fig. 3. This reflects the fact that it is much easier to estimate a distribution function than a density [6]. In fact, cumulative distributions could be efficiently and robustly estimated with an infinite number of bins.

4. The optimal number k used for the k-NN classifier depends on the noise level of the data. In the small sample regime with a high estimation variance a large value for k is helpful, while for large histogram sample sizes the choice of k plays a negligible role, see Fig. 4. Note that only 15 histograms for each class have been in the data base. For a larger number of samples per class, one would expect a larger number k to work better [6].

5. For the texture case, usually 12 Gabor filters have been sufficient and even outperform a larger number of filters as seen from Fig. 5. However, for very small sample sizes additional filters *implicitly* provide more samples which results in a better performance. We

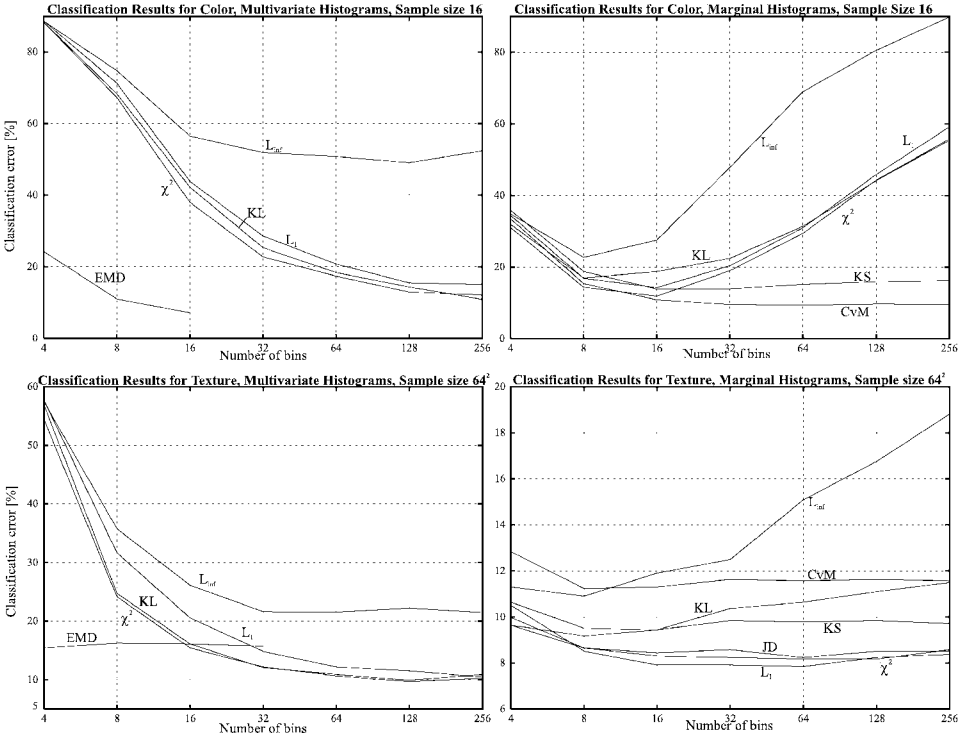

FIG. 3. Classification results depending on the number of bins used. In the experiments a $K = 1$ NN classifier has been used. For the color experiments, a sample size of 16 was used. For the texture experiments, a sample size of 64^2 and 12 filters were employed.

conclude that a small number of features is sufficient to distinguish a large number of texture classes.

5.2. Image Retrieval

As we saw in the results for classification, the EMD, WMV, CvM, and KS usually performed well for the small sample sizes, while JD, χ^2, and KL usually performed better for

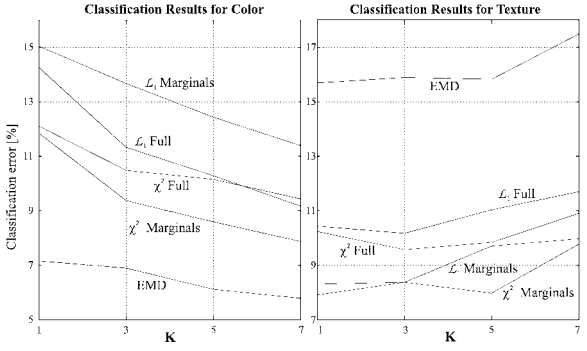

FIG. 4. Classification results depending on K of the K-NN classifier. In the experiments 16 bins have been used with the marginal histograms and 256 bins have been used full histograms. With the EMD, 16/32 locally adapted bins have been employed. For the color experiments, we used a sample size of 16. For the texture experiments, a sample size of 64^2 and 12 filters were used.

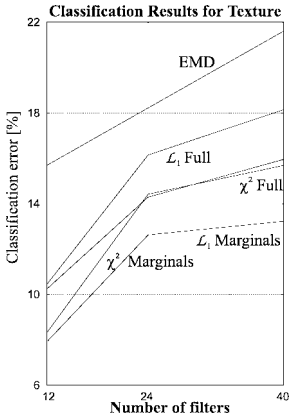

FIG. 5. Texture classification results depending on the number of filters. In the experiments, a sample size of 64^2 and 256 bins were used.

the larger sample sizes. This is confirmed by the retrieval results depicted in Fig. 6. Experiments with small sample size are closer to image retrieval, since they mimic the situation where similar images can have large variability, but should still be retrieved. Therefore, for better recall of a large number of similar images (fewer false negatives), the first class of measures performs better, while for better precision with a few, very similar images (fewer false positives), the second class of measures will probably perform better in real image retrieval systems with a heterogeneous content.

Figure 7 shows an example for color-based image retrieval on a real data base of 20,000 images from the Corel Photo Collection. The color content of the leftmost image of a red car was used as the query, and the eight images with the most similar color contents were returned and displayed in order of increasing distance for different dissimilarity measures. Every image in the data base was represented by a 256-bin multivariate adaptive histogram. Note that this experiment solely meant to illustrate the differences that can be observed using different dissimilarity measures for color.

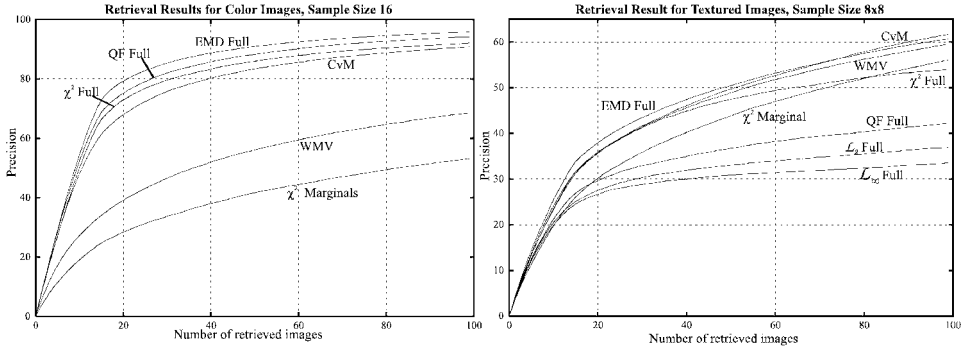

FIG. 6. Precision curves in [%] for selected similarity measures. (left) Color image retrieval for a sample size of 16; (right) textured image retrieval for a sample size of 8^2.

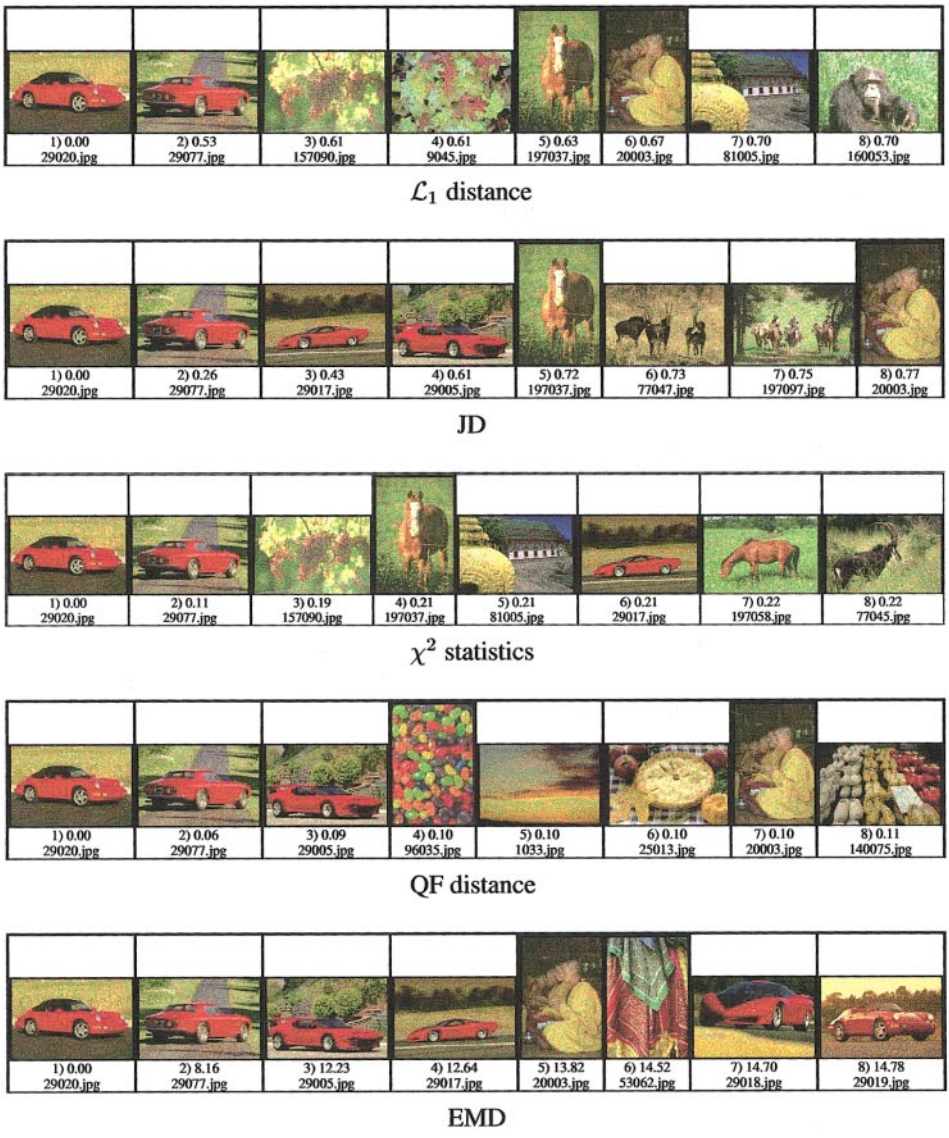

FIG. 7. The eight closest images for each of the red car images in the first column. The queries were processed by a color-based image retrieval system using different histogram dissimilarity measures.

It does not provide a full image retrieval system which should take image structure into account.

5.3. Unsupervised Segmentation

A major technical advantage of image segmentation compared to the other applications is the fact that the binning can be adapted specifically to the image at hand. This leads to an increased accuracy in representing multidimensional distributions. Consequently, adaptive

TABLE 2

Errors by Comparison with Ground Truth over 100 Randomly Generated Images with $K = 5$ Textures, 512^2 Pixels, and 128^2 Sites

	Median	20% quantile
χ^2 full	6.6%	10%
JD full	6.8%	10%
$\mathcal{L}_1$ full	6.8%	9%
χ^2 marginal	8.1%	13%
JD marginal	8.1%	12%
$\mathcal{L}_1$ marginal	8.2%	12%
KS marginal	10.8%	20%
CvM marginal	10.9%	22%

multivariate binning significantly outperforms marginal histograms in the unsupervised segmentation task. This is illustrated in Fig. 8 for an example image and confirmed by the benchmark results on the database with 100 images presented in Table 2. χ^2, JD, and $\mathcal{L}_1$ exhibit very similar performance both with marginal and multidimensional histograms. The best performance was achieved by χ^2 on adaptive multivariate histograms with a median error of 6.6% as compared to 10.8% for the Kolmogorov–Smirnoff test, which was utilized in [11]. Thus, employing the benchmark results to select a proper dissimilarity measure may substantially improve the quality of unsupervised segmentation. For segmentation, the EMD suffers from its high computational complexity and has, therefore, been excluded from the experiments.

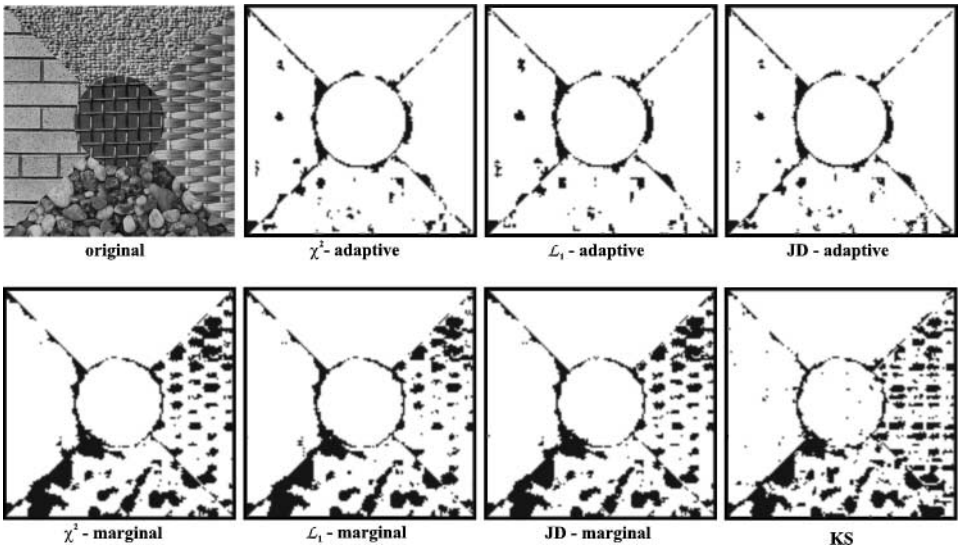

FIG. 8. Examples of segmentation results with $K = 5$ clusters for the different similarity measures under consideration. Misclassified image sites are depicted in black.

6. CONCLUSION

In this paper, a thorough quantitative performance evaluation has been presented for distribution-based image dissimilarity measures. As seen from the result section, there is no measure with best overall performance, but the selection rather depends on the specific task. While marginal histograms and aggregate measures are best for large feature spaces and small samples, multivariate histograms perform very well for large sample sizes. Multivariate histograms are especially effective if the number of classes to be distinguished is small or the binning can be efficiently adapted to the distribution. As a consequence, multivariate histograms performed best for color classification and color retrieval as well as texture segmentation. If disk space is an important issue, the EMD is especially attractive as it allows superior classification and retrieval performance with a much more compact representation.

As a final reminder, the reader should interpret the absolute performance numbers presented in this work with care, since they are highly data dependent. In contrast, most of the conclusion drawn about relative performance of measures have an underlying statistical explanation and are thus more likely to generalize to new problem instances.

ACKNOWLEDGMENT

This work has been supported by the NSF (Grant IRI-9712833) and by the German Research Foundation (DFG # BU 914/3–1 and DFG PU 165/1).

REFERENCES

1. J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu, Virage image search engine: An open framework for image management, in *SPIE Conference on Storage and Retrieval for Image and Video Databases IV*, Vol. 2670, March 1996, pp. 76–87.
2. J. Canny, A computational approach to edge detection, *IEEE Trans. Pattern Anal. Machine Intell.* **8**(6), 1986, 679–698.
3. B. Chaudhuri and N. Sarkar, Texture segmentation using fractal dimension, *IEEE Trans. Pattern Anal. Machine Intell.* **17**(1), 1995, 72–77.
4. K. Clarkson, Nearest neighbor queries in metric spaces, in *ACM Symposium on the Theory of Computing*, pp. 609–617, 1997.
5. G. Cross and A. Jain, Markov random field texture models, *IEEE Trans. Pattern Anal. Machine Intell.* **5**, 1983, 25–39.
6. L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.
7. J. Du Buf, M. Kardan, and M. Spann, Texture feature performance for image segmentation, *Pattern Recognit.* **23**, 1990, 291–309.
8. D. Dunn, W. Higgins, and J. Wakeley, Texture segmentation using 2-D Gabor elementary functions, *IEEE Trans. Pattern Anal. Machine Intell.* **16**(2), 1994, 130–149.
9. 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, *IEEE Comput.* September 1995, pp. 23–32.
10. D. Forsyth, J. Malik, M. Fleck, H. Greenspan, and T. Leung, Finding pictures of objects in large collections of images, in *International Workshop on Object Recognition for Computer Vision*, Cambridge, UK, April 1996.
11. D. Geman, S. Geman, C. Graffigne, and P. Dong, Boundary detection by constrained optimization, *IEEE Trans. Pattern Anal. Machine Intell.* **12**(7), 1990, 609–628.

12. J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, Efficient color histogram indexing for quadratic form distance functions, *IEEE Trans. Pattern Anal. Machine Intell.* **17**(7), 1995, 729–736.
13. R. Haralick, K. Shanmugan, and I. Dinstein, Textural features for image classification, *IEEE Trans. Systems Man Cybernet.* **3**(1), 1973, 610–621.
14. T. Hofmann, J. Puzicha, and J. Buhmann, Textured image segmentation in a deterministic annealing framework, *IEEE Trans. Pattern Anal. Machine Intell.* **20**(8), 1998.
15. A. Jain and F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, *Pattern Recognit.* **24**(12), 1991, 1167–1186.
16. A. Laine and J. Fan, Texture classification by wavelet packet signatures, *IEEE Trans. Pattern Anal. Machine Intell.* **15**, 1993, 1186–1191.
17. J. Malik, S. Belongie, J. Shi, and T. Leung, Textons, contours and regions: Cue integration in image segmentation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV'99)*, 1999.
18. B. Manjunath and W. Ma, Texture features for browsing and retrieval of image data, *IEEE Trans. Pattern Anal. Machine Intell.* **8**(18), 1996, 837–842.
19. J. Mao and A. Jain, Texture classification and segmentation using multiresolution simultaneous autoregressive models, *Pattern Recognit.* **25**, 1992, 173–188.
20. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, and Y. Heights, Querying images by content, using color, texture, and shape, in *SPIE Conference on Storage and Retrieval for Image and Video Databases*, Vol. 1908, April 1993, pp. 173–187.
21. P. Ohanian and R. Dubes, Performance evaluation for four classes of textural features, *Pattern Recognit.* **25**, 1992, 819–833.
22. T. Ojala, M. Pietikäinen, and D. Harwood, A comparative study of texture measures with classification based feature distributions, *Pattern Recognit.* **29**(1), 1996, 51–59.
23. A. Pentland, R. W. Picard, and S. Sclaroff, Photobook: content-based manipulation of image databases, *Inernat. J. Comput. Vision* **18**(3), 1996, 233–254.
24. R. W. Picard and T. P. Minka, Vision texture for annotation, *Multimedia Syst.* **3**, 1995, 3–14.
25. O. Pichler, A. Teuner, and B. Hosticka, A comparison of texture feature extraction using adaptive Gabor filtering, pyramidal and tree-structured wavelet transforms, *Pattern Recognit.* **29**(5), 1996, 733–742.
26. C. Poynton, *A Technical Introduction to Digital Video*, Wiley, New York, 1996.
27. J. Puzicha, T. Hofmann, and J. Buhmann, Non-parametric similarity measures for unsupervised texture segmentation and image retrieval, in *Proc. CVPR'97*, 1997, pp. 267–272.
28. J. Puzicha, T. Hofmann, and J. Buhmann, Histogram clustering for unsupervised image segmentation, in *Proc. CVPR'99*, 1999, pp. 602–608.
29. J. Puzicha, T. Hofmann, and J. Buhmann, Histogram clustering for unsupervised segmentation and image retrieval, *Pattern Recognit. Lett.* **20**(9), 1999, 899–909.
30. Y. Rubner, C. Tomasi, and L. J. Guibas, A metric for distributions with applications to image databases, in *Proc. ICCV'98*, 1998, pp. 59–66.
31. M. Ruzon and C. Tomasi, Color edge detection with the compass operator, in *Proc. CVPR'99*, 1999, pp. 160–166.
32. R. Shepard, Toward a universal law of generalization for psychological science, *Science* **237**, 1987, 1317–1323.
33. M. Swain and D. Ballard, Color indexing, *Internat. J. Comput. Vision* **7**(1), 1991, 11–32.
34. H. Voorhees and T. Poggio, Computing texture boundaries from images, *Nature* **333**, 1988, 364–367.
35. G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, Wiley, New York, 1982.