Empirical Evaluation of Dissimilarity Measures for Color and Texture

Jan Puzicha, Joachim M. Buhmann
Institut für Informatik III
Universität Bonn
D-53117 Bonn, Germany

Yossi Rubner, Carlo Tomasi
Computer Science Department
Stanford University
Stanford, CA, 94305

Abstract

This paper empirically compares nine image dissimilarity measures that are based on distributions of color and texture features summarizing over 1,000 CPU hours of computational experiments. Ground truth is collected via a novel random sampling scheme for color, and via 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 measures. 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* [4, 10], a new image sample is 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* [14, 1, 7, 11], the user searches a large collection of images for instances that are similar to a specified query. The search is based on perceptual similarities of attributes such as color, texture, shape, and composition. In *unsupervised segmentation* [2, 6, 5], an input image is divided into regions that are homogeneous according to some perceptual attribute. No predefined attribute classes are available in this case.

In recent years, dissimilarity measures that are based on empirical estimates of the *distribution* of feature have been developed for classification [10], image retrieval [1, 14, 11, 12] and unsupervised segmentation [2, 5]. Preliminary benchmark studies have confirmed that distribution–based dissimilarity measures exhibit good performance in image retrieval [7, 11], unsupervised texture segmentation [5], and in conjunction with a k–nearest-neighbor classifier, color– or texture–based object recognition [14, 10]. 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 [7, 11] 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.

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 in this paper summarize over 1,000 hours of CPU time, spent in an exhaustive exploration of a rather large space of parameters. First, in sections 2 and 3, we review and categorize distribution-based dissimilarity measures, showing strengths and limitations of each with respect to the different vision tasks mentioned above. Next, 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 to establish 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. We found no all-around 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 our representation of distributions in these spaces.

Color: For human color perception it is sufficient to represent all colors by a three dimensional space [16]. We use the CIE $L^*a^*b^*$ color space which was designed using psychophysical experiments to be *uniform*, in that the perceived differences between individual nearby colors correspond to the Euclidean distances between the color coordinates. Some similarity measures take advantage of the uniformity of a color space.

Texture: Over the past decades numerous approaches for the representation of textured images have been proposed [4, 2, 6, 7]. While color is a purely pointwise property of images, texture involves a notion of spatial extent: a single point has no texture. For each image point, frequency-domain texture descriptors refer instead to the frequency content in a local neighborhood of the point. *Gabor filters* are often used for texture analysis and have been shown to exhibit excellent discrimination properties over a broad range of textures [6, 7, 5]. In this paper we used the family of Gabor filter in log-polar space as derived in [7]. 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.

Distribution of Features: Color and texture descriptors vary substantially over an image or image part[1], 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 non–parametric estimators of empirical feature distributions. However, for high–dimensional feature spaces a regular binning often results in poor performance: coarse binning dulls resolving power, while fine binning leads 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. The binning is induced by a set of *prototypes* $\{\vec{c}_i\}$ and the corresponding Voronoi tessellation. Adaptive histograms are formally defined by

$$f(i;I) \quad = \quad \left| \{ \vec{x} \ : \ i = \arg\min_j \| \vec{I}(\vec{x}) - \vec{c}_j \| \} \right| \quad (1)$$

Here $\vec{I}(\vec{x})$ denotes the feature vector at image location $\vec{x}$. The histogram entry $f(i;I)$ 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. K–means [8].

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

$$f^r(i;I) \quad = \quad \left| \left\{ \vec{x} : \ t^r_{i-1} < I^r(\vec{x}) \leq t^r_i \right\} \right| \ . \quad (2)$$

Here, bin i is defined as the feature interval $(t^r_{i-1}, t^r_i]$ of dimension r. The *cumulative histogram* for marginal histograms is defined as

$$F^r(i;I) = \left| \{ \vec{x} : \ I^r(\vec{x}) \leq t^r_i \} \right| \ . \quad (3)$$

3. Dissimilarity Measures

In the following, $D(I,J)$ denotes a dissimilarity measure between the images I and J. A superscript $D^r(I,J)$ indicates that the respective measure is applied only to the marginal distributions along dimension r. We distinguish the following four categories of dissimilarity measures:

Heuristic histogram distances have been proposed mostly in the context of image retrieval:
(i) The *Minkowski-form distance* $\mathcal{L}_p$ is defined by:

$$D(I,J) = \left(\sum_i |f(i;I) - f(i;J)|^p \right)^{1/p} \ . \quad (4)$$

For example, the $\mathcal{L}_1$ distance has been proposed for computing the dissimilarity scores between color images [14], and the $\mathcal{L}_\infty$ was used for texture dissimilarity [15]. *Histogram Intersection* (HI) as proposed in [14] provides a generalization of $\mathcal{L}_1$ to partial matches.
(ii) The *Weighted–Mean–Variance* (WMV) has been proposed in [7]. This distance is defined by

$$D^r(I,J) = \frac{|\mu_r(I) - \mu_r(J)|}{|\sigma(\mu_r)|} + \frac{|\sigma_r(I) - \sigma_r(J)|}{|\sigma(\sigma_r)|} \ , \quad (5)$$

where $\mu_r(I), \mu_r(J)$ are the empirical means and $\sigma_r(I), \sigma_r(J)$ are the standard deviations of the distributions. $\sigma(\cdot)$ denotes an estimate of the standard deviation of the respective entity. For texture–based image retrieval this measure, based on a Gabor filter image representation, has outperformed several parametric models. [7]

Non–parametric test statistics provide a sound basis for probabilistic procedures that test the hypothesis

[1] In the following, we restrict the notation to complete images I for convenience. However, the adaptation to image regions as needed for segmentation is straight forward.

	$\mathcal{L}^p$	WMV	KS/ CvM	χ^2	KL	JD	QF	EMD
Symmetric	yes	yes	yes	yes	no	yes	yes	yes
Triangle inequality	valid	valid	valid	invalid	invalid	invalid	see text	see text
Computational complexity	medium	low	medium	medium	medium	medium	high	high
Exploits ground distance	no	no	yes	no	no	no	yes	yes
Individual binning	no	yes	no	no	no	no	no	yes
Multiple dimensions	yes	yes	no	yes	yes	yes	yes	yes
Partial matches	see text	no	no	no	no	no	no	yes
Non-parametric	yes	no	yes	yes	yes	yes	yes	yes

Table 1. Characteristics and advantages of the different distribution–based dissimilarity measures.

that two empirical distributions have been generated from the same underlying true distribution.

(i) The *Kolmogorov–Smirnov distance* (KS) has originally been proposed in [2] for image segmentation. It is defined as the maximal discrepancy between the cumulative distributions,

$$D^r(I,J) = \max_i |F_r(i;I) - F_r(i;J)| \qquad (6)$$

and has the desirable property to be invariant to arbitrary monotonic feature transformations.

(ii) A *statistic of the Cramer/von Mises type* (CvM) is also defined based on cumulative distributions:

$$D^r(I,J) = \sum_i (F_r(i;I) - F_r(i;J))^2 \quad . \qquad (7)$$

(iii) The χ^2*–statistic* is given by

$$D(I,J) = \sum_i \frac{\left(f(i;I) - \hat{f}(i)\right)^2}{\hat{f}(i)}, \text{ where} \qquad (8)$$

$\hat{f}(i) = [f(i;I) + f(i;J)]/2$ denotes the joint estimate.

Information–theoretic divergences measure how compact one distribution can be coded using the other one as the codebook. Here we examine two special cases:

(i) The *Kullback–Leibler divergence* (KL) suggested in [10] as an image dissimilarity measure is defined by

$$D(I,J) = \sum_i f(i;I) \log \frac{f(i;I)}{f(i;J)} \quad . \qquad (9)$$

(ii) The *Jeffrey–divergence* (JD) is defined by

$$D(I,J) = \sum_i f(i;I) \log \frac{f(i;I)}{\hat{f}(i)} + f(i;J) \log \frac{f(i;J)}{\hat{f}(i)} \quad .$$

In contrast to the KL–divergence, JD is symmetric and numerically more stable when comparing two empirical distributions.

Ground distance measures are based on perceptually meaningful distance measures between individual features. Employing this *ground distance* may improve the dissimilarity measure between two distributions. 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. However, these measures are defined only in one dimension and cannot exploit the ground distance in the full feature space.

(i) The *Quadratic Form (QF)* distance [3] incorporates cross-bin information via a similarity matrix $\mathbf{A} = [a_{ij}]$ where a_{ij} denote similarity between bins i and j.

$$D(I,J) = \sqrt{(\vec{f}_I - \vec{f}_J)^T \mathbf{A} (\vec{f}_I - \vec{f}_J)} \ , \qquad (10)$$

where $\vec{f}_I$ and $\vec{f}_J$ are vectors that list all the entries in $f(i;I)$ and $f(i;J)$ respectively. We refer to [9] for more details including efficient implementations.

(ii) The *Earth Movers Distance (EMD)* [12] is based on the minimal cost to transform one distribution to the other. If the cost of moving a single feature unit in the feature space is the ground distance, then the distance between two distributions is given by the minimal sum of the costs incurred to move all the individual features. The EMD can be defined as the solution of a transportation problem which can be solved by linear optimization:

$$D(I,J) = \frac{\sum_{i,j} g_{ij} d_{ij}}{\sum_{i,j} g_{ij}} \qquad (11)$$

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:

$$\sum_i g_{ij} \leq f(j;J) \ , \qquad \sum_j g_{ij} \leq f(i;I) \ ,$$
$$\sum_{i,j} g_{ij} = \min\left(f(j;I), f(i;J)\right) \ , \qquad (12)$$

for all i and j. The denominator in (11) is a normalization factor that permits matching parts of distributions with different total mass. If the ground distance is a metric and the two distributions have the same amounts of total mass, the EMD defines a metric. As a key advantage of the EMD each image may

be represented by a different binning that adapts to its specific distribution. When marginal histograms are used, the dissimilarity values obtained for the individual dimensions must be combined into a joint overall dissimilarity value. In [11] the Minkowski norms $D(I,J) = (\sum_r (D^r(I,J))^p)^{1/p}$ were investigated, including the limiting case $p = \infty$ utilized in [2]. Based on their results $p = 1$ is used in the sequel.

3.1. Properties

Table 1 compares the properties of the different measures. KS, CvM and WMV are defined only for marginal distributions. Metric dissimilarity measures enable more efficient indexing algorithms for image retrieval, since the triangle inequality entails lower bounds that can be exploited to substantially alleviate the computational burden. For the χ^2, KL, JD the triangle inequality does not hold, while for the QF and the EMD it holds only for specific ground distances. All the evaluated measures are symmetric except the HI and the KL divergence. A useful property for image retrieval is the ability to handle *partial matches*, i.e. to compute the dissimilarity score only with respect to the most similar image part [12]. The ability for partial matching is of minor importance for the other applications. Only the HI and the EMD allow for partial matches directly. Computational complexity is an important consideration. For applications such as image retrieval, it is important to differentiate between online and off-line complexity. Especially for the WMV the standard deviations can be computed in advance and the dissimilarity scores for a new query can be evaluated efficiently. The computational complexity of the EMD is the highest among the evaluated measures, as for each dissimilarity calculation a linear optimization is necessary. However, while using the EMD on large histograms is prohibitive for certain applications, its ability to represent different images by a different binning often yields good results even with small number of bins, and consequently less computation. In our experiments we have limited the number of bins for the EMD to 32 bins, while for the other dissimilarity measures we used up to 256 bins.

4. Benchmark Methodology

Any systematic comparison of dissimilarity measures should conform at least to the following guidelines:
(i) 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.
(ii) 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.
(iii) Processing steps that affect performance independently of each other 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 [5].
(iv) *Ground truth* should be available which is a set of data for which the correct solution of 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 for a statistically significant performance evaluation. In the following, we summarize our choice of ground truth for color and texture.
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 database 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 from the two modalities. 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. For each of the five set sizes, this resulted into a ground-truth data set of $16 \times 94 = 1504$ samples in 94 different classes, one class per image. For the QF and the EMD that employ a ground distance, we use

$$a_{ij} = \exp(-\alpha \|\vec{c_i} - \vec{c_j}\|) \text{ and } d_{ij} = 1 - a_{ij} \quad (13)$$

as the measure of similarity and dissimilarity of bins i and j, where $\|\vec{c_i} - \vec{c_j}\|$ 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, which otherwise dominate the result. This

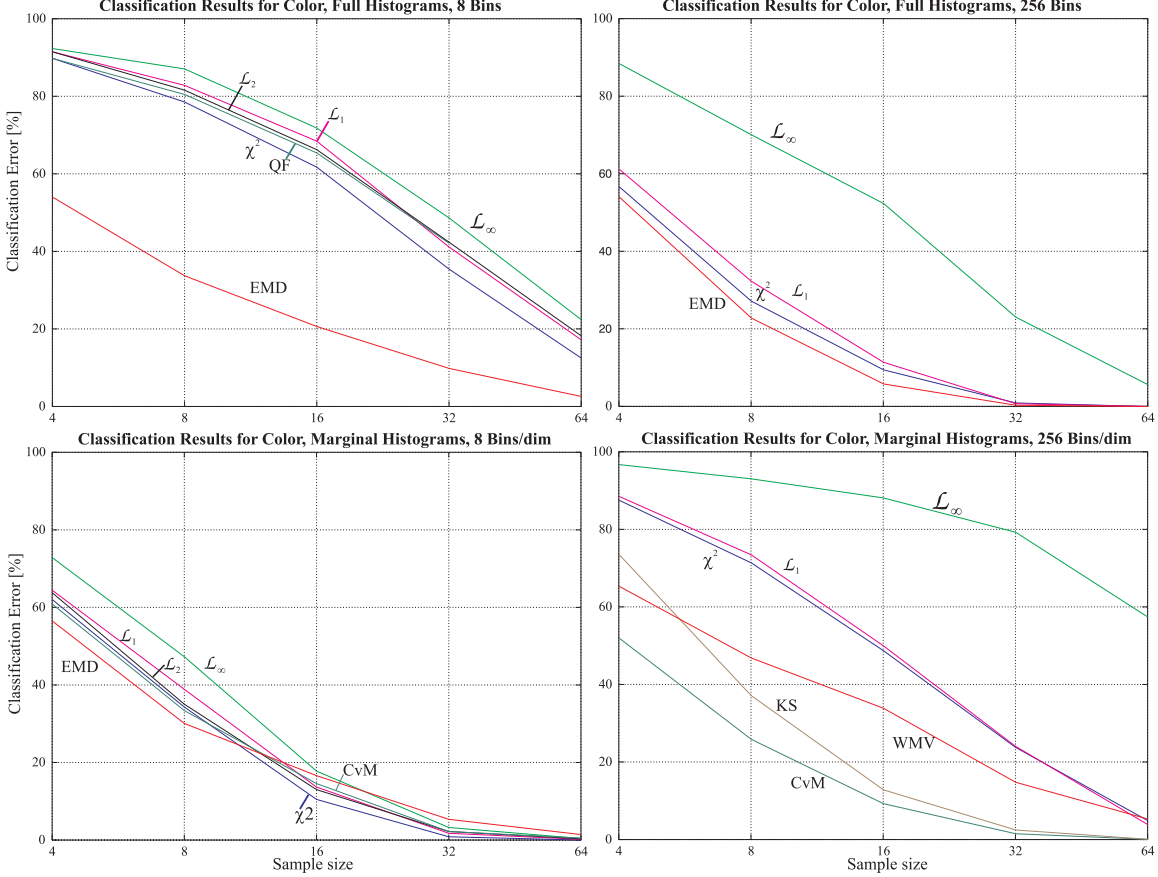

Figure 1. Classification results for the *color* database for different sample sizes and different binning. For each result, an optimal value $k \in \{1, 3, 5, 7\}$ for the k–nearest neighbor classifier has been chosen.

agrees with results from psychophysics [13]. Here we set α to half the standard deviation of all the feature values in the database. This makes closeness a relative notion, and was found empirically to give good results.

Texture: In our 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 pre–selection of images. We selected 94 Brodatz textures *a priori* by visual inspection. We excluded the textures d25, d30, d31, d39-d45, d48, d59, d61, d88, d89, d91, d94, d97 due to missing micro–pattern 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, non–overlapping blocks sizes 8×8, 16×16, $\ldots$, 256×256 pixels[2]. For each sample size this resulted in a ground truth data set of $16 \times 94 = 1504$ samples in 94 different classes, just as for color. For the QF and the EMD we again employ (13), with the only difference that $\|\vec{c_i} - \vec{c_j}\|$ 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) dimensions, so we simply sum them.

Performance Evaluation for classification, retrieval, and segmentation. For *classification*, a k–NN classifier is used, with k having the values 1, 3, 5, and 7. We use only odd values to reduce the chances of ties. As a performance measure we use the average misclassification rate in percent applying a leave–one–out estimation procedure.

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

[2]For a sample size of 256×256 we only extracted 4 samples per class because of the limited size of the original image.

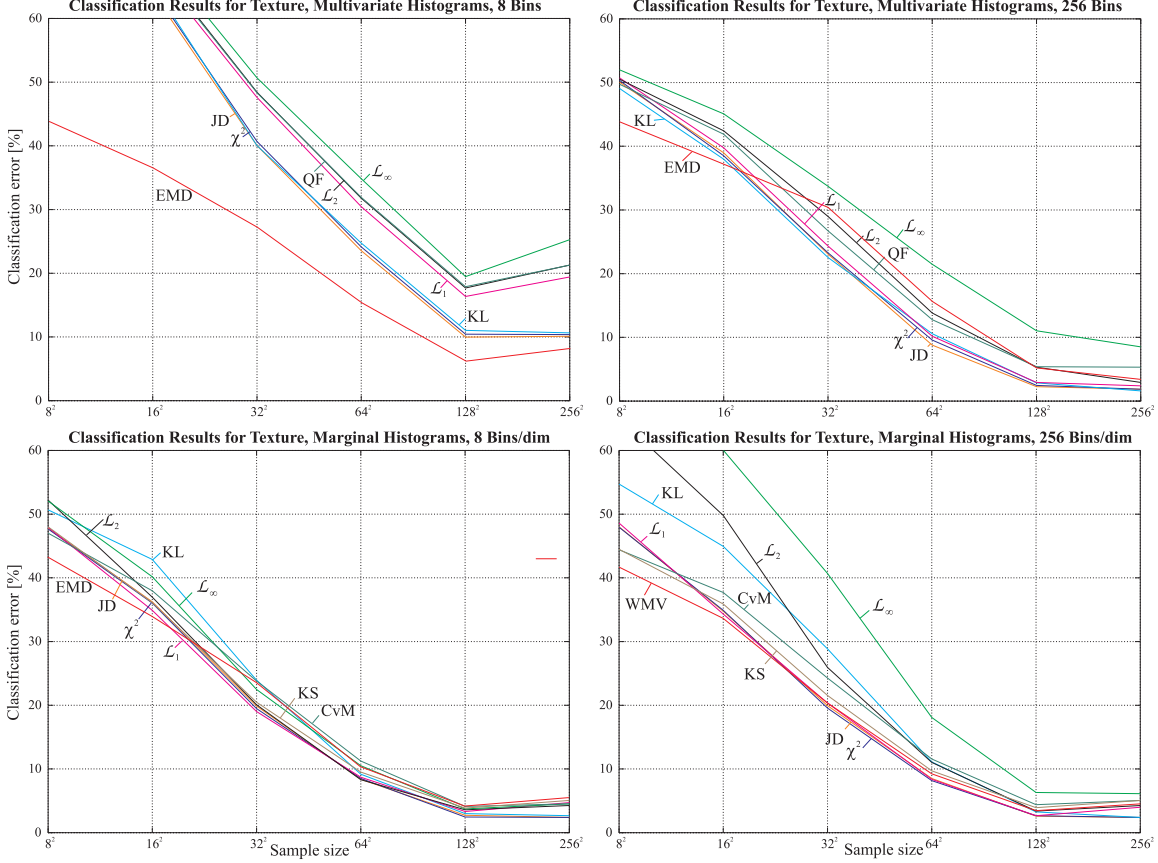

Figure 2. Results for *texture* classification for different sample sizes and different binning. In each case, the best possible k and the best number of filters has been chosen.

For *unsupervised texture segmentation* we followed the approach of [5] and used a database of random mixtures (512×512 pixels each) containing 100 entities of five Brodatz textures each (see Figure 4). Segmentations are computed on a regular sub-grid of size 128×128 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 [6]. For the multivariate histograms, the binning has been adapted to the specific image. Each local histogram is then compared with 80 randomly selected images sites using the dissimilarity measure. To compute an optimal segmentation we implemented the approach of [5] which groups image sites with a high average similarity to obtain a segmentation. As a performance measure we report the average 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 failures as structural segmentation errors with typically entire textures being misclassified.

5. Results and Interpretation

Classification The classification performance has been estimated in a leave–one–out procedure for all combinations of parameters $k \in \{1, 3, 5, 7\}$, number of bins $\in \{4, 8, 16, 32, 64, 128, 256\}$ [3]. In the texture case, we tried three different filter banks with $12, 24$ and 40 filters, respectively. The experiments resulted in an enormous amount of information, computed in over 1,000 CPU hours. Due to limitations in space, we present here only the main results, and plot a few informative cuts from the high-dimensional parameter space. The classification results are summarized in Figure 1 (color) and Figure 2 (texture). We plot the classification error of the dissimilarity measures as a function of the sample size both for the full distribution (top) and for the marginal cases (bottom). The results are further separated into two cases: small histograms with 8 bins (left), and large histograms with 256 bins (right). An exception to these histogram sizes is the EMD which uses locally adapted histograms. As

[3] For EMD because of computational limitations and the additional information carried by the local binning, we used only number of bins $\in \{4, 8, 16, 32\}$.

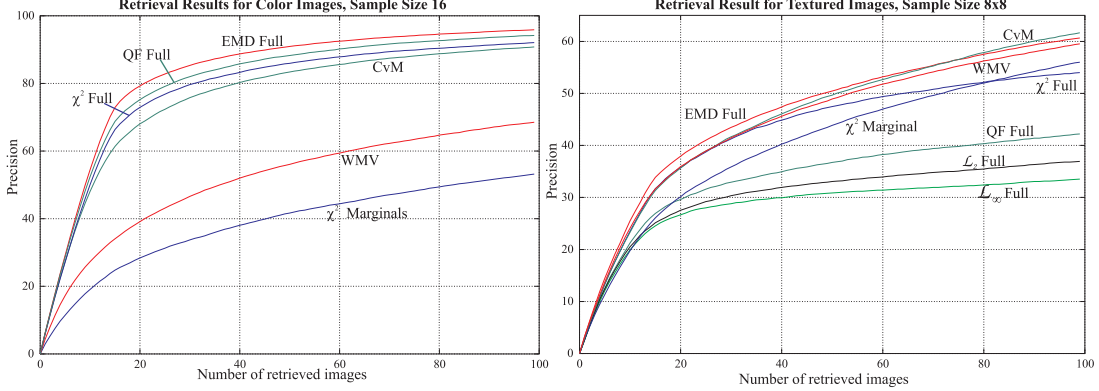

Figure 3. Precision curves in $[\%]$ for selected similarity measures. Left: color retrieval for a sample size of 16. Right: textured image retrieval for a sample size of 8×8.

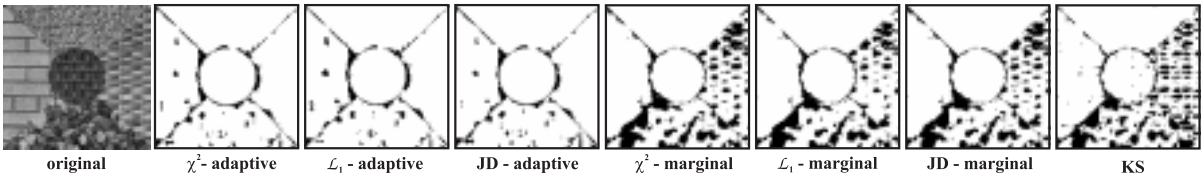

Figure 4. Examples of segmentation results with $K = 5$ clusters for the different similarity measures under consideration. Misclassified image sites are depicted in black.

discussed in Section 3, these contain more information than fixed histograms. For a fair comparison, we use 4 bins for the small histogram case for the EMD (in contrast to 8 bins), and 32 bins for the large histogram (in contrast to 256 bins). The following main conclusions can be drawn.

(i) Two regimes can be distinguished based on the sample size:

For small sample sizes, the WMV measure performs best in the texture case (last plot in figure 1). This behaviour is explained by the fact that WMV only estimates the means and variances of the marginal 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. The measures which are based on cumulative distributions (KS and CvM) and which thus incorporate ground distance information also perform well using marginal distributions. The EMD performed exceptionally well with full distributions, even for the hard case of small histograms where other measures scored poorly. This is explained by the local binning that provides additional information, not available to the other measures.

For large sample sizes, 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. The $\mathcal{L}_1$ does best from the class of heuristic measures.

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

(iii) The maximally allowed number of bins performs best for multidimensional histograms. More bins might result in an increased performance, up to a point where close features fall in separate neighboring bins, but also result 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 play a negligible role.

For the texture case, usually 12 Gabor filters have been sufficient. However, for small sample sizes additional filters *implicitly* provide more samples which results in a better performance. We conclude that a small number of features is sufficient to distinguish a large number of texture classes.

Image Retrieval As we saw in the results for classification, the EMD, WMV, CvM, and KS performed very well for the small sample sizes, while JD, χ^2, and KL usually performed better for the larger sample sizes. This is confirmed by the retrieval results

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

Table 2. Errors by comparison with ground truth over 100 randomly generated images with $K = 5$ textures, 512×512 pixels and 128×128 sites.

depicted in Figure 3. Small sample size is closer to image retrieval, 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.

Unsupervised Segmentation As a major difference in segmentation the binning can be adapted to the image at hand. This leads to an increased accuracy in representing multidimensional distributions. Consequently, adaptive multivariate binning significantly outperforms marginal histograms in the unsupervised segmentation task. This is illustrated in Figure 4 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 [2]. Thus, employing the benchmark results to select a proper dissimilarity measure substantially improves the quality of unsupervised segmentation. For segmentation, the EMD suffers from its high computational complexity and has, therefore, been excluded from the experiments.

6. Conclusion

In this paper, a thorough quantitative performance evaluation has been presented for distribution–based image dissimilarity measures. No measure exhibits 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 storage space is an important issue, the EMD is especially attractive since it allows superior classification and retrieval performance with a much more compact representation, but at a higher computational cost.

References

[1] M. Flickner et al. Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–32, Sept. 1995.

[2] D. Geman et al. Boundary detection by constrained optimization. *IEEE Trans. PAMI*, 12(7):609–628, 1990.

[3] J. Hafner et al. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. PAMI*, 17(7):729–736, 1995.

[4] R. Haralick, K. Shanmugan, and I. Dinstein. Textural features for image classification. *IEEE Trans. Systems, Man and Cybernetics*, 3(1):610–621, 1973.

[5] T. Hofmann, J. Puzicha, and J. Buhmann. Textured image segmentation in a deterministic annealing framework. *IEEE Trans. PAMI*, 20(8), 1998.

[6] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[7] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI*, 8(18):837–842, 1996.

[8] N. M. Nasrabad and R. A. King. Image coding using vector quantization: A review. *IEEE Trans. on Communication*, 36(8):957–971, August 1988.

[9] W. Niblack et al. Querying images by content, using color, texture, and shape. In *SPIE Conference on Storage and Retrieval for Image and Video Databases*, volume 1908, pages 173–187, April 1993.

[10] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[11] J. Puzicha, T. Hofmann, and J. Buhmann. Non–parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proc. CVPR'97*, pages 267–272, 1997.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision*, pages 59–66, Bombay, India, January 1998.

[13] R. N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323, 1987.

[14] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[15] H. Voorhees and T. Poggio. Computing texture boundaries from images. *Nature*, 333:364–367, 1988.

[16] G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and Sons, New York, NY, 1982.