International Journal of Computer Vision 40(2), 99–121, 2000

The Earth Mover's Distance as a Metric for Image Retrieval

YOSSI RUBNER, CARLO TOMASI AND LEONIDAS J. GUIBAS
Computer Science Department, Stanford University, Stanford, CA 94305, USA
rubner@cs.stanford.edu
tomasi@cs.stanford.edu
guibas@cs.stanford.edu

Abstract. We investigate the properties of a metric between two distributions, the *Earth Mover's Distance* (EMD), for content-based image retrieval. The EMD is based on the minimal cost that must be paid to transform one distribution into the other, in a precise sense, and was first proposed for certain vision problems by Peleg, Werman, and Rom. For image retrieval, we combine this idea with a representation scheme for distributions that is based on vector quantization. This combination leads to an image comparison framework that often accounts for perceptual similarity better than other previously proposed methods. The EMD is based on a solution to the transportation problem from linear optimization, for which efficient algorithms are available, and also allows naturally for partial matching. It is more robust than histogram matching techniques, in that it can operate on variable-length representations of the distributions that avoid quantization and other binning problems typical of histograms. When used to compare distributions with the same overall mass, the EMD is a true metric. In this paper we focus on applications to color and texture, and we compare the retrieval performance of the EMD with that of other distances.

Keywords: image retrieval, perceptual metrics, color, texture, Earth Mover's Distance

1. Introduction

Multidimensional distributions are often used in computer vision to describe and summarize different features of an image. For example, the one-dimensional distribution of image intensities describes the overall brightness content of a gray-scale image, and a three-dimensional distribution can play a similar role for color images. The texture content of an image can be described by a distribution of local signal energy over frequency. These descriptors can be used in a variety of applications including, for example, image retrieval.

It is often advantageous to 'compress' or otherwise approximate an original distribution by another distribution with a more compact description. This yields important savings in storage and processing time, and most importantly, as we will see, a certain perceptual robustness to the matching. Multidimensional distributions are usually compressed by partitioning the underlying space into a fixed number of bins, usually of a predefined size: the resulting quantized data structure is a histogram. However, even when the binning is adaptive, based on the overall distribution of the features of all the images in the database, often for specific images only a small fraction of the bins in a histogram contain significant information. For instance, when considering color, a picture of a desert landscape contains mostly blue pixels in the sky region and yellow-brown pixels in the rest. A finely quantized histogram in this case is highly inefficient. On the other hand, a multitude of colors is a characterizing feature for a picture of a carnival in Rio, and a coarsely quantized histogram would be inadequate. In brief, because histograms are fixed-size structures, they cannot achieve a balance between expressiveness and efficiency.

In contrast, we propose *variable-size descriptions* of distributions. In our *signatures*, as we call these descriptions, the dominant clusters are extracted from the original distribution using a clustering algorithm such as vector quantization, and are used to form its

compressed representation. A signature is a set of the main clusters or modes of a distribution, each represented by a single point (the cluster center) in the underlying space, together with a weight that denotes the size of that cluster. Simple images have short signatures, complex images have long ones. Of course, in some applications, fixed-size histograms may still be adequate, and can be considered as special cases of signatures.

In addition to histograms and signatures which are based on global or local tessellation of the space into non-overlapping regions, there are other techniques to describe non-parametric distributions. For example, in kernel density estimation (Duda and Hart, 1973), each data point is replaced by some kernel and the density estimations is regarded as the superposition of all these kernels. These techniques are out of the scope of this paper.

Given two distributions, it is often useful to define a quantitative measure of their dissimilarity, with the intent of approximating perceptual dissimilarity as well as possible. This is particularly important in image retrieval applications, but has fundamental implications also for the understanding of texture discrimination and color perception. Defining a distance between two distributions requires first a notion of distance between the basic features that are aggregated into the distributions. We call this distance the *ground distance*. For instance, in the case of color, the ground distance measures dissimilarity between individual colors. Fortunately, color ground distance has been carefully studied in the literature of psychophysics, and has led to measures like the CIE-Lab color space (Wyszecki and Stiles, 1982). To be sure, this space was designed based on psychophysical experiments where colors were presented in pairs and on a neutral background. While this limits the appropriateness of this space for the more complex situations encountered in retrieval, we believe that it is hard to do better than CIE-Lab without explicitly modelling the geometric layout of colors in images. While RGB space has proven clearly inadequate in our experiments, it is possible that other spaces, such as HSV, may lead to performance similar to that obtained with CIE-Lab.

In this paper, we address the problem of lifting these distances from individual features to full distributions. In other words, we want to define a consistent measure of distance, or dissimilarity, between two distributions of mass in a space that is itself endowed with a ground distance. For color, this means finding distances between image color distributions. For texture, we locally describe the texture content of a small neighborhood in an image as distribution of energy in the frequency domain. The "lifted" distance is a distance between distributions of such local descriptors over the entire images, regarded as distribution of textures.

Mathematically, it would be convenient if these distribution distances were true metrics, which would lead to more efficient data structures and search algorithms (Bozkaya and Ozsoyoglu, 1997; Clarkson, 1997). Practically, it is important that distances between distributions correlate with human perception. In this paper we strive to achieve both goals. For the first we have proof, for the second we show experiments. We also would like these distances to allow for partial matches when one distribution is compared to a subset of the other. For partial matches, the distances we define are not metric. Concerning this point, we refer to Tversky's discussion (Tversky, 1977) of the non-metric nature of perceptual distances. From a practical point of view, our measure deals naturally both with full, metric matches and with partial, non-metric matches.

In this paper we capitalize on the old *transportation problem* (Rachev, 1984; Hitchcock, 1941) from linear optimization, which was first introduced into computer vision by Peleg et al. (1989) to measure the distance between two gray-scale images. For image retrieval, we use this distance measure to compare two signatures in color or texture space. As discussed in more detail in the next section, this leads to very different computational properties, mainly because signatures rather than pixels are compared to each other. We give the name of *Earth Mover's Distance* (EMD), suggested by Stolfi (1994), to this metric in this new context. The transportation problem is to find the minimal cost that must be paid to transform one distribution into the other. The EMD is based on a solution to the transportation problem for which efficient algorithms are available, and it has many desirable properties for image retrieval, as we will see. It is also more robust in comparison to other histogram matching techniques, in that it suffers from no arbitrary quantization problems due to rigid binning, and it tolerates well some amount of deformations that shift features in the feature space. This robustness results in increased precision for image retrieval. It allows for partial matching, and hence naturally supports partial image retrieval queries. It can be applied to signatures with different sizes, which leads to better storage utilization. When used to compare

distributions that have the same overall mass, the EMD is a true metric.

In this paper we focus on applications of the EMD to color and texture images. In the next section, we introduce histograms and survey some of the existing measures of dissimilarity and their drawbacks. Then, in Sections 3 and 4, we introduce the concepts of a signature and of the Earth Mover's Distance (EMD), which we apply to color and texture in Section 5. We compare the results of image retrieval using the EMD with those obtained with other metrics, and demonstrate the unique properties of the EMD for texture-based retrieval. Section 6 concludes with a summary and plans for future work.

2. Previous Work

Image retrieval systems usually represent image features by multi-dimensional histograms. For example, the color content of an image is defined by the distribution of its pixels in some color space (Swain and Ballard, 1991; Hafner et al., 1995; Belongie et al., 1998). Texture features are commonly defined by energy distributions in the spatial frequency domain (Farrokhnia and Jain, 1991; Bigün and Buf, 1994; Manjunath and Ma, 1996). Image databases are indexed by histograms of these distributions, and those images that have the closest histograms to that specified in the query are retrieved. For such a search, a measure of dissimilarity between histograms must be defined. In this section we formally define histograms, and discuss some of the most common histogram dissimilarity measures that are used for image retrieval. In Section 4 we define the EMD. In addition to histograms, this distance is well defined also for signatures, defined in Section 3. In Section 5 we also compare the EMD with the other methods surveyed below.

A *histogram* $\{h_\mathbf{i}\}$ is a mapping from a set of d-dimensional integer vectors $\mathbf{i}$ to the set of nonnegative reals. These vectors typically represent bins (or their centers) in a fixed partitioning of the relevant region of the underlying feature space, and the associated reals are a measure of the mass of the distribution that falls into the corresponding bin. For instance, in a grey-level histogram, d is equal to one, the set of possible grey values is split into N intervals, and $h_\mathbf{i}$ is the number of pixels in an image that have a grey value in the interval indexed by $\mathbf{i}$ (a scalar in this case). The fixed partitioning of the feature space does not have to be regular. If the distribution of features of all the images is known *a priori*, adaptive binning can be used.

Several measures have been proposed for the dissimilarity between two histograms $H = \{h_\mathbf{i}\}$ and $K = \{k_\mathbf{i}\}$. We divide them into two categories. The *bin-by-bin* dissimilarity measures only compare contents of corresponding histogram bins, that is, they compare $h_\mathbf{i}$ and $k_\mathbf{i}$ for all $\mathbf{i}$, but not $h_\mathbf{i}$ and $k_\mathbf{j}$ for $\mathbf{i} \neq \mathbf{j}$. The *cross-bin* measures also contain terms that compare non-corresponding bins. To this end, cross-bin distances make use of the *ground distance* $d_{\mathbf{ij}}$, defined as the distance between the representative features for bin $\mathbf{i}$ and bin $\mathbf{j}$. Predictably, bin-by-bin measures are more sensitive to the position of bin boundaries.

2.1. Bin-By-Bin Dissimilarity Measures

In this category only pairs of bins in the two histograms that have the same index are matched. The dissimilarity between two histograms is a combination of all the pairwise comparisons. A ground distance is used by these measures only implicitly and in an extreme form: features that fall into the same bin are close enough to each other to be considered the same, and those that do not are too far apart to be considered similar. In this sense, bin-by-bin measures imply a binary ground distance with a threshold depending on bin size.

Minkowski-Form Distance:

$$d_{L_r}(H, K) = \left(\sum_\mathbf{i} |h_\mathbf{i} - k_\mathbf{i}|^r \right)^{1/r}.$$

The L_1 distance is often used for computing dissimilarity between color images (Swain and Ballard, 1991). Other common usages are L_2 and L_∞. In Stricker and Orengo (1995) it was shown that for image retrieval the L_1 distance results in many false negatives because neighboring bins are not considered.

Histogram Intersection:

$$d_\cap(H, K) = 1 - \frac{\sum_\mathbf{i} \min(h_\mathbf{i}, k_\mathbf{i})}{\sum_\mathbf{i} k_\mathbf{i}}.$$

The histogram intersection (Swain and Ballard, 1991) is attractive because of its ability to handle partial matches when the areas of the two histograms (the sum over all the bins) are different. It is shown in Swain and Ballard (1991) that when the areas of the two histograms are equal, the histogram intersection is equivalent to the (normalized) L_1 distance.

Kullback-Leibler Divergence and Jeffrey Divergence: The Kullback-Leibler (K-L) divergence (Kullback, 1968) is defined as follows:

$$d_{KL}(H, K) = \sum_{\mathbf{i}} h_{\mathbf{i}} \log \frac{h_{\mathbf{i}}}{k_{\mathbf{i}}} .$$

From the information theory point of view, the K-L divergence has the property that it measures how inefficient on average it would be to code one histogram using the other as the code-book (Cover and Thomas, 1991). However, the K-L divergence is non-symmetric and is sensitive to histogram binning. The empirically derived Jeffrey divergence is a modification of the K-L divergence that is numerically stable, symmetric and robust with respect to noise and the size of histogram bins (Puzicha et al., 1997). It is defined as:

$$d_J(H, K) = \sum_{\mathbf{i}} \left(h_{\mathbf{i}} \log \frac{h_{\mathbf{i}}}{m_{\mathbf{i}}} + k_{\mathbf{i}} \log \frac{k_{\mathbf{i}}}{m_{\mathbf{i}}} \right),$$

where $m_{\mathbf{i}} = \frac{h_{\mathbf{i}}+k_{\mathbf{i}}}{2}$.

χ^2 *Statistics*:

$$d_{\chi^2}(H, K) = \sum_{\mathbf{i}} \frac{(h_{\mathbf{i}} - m_{\mathbf{i}})^2}{m_{\mathbf{i}}},$$

where again $m_{\mathbf{i}} = \frac{h_{\mathbf{i}}+k_{\mathbf{i}}}{2}$. This distance measures how unlikely it is that one distribution was drawn from the population represented by the other.

These dissimilarity definitions can be appropriate in different areas. For example, the Kullback-Leibler divergence is justified by information theory and the χ^2 statistics by statistics. However, these measures do not necessarily match perceptual similarity well. Their major drawback is that they account only for the correspondence between bins with the same index, and do not use information across bins. This problem is illustrated in Fig. 1(a) which shows two pairs of one-dimensional gray-scale histograms. For instance, the L_1 distance between the two histograms on the left is larger than the L_1 distance between the two histograms on the right, in contrast to perceptual dissimilarity. The desired distance should be based on correspondences between bins in the two histograms and on the ground distance between them as shown in part (c) of the figure.

Another drawback of bin-by-bin dissimilarity measures is their sensitivity to bin size. A binning that is too coarse will not have sufficient discriminative power, while a binning that is too fine will place similar features in different bins which will never be matched. On the other hand, cross-bin dissimilarity measures, described next, always yield better results with smaller bins.

2.2. *Cross-Bin Dissimilarity Measures*

When a ground distance that matches perceptual dissimilarity is available for single features, incorporating this information results in perceptually more meaningful dissimilarity measures.

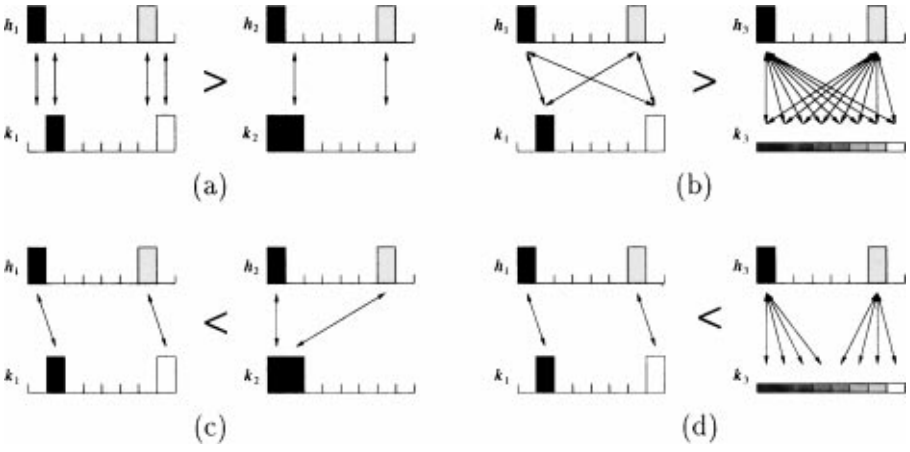


Figure 1. Examples where the L_1 distance (as a representative of bin-by-bin dissimilarity measures) and the quadratic-form distance do not match perceptual dissimilarity. Assuming that histograms have unit mass, (a) $d_{L_1}(\mathbf{h}_1, \mathbf{k}_1) = 2$, $d_{L_1}(\mathbf{h}_2, \mathbf{k}_2) = 1$. (b) $d_A(\mathbf{h}_1, \mathbf{k}_1) = 0.1429$, $d_A(\mathbf{h}_3, \mathbf{k}_3) = 0.0893$. Perceptual dissimilarity is based on correspondence between bins in the two histograms. Figures (c) and (d) show the desired correspondences for (a) and (b) respectively.

Quadratic-form distance: this distance was suggested in Niblack et al. (1993) for color based retrieval:

$$d_A(H, K) = \sqrt{(\mathbf{h} - \mathbf{k})^T \mathbf{A} (\mathbf{h} - \mathbf{k})} \ ,$$

where $\mathbf{h}$ and $\mathbf{k}$ are vectors that list all the entries in H and K. Cross-bin information is incorporated via a similarity matrix $\mathbf{A} = [a_{ij}]$ where a_{ij} denote similarity between bins i and j. Here i and j are sequential (scalar) indices into the bins.

For our experiments, we followed the recommendation in Niblack et al. (1993) and used $a_{ij} = 1 - d_{ij}/d_{max}$ where d_{ij} is the ground distance between bins i and j of the histogram, and $d_{max} = \max(d_{ij})$. Although in general the quadratic-form is not a metric, it can be shown that with this choice of $\mathbf{A}$ the quadratic-form is indeed a metric.

The quadratic-form distance does not enforce a one-to-one correspondence between mass elements in the two histograms: The same mass in a given bin of the first histogram is simultaneously made to correspond to masses contained in different bins of the other histogram. This is illustrated in Fig. 1(b) where the quadratic-form distance between the two histograms on the left is larger than the distance between the two histograms on the right. Again, this is clearly at odds with perceptual dissimilarity. The desired distance here should be based on the correspondences shown in part (d) of the figure.

Similar conclusions were obtained in Stricker and Orengo (1995) where it was shown that using the quadratic-form distance in image retrieval results in false positives, because it tends to overestimate the mutual similarity of color distributions without a pronounced mode.

Match distance:

$$d_M(H, K) = \sum_i |\hat{h}_i - \hat{k}_i| \ ,$$

where $\hat{h}_i = \sum_{j \le i} h_j$ is the cumulative histogram of $\{h_i\}$, and similarly for $\{k_i\}$.

The match distance (Shen and Wong, 1983; Werman et al., 1985) between two one-dimensional histograms is defined as the L_1 distance between their corresponding cumulative histograms. For one-dimensional histograms with equal areas, this distance is a special case of the EMD which we present in Section 4 with the important differences that the match distance cannot handle partial matches, or handle other ground distances. The match distance does not extend to higher dimensions because the relation $\mathbf{j} \le \mathbf{i}$ is not a total ordering in more than one dimension, and the resulting arbitrariness causes problems.

Kolmogorov-Smirnov distance:

$$d_{KS}(H, K) = \max_i(|\hat{h}_i - \hat{k}_i|).$$

Again, $\hat{h}_i$ and $\hat{k}_i$ are cumulative histograms.

The Kolmogorov-Smirnov distance is a common statistical measure for unbinned distributions. Similarly to the match distance, it is defined only for one dimension.

2.3. *Parameter-Based Dissimilarity Measures*

These methods first compute a small set of parameters from the histograms, either explicitly or implicitly, and then compare these parameters. For instance, in Stricker and Orengo (1995) the distance between distributions is computed as the sum of the weighted distances of the distributions' first three moments. In Das et al. (1997), only the peaks of color histograms are used for color image retrieval. In Liu and Picard (1996), textures are compared based on measures of their periodicity, directionality, and randomness, while in Manjunath and Ma (1996) texture distances are defined by comparing their means and standard deviations in a weighted-L_1 sense.

Additional dissimilarity measures for image retrieval are evaluated and compared in Smith (1997) and Puzicha et al. (1997).

3. Histograms vs Signatures

In Section 2 we defined a histogram as deriving from a fixed partitioning of the domain of a distribution. Of course, even if bin sizes are fixed, they can be different in different parts of the underlying feature space. Even so, however, for some images often only a small fraction of the bins contain significant information, while most others are hardly populated. A finely quantized histogram is highly inefficient in this case. On the other hand, for images that contain a large amount of information, a coarsely quantized histogram would be inadequate. In brief, because histograms are fixed-size structures, they cannot achieve a good balance between expressiveness and efficiency.

A *signature* $\{\mathbf{s}_j = (\mathbf{m}_j, w_{\mathbf{m}_j})\}$, on the other hand, represents a set of feature clusters. Each cluster is

represented by its mean (or mode) $\mathbf{m}_j$, and by the fraction $w_{\mathbf{m}_j}$ of pixels that belong to that cluster. The integer subscript j ranges from one to a value that varies with the complexity of the particular image. While j is simply an integer, the representative $\mathbf{m}_j$ is a d-dimensional vector. In general, vector quantization algorithms (Nasrabad and King, 1988) can be used for the clustering, as long as they are applied on every image independently, and they adjust the number of clusters to the complexities of the individual images. For image retrieval, where the number of images is large, we derived a fast clustering algorithm described in Section 5.1.

Since the definition of cluster is open, a histogram $\{h_{\mathbf{i}}\}$ can be viewed as a signature $\{\mathbf{s}_j = (\mathbf{m}_j, w_{\mathbf{m}_j})\}$ in which the vectors $\mathbf{i}$ index a set of clusters defined by a fixed *a priori* partitioning of the underlying space. If vector $\mathbf{i}$ maps to cluster j, the point $\mathbf{m}_j$ is the central value in bin $\mathbf{i}$ of the histogram, and w_j is equal to $h_{\mathbf{i}}$.

We show in Section 5.1 that representing the content of an image database by signatures leads to better results for queries than with histograms. This is the case even when the signatures contain on the average significantly less information than the histograms. By "information" here we refer to the minimal number of bits needed to store the signatures and the histograms.

4. The Earth Mover's Distance

The ground distance between two single perceptual features can often be found by psychophysical experiments. For example, perceptual color spaces were devised in which the Euclidean distance between two single colors approximately matches human perception of their difference. This becomes more complicated when sets of features, rather than single colors, are being compared. In Section 2 we showed the problems caused by dissimilarity measures that do not handle correspondences between different bins in the two histograms. This correspondence is key to a perceptually natural definition of the distances between sets of features. This observation led to distance measures based on bipartite graph matching (Peleg et al., 1989; Zikan, 1990), defined as the minimum cost of matching elements between the two histograms.

In Peleg et al. (1989) the distance between two gray-scale images is computed as follows: every pixel is represented by n "pebbles" where n is an integer representing the gray level of that pixel. After normalizing the two images to have the same number of pebbles, the distance between them is computed as the minimum cost of matching the pebbles between the two images. The cost of matching two single pebbles is based on their distance in the image plane. In this section we adapt this idea to produce the *Earth Mover's Distance* (EMD), a useful metric between signatures for image retrieval in different feature spaces. The main differences between the two approaches are that we solve the transportation problem in contrast to the matching problem. This significantly increases the efficiency due to the ability to cluster pixels in the feature space and to transport together large chunks of "mass", and leads to implementations that are fast enough for on-line image retrieval systems. In addition, as we show, our formulation allows for partial matches, which are important for image retrieval applications. Finally, instead of computing image distances based on the cost of moving pixels in the image space, we are computing the distances in other feature spaces where the ground distances can be perceptually better defined.

Intuitively, given two distributions, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. Then, the EMD measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of ground distance.

Computing the EMD is based on a solution to the well-known *transportation problem* (Hitchcock, 1941) a.k.a. the Monge-Kantorovich problem which goes back to 1781 when it was first introduced by Monge (Rachev, 1984) Suppose that several *suppliers*, each with a given amount of goods, are required to supply several *consumers*, each with a given limited capacity. For each supplier-consumer pair, the cost of transporting a single unit of goods is given. The transportation problem is then to find a least-expensive flow of goods from the suppliers to the consumers that satisfies the consumers' demand. Signature matching can be naturally cast as a transportation problem by defining one signature as the supplier and the other as the consumer, and by setting the cost for a supplier-consumer pair to equal the ground distance between an element in the first signature and an element in the second. Intuitively, the solution is then the minimum amount of "work" required to transform one signature into the other.

This can be formalized as the following linear programming problem: Let $P = \{(\mathbf{p}_1, w_{\mathbf{p}_1}), \ldots, (\mathbf{p}_m, w_{\mathbf{p}_m})\}$ be the first signature with m clusters, where $\mathbf{p}_i$ is the cluster representative and $w_{\mathbf{p}_i}$ is the weight of the cluster; $Q = \{(\mathbf{q}_1, w_{\mathbf{q}_1}), \ldots, (\mathbf{q}_n, w_{\mathbf{q}_n})\}$ the second

signature with n clusters; and $\mathbf{D} = [d_{ij}]$ the ground distance matrix where d_{ij} is the ground distance between clusters $\mathbf{p}_i$ and $\mathbf{q}_j$.

We want to find a flow $\mathbf{F} = [f_{ij}]$, with f_{ij} the flow between $\mathbf{p}_i$ and $\mathbf{q}_j$, that minimizes the overall cost

$$\text{WORK}(P, Q, \mathbf{F}) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij},$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m,\ 1 \leq j \leq n \quad (1)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \quad 1 \leq i \leq m \quad (2)$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j} \quad 1 \leq j \leq n \quad (3)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}\right), \quad (4)$$

Constraint (1) allows moving "supplies" from P to Q and not vice versa. Constraint (2) limits the amount of supplies that can be sent by the clusters in P to their weights. Constraint (3) limits the clusters in Q to receive no more supplies than their weights; and constraint (4) forces to move the maximum amount of supplies possible. We call this amount the *total flow*. Once the transportation problem is solved, and we have found the optimal flow $\mathbf{F}$, the earth mover's distance is defined as the resulting work normalized by the total flow:

$$\text{EMD}(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}},$$

The normalization factor is the total weight of the smaller signature, because of constraint (4). This factor is needed when the two signatures have different total weight, in order to avoid favoring smaller signatures. In general, the ground distance d_{ij} can be any distance and will be chosen according to the problem at hand. Examples are given in Section 5.

Thus, the EMD naturally extends the notion of a distance between single elements to that of a distance between sets, or distributions, of elements. The advantages of the EMD over previous definitions of distribution distances should now be apparent. First, the EMD applies to signatures, which subsume histograms as shown in Section 3. The greater compactness and flexibility of signatures is in itself an advantage, and having a distance measure that can handle these variable-size structures is important. Second, the cost of moving "earth" reflects the notion of nearness properly, without the quantization problems of most current measures. Even for histograms, in fact, items from neighboring bins now contribute similar costs, as appropriate. Third, the EMD allows for partial matches in a very natural way. This is important, for instance, in order to deal with occlusions and clutter in image retrieval applications, and when matching only parts of an image. Fourth, if the ground distance is a metric and the total weights of two signatures are equal, the EMD is a true metric, which allows endowing image spaces with a metric structure. A proof of this is given in Appendix A.

Of course, it is important that the EMD can be computed efficiently, especially if it is used for image retrieval systems where a quick response is required. Fortunately, efficient algorithms for the transportation problem are available. We used the transportation-simplex method (Hillier and Lieberman, 1990), a streamlined simplex algorithm that exploits the special structure of the transportation problem. A good initial basic feasible solution can drastically decrease the number of iterations needed. We compute the initial basic feasible solution by Russell's method (Russell, 1969).

A theoretical analysis of the computational complexity of the transportation simplex is hard, since it is based on the simplex algorithm which can have, in general, an exponential worst case (Klee and Minty, 1972). However, in practice, because of the special structure in our case and the good initial solution, the performance is much better. We empirically measure the time-performance of our EMD implementation by generating random signatures of sizes that range from 1 to 100. For each size we generate 100 pairs of random signatures and record the average CPU time for computing the EMD between the pairs. The results are shown in Fig. 2. This experiment was done on a SGI Indigo 2 with a 195 MHz CPU.

Other efficient methods to solve the transportation problem include interior-point algorithms (Karmarkar, 1984) which have polynomial time complexity, and by formalizing the transportation as the uncapacitated minimum cost network flow problem (Ahuja et al., 1993), it can be solved in our case of bipartite graph in $O(n^3 \log n)$, where n is the number of clusters in the

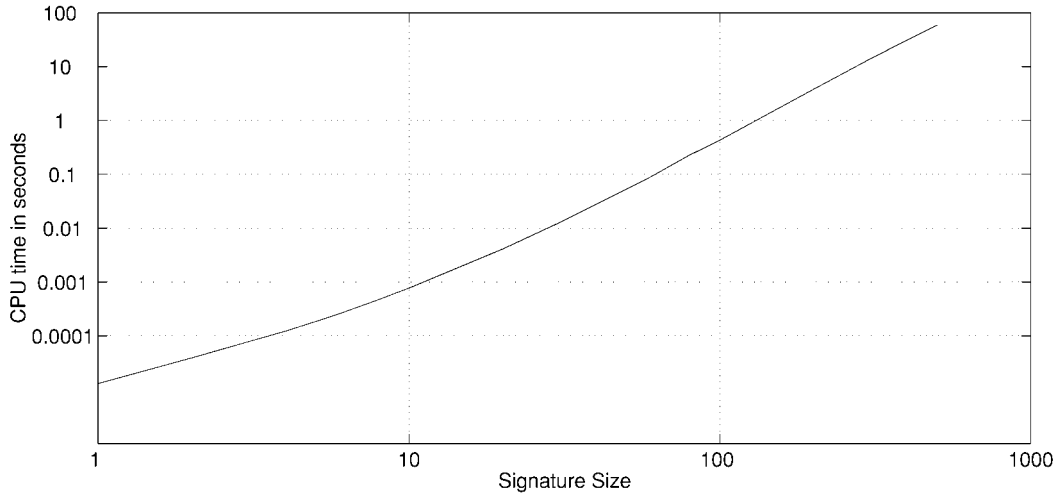


Figure 2. A log-log plot of the average computation time for random signatures as a function of signature size.

signatures (This bound assumes that the two signatures have the same size, and that the precision of the calculations is fixed and can be considered as a constant.) This is similar to the time complexity of our algorithm as can be inferred from the plot in Fig. 2.

Retrieval speed can be increased if lower bounds to the EMD can be computed at a low expense. These bounds can significantly reduce the number of EMDs that actually need to be computed by prefiltering the database and ignoring images that are too far from the query. An easy-to-compute lower bound for the EMD between signatures with equal total weights is the distance between their centers of mass, as long as the ground distance is induced by a norm. A proof of this is given in appendix B, along with the definition of norm-induced distance. Using this lower bound in our color-based image retrieval system significantly reduced the number of EMD computations. Figure 3 shows the average number of EMD computations per query as a function of the number of images retrieved. This graph was generated by averaging over 200 random queries on an image database with 20,000 images

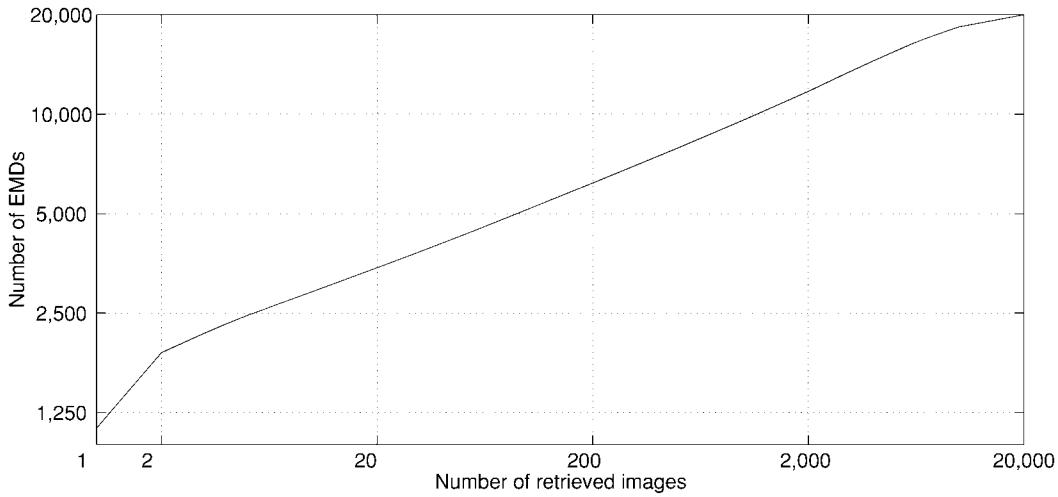


Figure 3. A log-log plot of the number of EMDs as a function of the number of images retrieved. The database contains 20,000 images.

using the color-based image retrieval system described in Section 5.1. The fewer the images that are returned by a query, the fewer the EMDs that need to be compared thanks to our lower bound, which guarantees that no image is missed as a result of the saving in computation.

A computational advantage of signatures over histograms is that distributions defined in high-dimensional feature spaces can be matched more efficiently. This is because the only computational factor is the number of significant clusters in the distributions and not the dimension of the underlying space, although sometimes the two correlate. The EMD is also robust to the clustering algorithm that is used to find the significant clusters. If a cluster in some signature is split into smaller fragments, the EMD will consider them as similar.

While the EMD works very well on signatures, it should not, in general, be applied to histograms. Small histogram invalidate the ground distance as the bin centers are rather far, while computing the EMD on large histograms can be very slow.

5. Examples

In this section we show a few examples of application of the earth mover's distance in the areas of color and texture analysis. Because of how the human vision system is built, color lives naturally in a three dimensional space. Color distributions, then, can describe the color contents of entire images. A color example is given in Section 5.1. Combining the color of the pixels together with their position in the image leads to a distance that considers the layout similarity together with the color similarity of the images. This is discussed in Section 5.2.

For texture, the situation is more complex. A texture can be described locally as a mixture of two-dimensional sinusoidal signals at different scales and orientations. Thus, the responses of a bank of filters, centered at a pixel, can be seen as a distribution of signal energy and phase in the frequency domain, which is the space of all two-dimensional sinusoidal signals. In keeping with most of the literature on texture, we ignore phase information. At a higher level, the texture content of a full image that might contain multiple textures can be seen as a distribution of such two-dimensional distributions. Defining a ground distance between the local representations of texture leads to an EMD between images of textures. Examples of distance computations between images with multiple textures are given in Section 5.3.

For both color and texture our ground distance is

$$d_{ij} = 1 - e^{-\alpha \|\mathbf{p}_i - \mathbf{q}_j\|}, \tag{5}$$

where $\|\cdot\|$ is an appropriate L_p-norm which we choose differently for color and texture, and α distinguishes between "close" and "far" distances in the feature space. We used

$$\alpha = \|[\sigma_1 \ldots \sigma_{dim}]^T\|,$$

where σ_i is the standard deviation of the i-th dimension components of the features from the overall distribution of all images in the database. *dim* is the number of dimensions in the feature space.

This ground distance has the property that for large distances it saturates to 1. This limits the effect that few, very different, features can have on the distance between overall similar distributions. As we show in appendix C this ground distance is metric and therefore the resulting EMD is metric for signatures of equal weights.

5.1. Color Distributions

To compute of the earth mover's distance between color images, we first convert the distribution of pixel colors to the CIE-Lab color space (Wyszecki and Stiles, 1982) which was expressly designed so that short Euclidean distances correlate strongly with human color discrimination performance, albeit for pairs of colors on a neutral background (recall the discussion of this point in the introduction). The L_2-norm is therefore a natural choice for the ground distance in Eq. (5).

We performed our color-based retrieval on a collection of 20,000 color images from the Corel Stock Photo Library. To compute the signature of a color image, we first slightly smooth each band of the image's RGB representation in order to reduce possible color quantization and dithering artifacts. We then transform the image into the CIE-Lab color space using D65 as the reference white (Poynton, 1996). At this point each image implies a distribution of points in the three-dimensional CIE-Lab color space where a point corresponds to a pixel in the image. We coalesce this distribution into clusters of similar colors (25 units in any of the L, a, b axes). Because of the large number of images to be processed in typical database applications,

clustering must be performed efficiently. To this end, we devised a novel two-stage algorithm based on a *k-d* tree (Bentley, 1975) where the splitting rule is to simply divide an interval into two equal sub-intervals. In the first phase, approximate clusters are found by excessive subdivisions stopping when the cells become smaller than the allowed cluster size. Since by this method clusters might be split over few cells, we use a second phase in order to recombine them. This is done by performing another *k-d* tree clustering of the cluster centroids from the first phase, after shifting the space coordinates by one half of the minimal allowed cell size (25 units). Each new cluster contributes a pair $(\mathbf{p}, w_{\mathbf{p}})$ to the signature representation of the image where $\mathbf{p}$ is the average color of the cluster, and the corresponding weight $w_{\mathbf{p}}$ is the fraction of image pixels that are in that cluster. At this point, we remove clusters with insignificant weights (less than 0.1%). In our database, the average signature has 8.8 clusters, which leads to typical query times of a few seconds.

The difficulty of establishing ground truth makes it hard to evaluate the performance of an image retrieval system. To evaluate the precision of a query, all the images which are perceived to have similar color content to the query should be taken into account. Evaluating the performance of retrieval systems is beyond the scope of this paper, as our goal is rather to compare the EMD to the other dissimilarity measures described in Section 2. For that purpose we conducted two sets of experiments where we created common ground truths on which we measured the performance of the different methods.

In our first experiment, we randomly chose 94 images from our database. We used the same number of images as in the texture case (see Section 5.3), so that we can compare the results in both cases. From each image we created disjoint sets of randomly sampled pixels and considered 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 to the query have to be identified. We used a set size of 8 pixels, and obtained for each image 16 disjoint sets of random samples, resulting in a ground truth data set of 1504 samples with 94 different classes, one class per image. We represented every sample in this database by a histogram with 128 bins which were adapted to the overall distribution in the database. To construct the histograms, we first ran a k-means algorithm on the combined distribution of all the samples in the database, resulting in the optimal 128 prototypes. Each pixel in every set was then assigned to the bin represented by the closest prototype. Now we used each of the 1504 samples as our query, asking all dissimilarity measures to retrieve and rank the most similar samples in the database. We averaged the results of all the 1504 queries. Figure 4

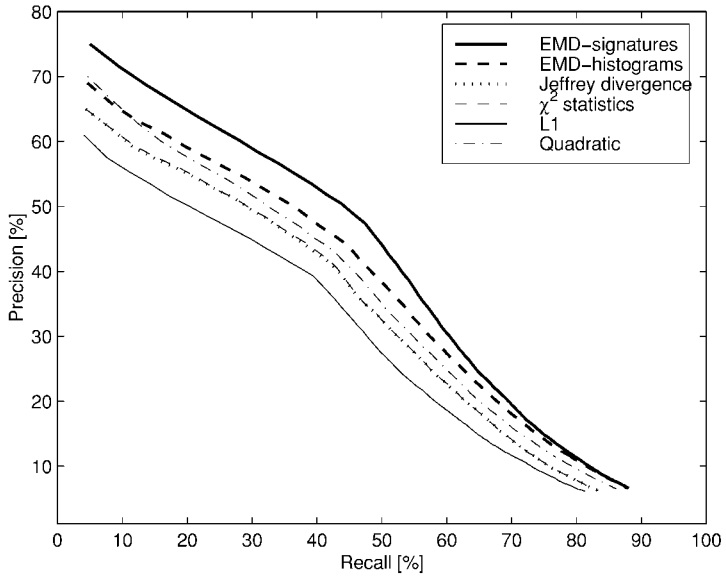


Figure 4. Precision vs. recall for color distributions. The database consists of 1504 samples, divided into 94 classes of 16 similar samples. All dissimilarity measures use histograms with 128 bins, except for the EMD which was applied also to signatures that have only 8 clusters.

shows the precision (retrieved and relevant/total retrieved) vs. recall (retrieved and relevant/total relevant). We also display in the graph the result of applying the EMD on signatures. Unlike the histograms that had 128 bins, the signatures had only 8 clusters. The EMD performs best even with the much smaller representation.

The second set of experiments was conducted on our full 20,000 image database. Unlike the first experiment were we used adaptive histograms, here we use histograms with regular binning. This is typical to image retrieval systems, where due to the large number of images, a specific image cannot take advantage of such adaptive histograms. Also, all the images are usually not available when the database is created. We run this experiment twice, once on color histograms with coarse binning, and once with fine binning. For the coarse binning, we divided the CIE-Lab color space into fixed-size bins of size $25 \times 25 \times 25$. This quantized the color space into 4 bins in the L channel and 8 bins in both the a and the b channels, for a total of 256 bins. However, most of these bins are always empty due to the fact that valid RGB colors can map only to a subset of this CIE-Lab space. In fact, only 130 bins can have non-zero values. Our histograms then have 130 bins. After thresholding away bins with insignificant weights (less than 0.1%), the average histogram has 15.3 non-zero bins. Notice that the amount of information contained in the signatures (8.8 clusters in average) is comparable to that contained in the histograms. For the fine binning, we divided the CIE-Lab color space into fixed-size bins of size $12.5 \times 12.5 \times 12.5$. This resulted in a total of 2048 bins of which only 719 can possibly have non zero values. Over our 20,000-image database the average fine histogram has 39 non-zero bins. Clearly, the amount of information in the average signature is now much smaller than that in these finer histograms. Again, we see that even with less information, signatures result in better retrieval precision than histograms.

For the EMD and the Quadratic Form, instead of using the ground distance that we used for the previous experiment (Eq. (5)), we simply use the Euclidean distance in the CIE-Lab color space. In addition to being faster to compute, we found that for real images, the Euclidean distance leads to better recall which is most important for retrieval system. Being induced by a norm, using the Euclidean distance also allows us to use the lower bound described in Section 4 which significantly reduces the computation needed.

Our goal in this experiment was to compare the different dissimilarity measures on images that are perceived as having similar color content. To do so, we looked for sets of images with high correlation between the semantic meaning of the images and their color distribution. For the first set, we identified all the images of red cars in the database (75 images) and marked them as relevant. From this set we chose the ten images that are shown in Fig. 5(a). In these ten images the red car had a Green/Gray background, was relatively big and not obscured by the background (for example, using an image with a small red car in front of a sunset is likely to return images of sunsets rather that images of red cars). We performed ten queries using a different "good" car every time, and averaged the number of relevant images for the different dissimilarity measures as a function of the number of images. An example of such a query is shown in Fig. 6. 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 histogram dissimilarity measures. For the second set, we similarly identified 157 images of brown horses in green fields. Again 10 "good" images of horses (Fig. 5(b)) were used for the query.

The results of these experiments are shown in Figs. 7 and 8. For the cars, the average number of relevant images for the different dissimilarity measures as a function of the number of images retrieved is shown in Fig. 7(a) and 7(b) for the coarse and fine histograms respectively. The EMD that was computed on the histograms outperformed the other histogram-based methods, and the EMD that was computed on the signatures performed best. Here, instead of a precision vs. recall graph, we only show the precision of the queries as our goal is to compare the different measures and not to evaluate the performance of a specific retrieval system. The colors of the cars are very similar in all the relevant images while the colors of the backgrounds have more variation. Although other images that do not have cars in them might match the color contents of the query images better, we still expect some of the cars to be retrieved when a large number of images is returned by the system.

For the horses, both the colors of the objects and the colors of the backgrounds are similar for all the relevant images. Figure 8 shows the results for the coarse and fine histograms respectively. Here again the EMD that was computed on the signatures performed

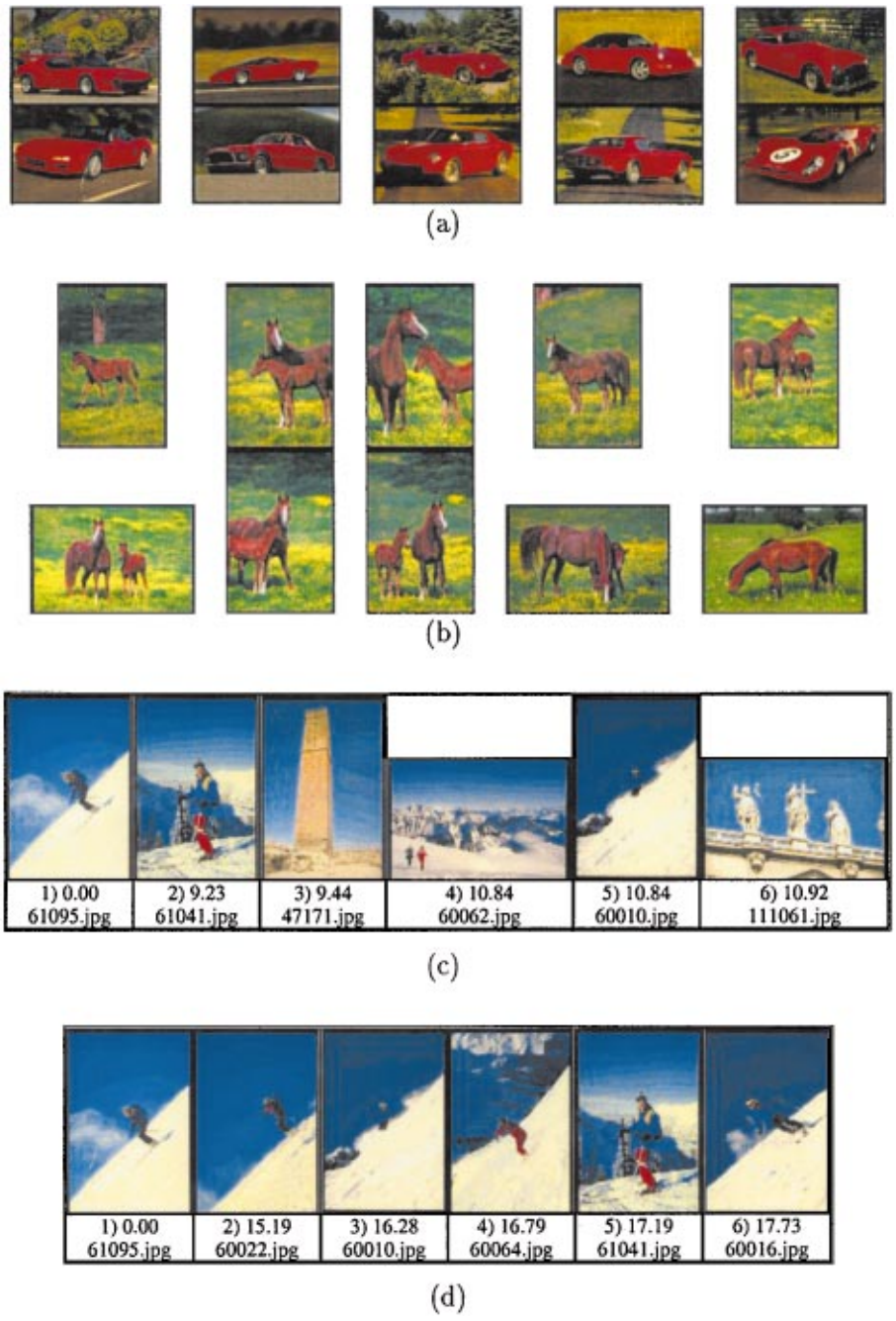


Figure 5. (a) Ten images of red cars used for the retrieval results in Fig. 7. (b) Ten images of horses used for the retrieval results in Fig. 8(c, d): The six best matches without position information (c) and with position information (d) when using the leftmost image of a skier as the query.

best. Among the histogram-based methods, in the experiment that used the coarse histograms, both the Jeffrey divergence and the χ^2 statistics outperformed the EMD that was computed on the histograms. In the experiment that used the fine histograms, the EMD outperformed all the other measures. This can be explained by the fact that, for coarser histograms, the ground distance is computed between more distant bin

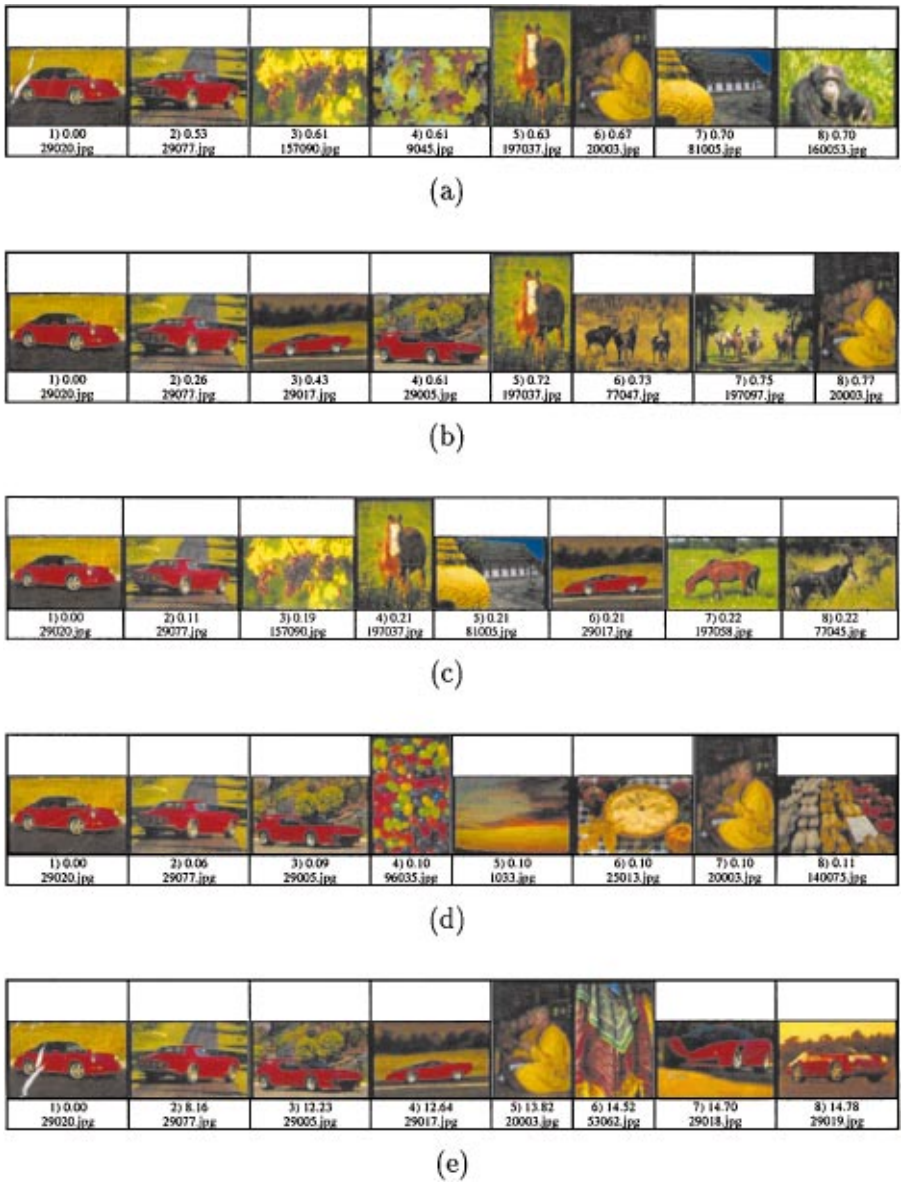
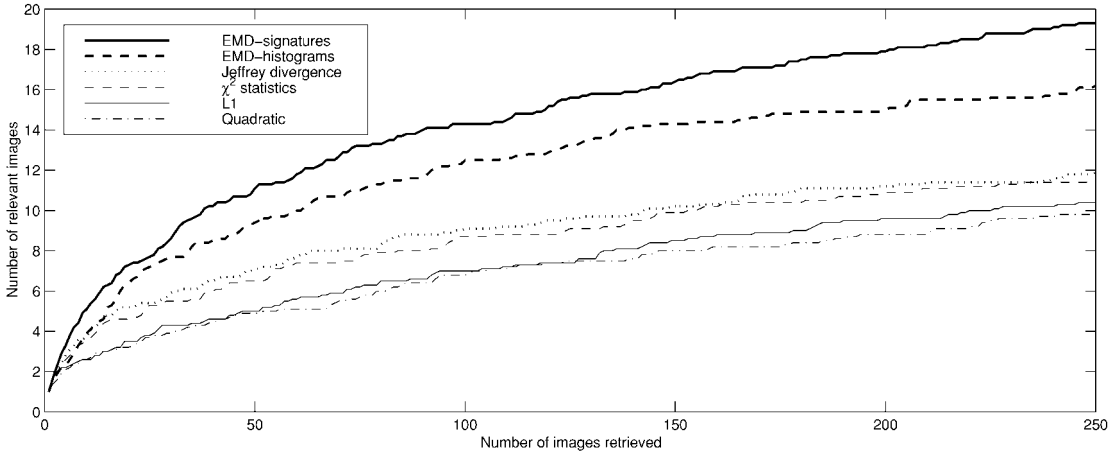


Figure 6. The eight closest images to the leftmost image of a red car. The queries were processed by a color-based image retrieval system using different histogram dissimilarity measures. (a) L_1 distance. (b) Jeffrey divergence. (c) χ^2 statistics. (d) Quadratic-form distance. (e) EMD.

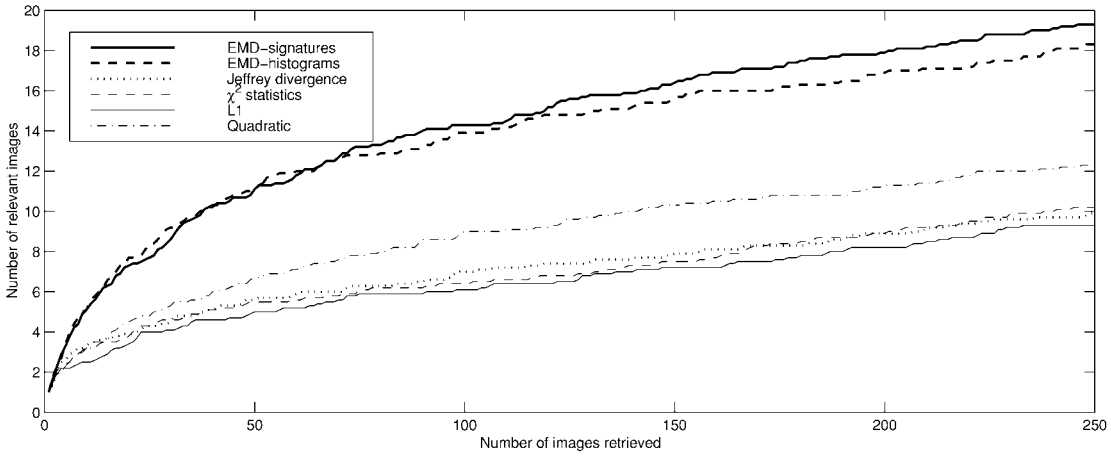
centers, and therefore becomes less meaningful. We recall that only small Euclidean distances in CIE-Lab space are perceptually meaningful. On the other hand, bin-by-bin distances break down as the histograms get finer, because similar features are split among different bins.

5.2. Joint Distribution of Color and Position

In many cases, global color distributions that ignore the actual positions of the colors in the image are not sufficient for good retrieval. For example, consider the following two color images: In the first, there are blue



(a)



(b)

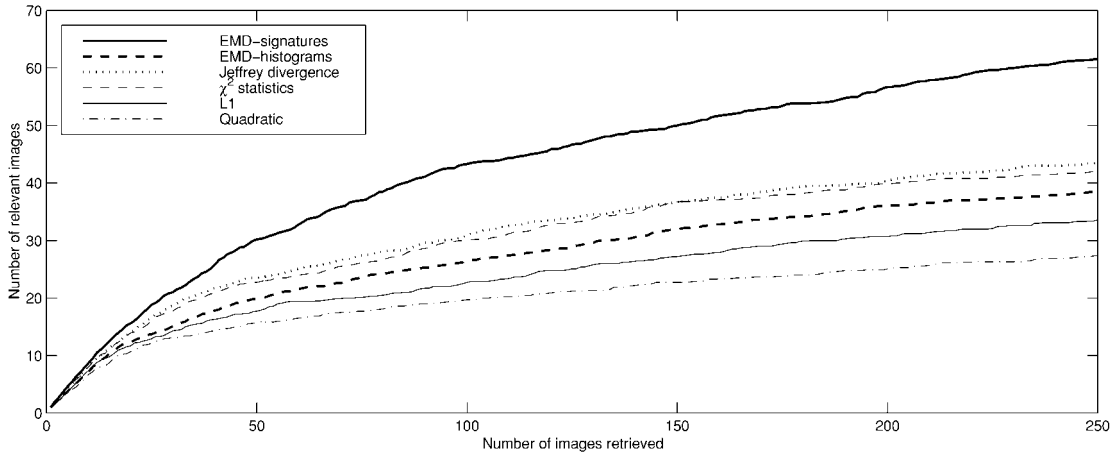
Figure 7. The average number of relevant images, for the different dissimilarity measures, that were returned by using the ten car images in Fig. 5(a) as the queries for the coarse histograms (a) and fine (b) histograms. The results obtained by using signatures is also shown in the two graphs for reference.

skies *on top* of a green field, while in the other there is a blue lake *below* green tree-tops. Although the color distributions might be very similar, the position of the colors in the image is very different and may have to be taken into account by the query. This can be achieved by modifying the color distance in Section 5.1 as follows: Instead of using the three-dimensional CIE-Lab color space, we use a five-dimensional space whose first three dimensions are the CIE-Lab color space, and the other two are the (x, y) position of each pixel. We normalize the image coordinates to be in the range of 0 to 100, and use the same clustering algorithm as used in Section 5.1. The average signature size in our 20,000 image database is now 18.5.

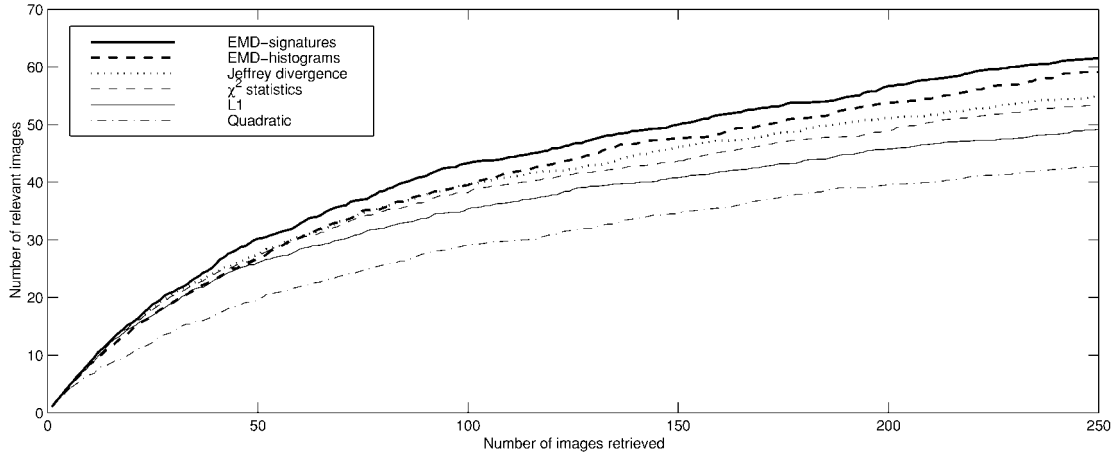
The ground distance is now defined as

$$[(\Delta L)^2 + (\Delta a)^2 + (\Delta b)^2 + \lambda((\Delta x)^2 + (\Delta y)^2)]^{\frac{1}{2}}.$$

The parameter λ defines the importance of the color positions relative to the color values. Figure 5 (c, d) shows the effect of position information where the leftmost image of a skier was used as the query. Part (c) shows the 6 best matches when position information was ignored ($\lambda = 0$). Part (d) uses position information



(a)



(b)

Figure 8. The average number of relevant images, for the different dissimilarity measures, that were returned by using the ten horse images in Fig. 5(b) as the queries for the coarse histograms (a) and fine (b) histograms. The results obtained by using signatures is also shown in the two graphs for reference.

($\lambda = 0.5$). Exact color matches are somewhat compromised in order to get more similar positional layouts.

5.3. Texture

While color is a purely pointwise property of images, texture involves a 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 of a neighborhood of it. Gabor functions are commonly used in texture analysis to capture this information (Bovik et al., 1990; Farrokhnia and Jain, 1991; Manjunath and Ma, 1996) because they are optimally localized in both the spatial and frequency domains (Gabor, 1946). There is also strong evidence that simple cells in the primary visual cortex can be modeled by Gabor functions tuned to detect different orientations and scales on a log-polar grid (Daugman, 1988).

In this paper we used a similar dictionary of Gabor filter as the one derived in Manjunath and Ma (1996) with four scales and six orientations. Applying these Gabor filters to an image results for every image pixel in a four by six array of numbers which can be seen also as a 24 dimensional vector. In order to treat all the Gabor responses from the different scales in a similar

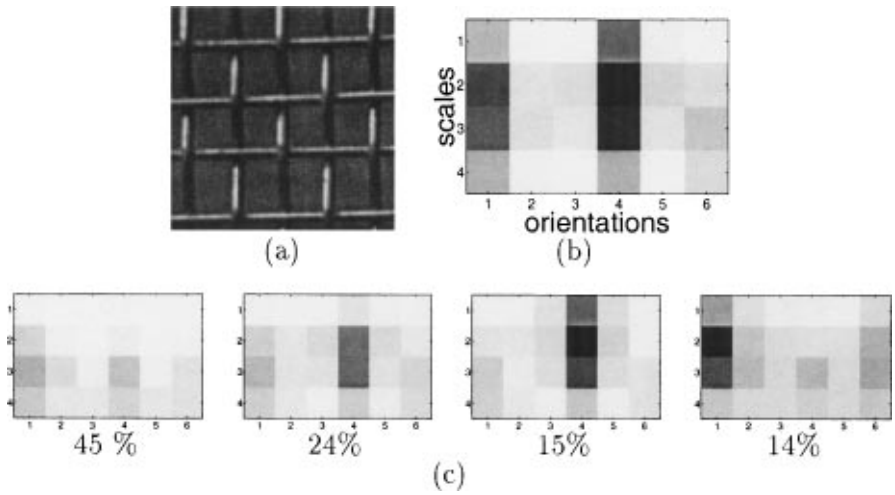


Figure 9. (a) Texture patch from the Brodatz album (Brodatz, 1966). (b) Average over all texture features. The Gabor filter bank consists of four scales and six orientations. (c) The texture signature.

way, we need to appropriately normalize the vector. Unlike (Manjunath and Ma, 1996), who normalizes each feature in the vector by the standard deviation of the respective feature over the entire database, we normalize the feature based on the radial frequency f of the corresponding Gabor filter. This follows (Field, 1987) where it is shown that the magnitude of the power spectrum of natural images falls of as $1/f$. With this normalization, similar amounts of energy will be captured on average by all filters of all scales. In principle, a normalization that is based on the standard deviations requires the knowledge of the entire database and will overemphasize features that are dominated by noise. The normalized texture vector is our *texture feature*. Figure 9 shows an example of a texture feature. Part (b) shows the spatial average of each of the 24 filter responses over the image in part (a) of the figure. Darker squares represent stronger responses. Notice the two strong responses that correspond to the texture's vertical and horizontal components at an intermediate scale.

The texture content of an entire image is represented by a distribution of texture features. In general, this distribution will be simple for images of one uniform texture, and more complex for images with multiple textures such as natural images. To make the representation more compact, we find the dominant clusters in the 24 dimensional space. This is done using the same clustering algorithm described in Section 5.1.

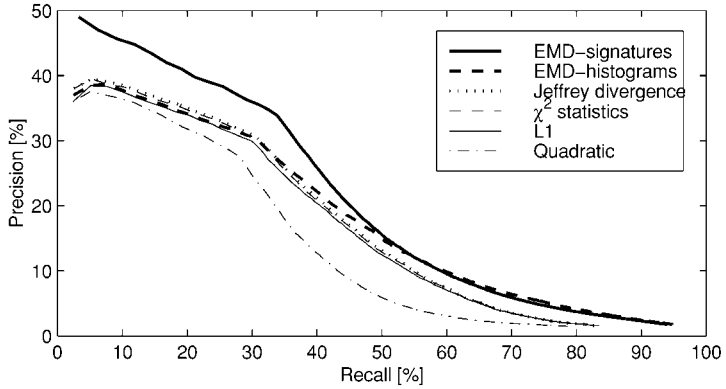


Figure 10. Precision vs. recall for texture distributions. The database consists of 1504 samples, divided into 94 classes of 16 similar samples. All dissimilarity measures use histograms with 128 bins, except for the EMD which uses also signatures with 8 clusters.

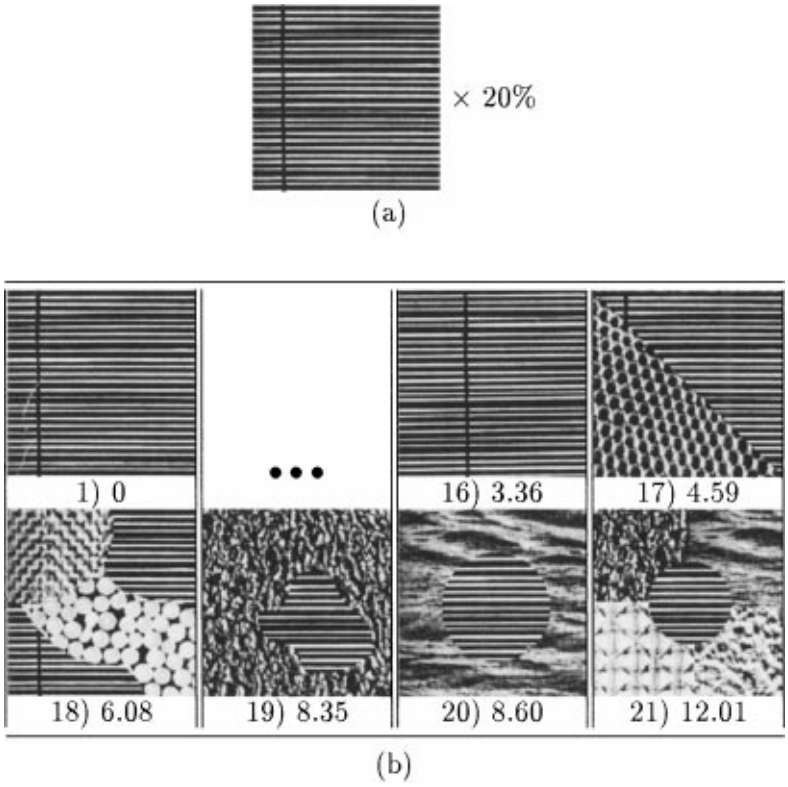


Figure 11. Partial texture query. The query was 20% of the texture in part (a) and 80% "don't care". (b) The 21 best matches: the 16 patches from the same texture (only the first and last ones are shown), followed by all the compositions that contain some part of the queried texture.

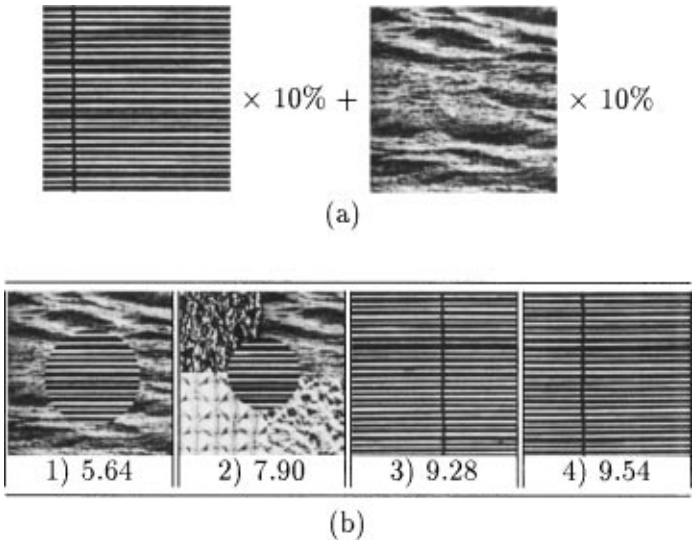


Figure 12. Another partial query. The query now contains 10% of each of the two patches in part (a) and 80% "don't care". (b) The two best matches are the two compositions that contain the textures in the query, followed by the patches that contain only one of the queried textures.

The resulting set of cluster centers together with the cluster weights is the *texture signature*. An example of a texture signature with four clusters is shown in Fig. 9(c) together with the clusters weights.

In order to compare the different dissimilarity measures for texture, we selected 94 Brodatz album (Brodatz, 1966) 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.) We divided each of the textures into 4 by 4 non-overlapping patches. Every patch is 128 by 128 pixels. Similarly to the color case, the database

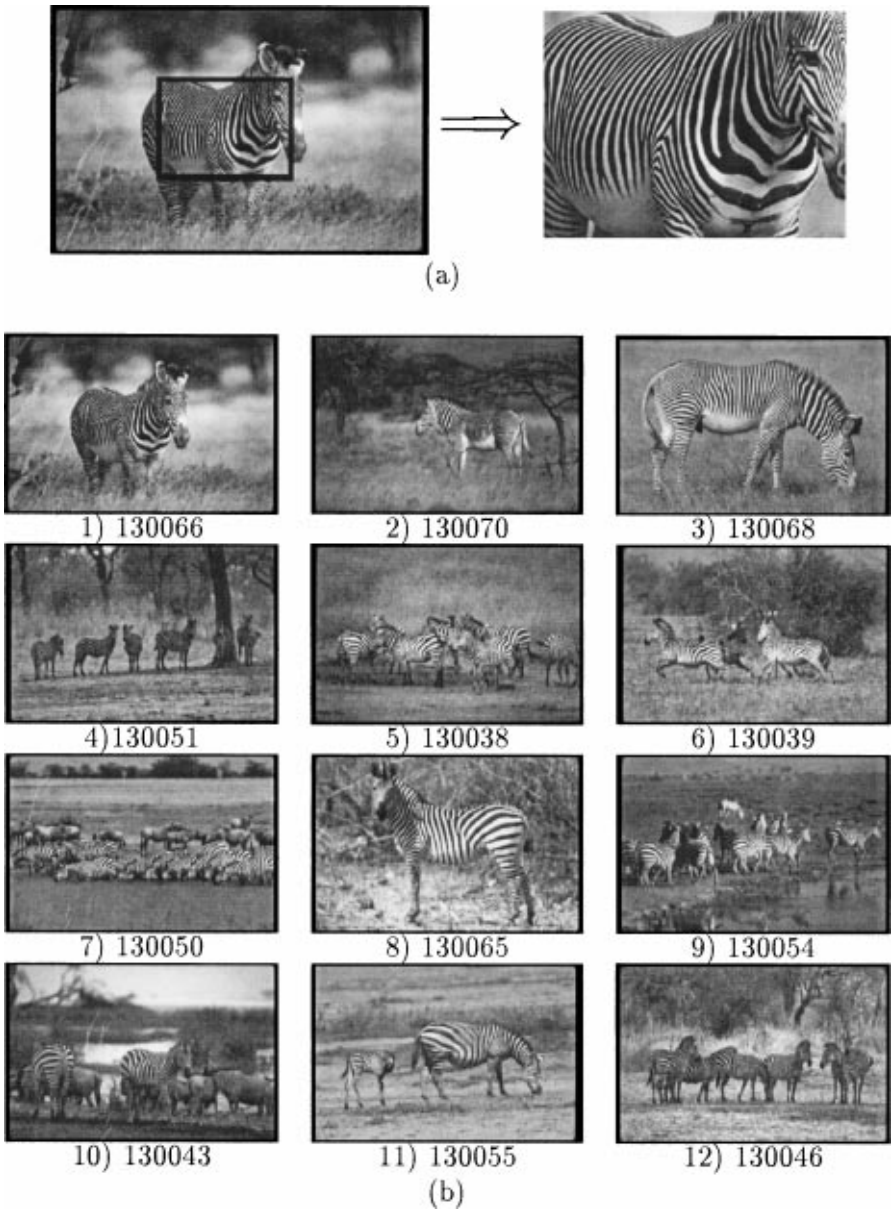


Figure 13. Looking for zebras. (a) An image of a zebra and a block of zebra stripes extracted from it. (b) The twelve best matches to a query asking for images with at least 10% of the texture in (a). The large numbers in the thumbnail captions are indices into Corel CDs. The first three digits ("130" in this case) refer to the same set of 100 images.

contains 1504 texture patches with 94 different classes, each with 16 patches. We used each of the patches in the database as a query, and averaged the results over all the patches. The results of the different dissimilarity measures are shown in Fig. 10. Again, we use globally adapted histograms with 128 bins. For the EMD we use signatures with 8 clusters. We use the ground distance as in Eq. (5), with the L_1-norm, so that the different responses from the different Gabor filters are added together.

An important advantage of the EMD over other measures for texture similarity is its ability to handle images that contain more than one texture without first segmenting the images as needed when using other measures. Using the EMD for partial matches can find images that contain specific textures. Figure 11 shows

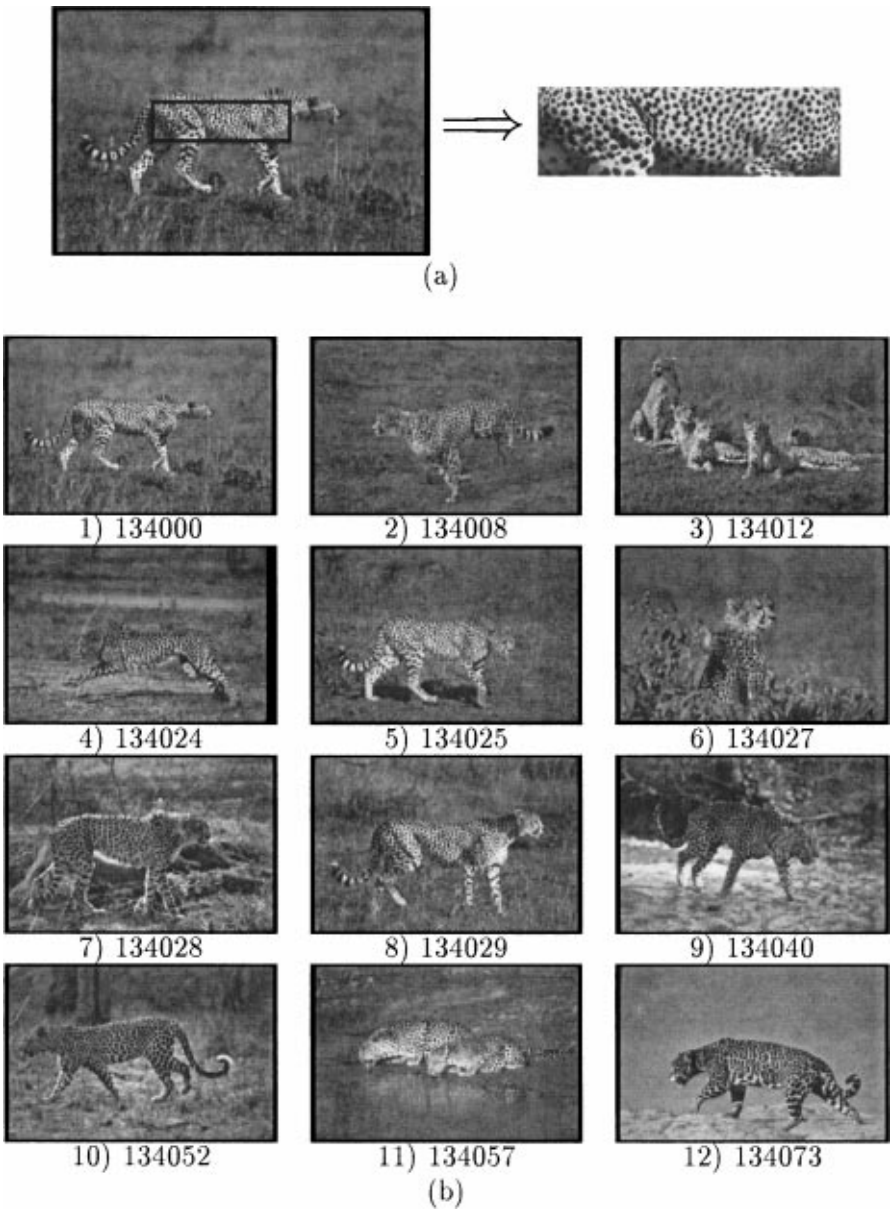


Figure 14. Looking for cheetahs. (a) The query. (b) The twelve best matches with at least 10% of the query texture. The last four images are leopards and jaguars which have similar texture as cheetahs. However, cheetahs come first.

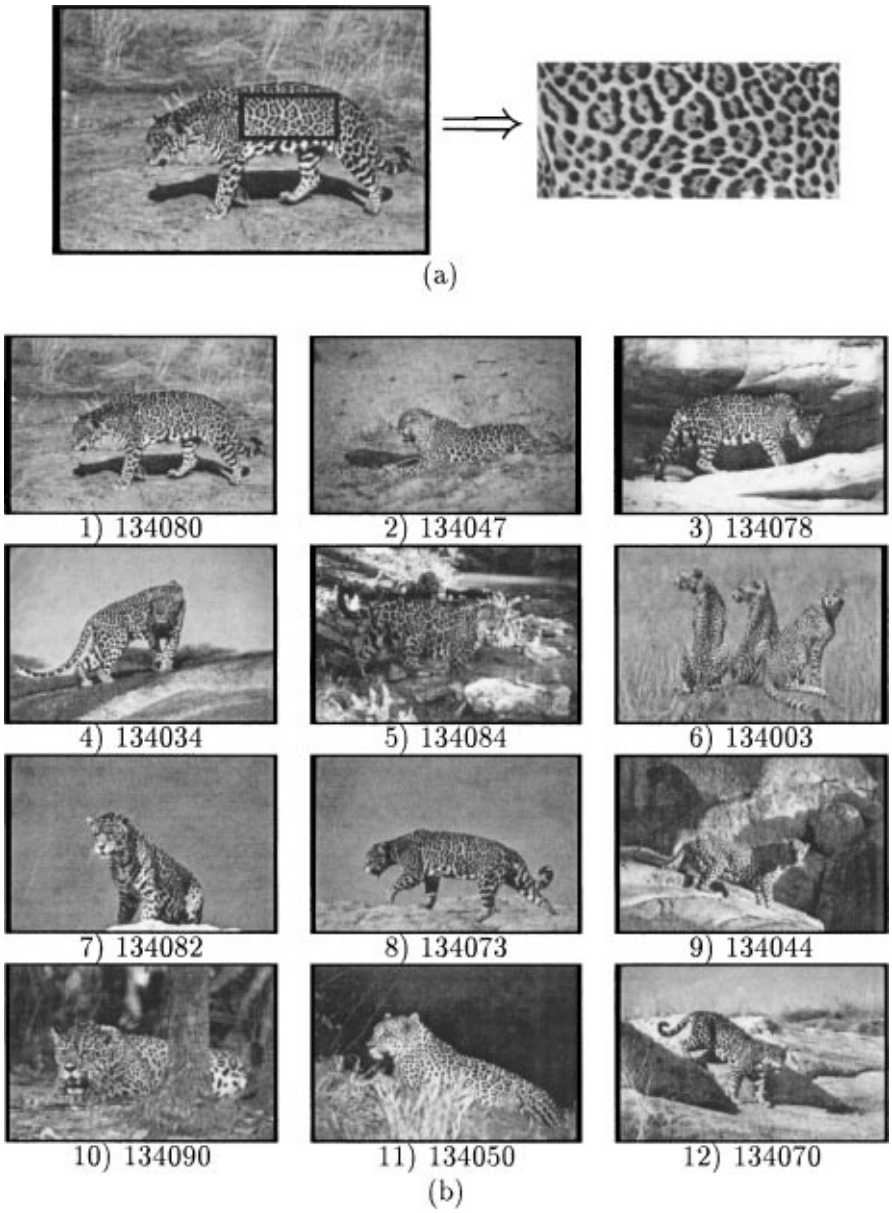


Figure 15. Looking for leopards and jaguars. (a) The query. (b) The twelve best matches with at least 10% of the query texture. All but the sixth image are leopards and jaguars (which are hard to tell apart). The sixth image is of cheetahs.

an example of a partial query. Here we added images with compositions of textures to our texture database. The query was 20% of the texture in part (a) and 80% "don't care". The best matches are shown in part (b) with the 16 patches from the same texture at the beginning followed by all the compositions that contain some part of the queried texture. We emphasize again that no segmentation was performed.

Figure l2 demonstrates a partial query with more than one texture.

In the next experiment we created a database of 500 gray scale images of animals from the Corel Stock Photo Library. This library consists of 20,000 images organized into sets of 100 images each. We created our database using the following sets: 123000 (Backyard Wildlife), 134000 (Cheetahs, Leopards and Jaguars),

130000 (African Specialty Animals), 173000 (Alaskan Wildlife), 66000 (Barnyard Animals). Images are 768 by 512 pixels. We preprocessed the images by the usual clustering procedure, and obtained an average signature size of 32 clusters. Since most of the queries consist of a single, or a few textures, their signatures are significantly smaller and the EMD computation is more efficient.

Figure 13(a) shows an example of a query that used a rectangular patch from an image of a zebra. We asked for images with at least 20% of this texture. The 12 best matches are shown in part (b) ranked by their similarity to the query. The 16 best matches were all images of zebras. The database contains a total of 34 images of zebras. Notice the various backgrounds in the retrieved images. They were ignored by the query because of the EMD's ability to handle partial queries. Notice also that in some of the retrieved images there are a few small zebras, which only when combined together provide a significant amount of "zebra texture". Methods based on segmentation are likely to have problem with such images.

Next we searched for images of cheetahs. The database has 33 images of cheetahs, and 64 more images of leopards and jaguars that have similar texture as cheetahs. Figure 14 shows the query and the best matches. The first eight images are indeed cheetahs. The next four matches are images of leopards and jaguars.

To check if our method can distinguish between different wild cats, we looked for images of jaguars. Figure 15 shows the query results. From the best twelve matches, eleven are jaguars and leopards which are almost indistinguishable. Only the sixth match was an image of a cheetah.

6. Conclusions

The earth mover's distance is a general and flexible metric and has desirable properties for image retrieval. It allows for partial matches, and it can be applied to variable-length representations of distributions. Lower bounds are readily available for it, and it can be computed efficiently, when the signatures are not too large. The EMD should be applied to signatures, not to global histograms, as histograms with few bins will invalidate the ground distances, while EMDs on histograms with many bins will be slow to compute. Because of these advantages, we believe that the EMD can be of use both for understanding distributions related to vision problems, as exemplified by our case studies with color and texture, and as a building block of image retrieval systems.

Our analysis of texture similarity in particular has brought forth a number of interesting open problems. For instance, how can the distance between two signatures be computed if either of them is allowed to undergo a transformation from a predefined group at no cost? An answer to this question would lead to a more direct approach to the issue of invariance when comparing textures or other features.

Finally, it would be interesting to apply the earth mover's distance to other vision problems such as classification and recognition based on other types of visual cues. In addition, we surmise that the EMD may be a useful metric also for problems outside the realm of computer vision.

Appendix A. Metric Proof

In this appendix we prove that when the signatures have equal weights and the ground distance $d(\cdot, \cdot)$ is metric, the EMD is a true metric. Non-negativity and symmetry hold trivially in all cases, so we only need to prove that the triangle inequality holds. Without loss of generality we assume here that the total sum of the flows is 1. Let $\{f_{ij}\}$ be the optimal flow from P to Q and $\{g_{ij}\}$ be the optimal flow from Q to R. Consider the flow $P \mapsto Q \mapsto R$. We now show how to construct a feasible flow from P to R that represents no more work than that of moving mass optimally from P to R through Q. Since the EMD is the least possible amount of feasible work, this construction proves the triangle inequality.

The largest weight that moves as one unit from $\mathbf{p}_i$ to $\mathbf{q}_j$ and from $\mathbf{q}_j$ to $\mathbf{r}_k$ defines a flow which we call b_{ijk} where i, j and k correspond to $\mathbf{p}_i$, $\mathbf{q}_j$ and $\mathbf{r}_k$ respectively. Clearly $\sum_k b_{ijk} = f_{ij}$, the flow from P to Q, and $\sum_i b_{ijk} = g_{ij}$, the flow from Q to R. We define

$$h_{ik} \stackrel{\triangle}{=} \sum_j b_{ijk}$$

to be a flow from $\mathbf{p}_i$ to $\mathbf{r}_k$. This flow is a feasible one because it satisfies the constraints (1)–(4) in Section 4. Constraint (1) holds since by construction $b_{ijk} > 0$. Constraints (2) and (3) hold because

$$\sum_k h_{ik} = \sum_{j,k} b_{ijk} = \sum_j f_{ij} = w_{\mathbf{p}_i},$$

and

$$\sum_i h_{ik} = \sum_{i,j} b_{ijk} = \sum_j g_{jk} = w_{\mathbf{r}_k},$$

and constraint (4) holds because the signatures have equal weights. Since EMD(P, R) is the minimal flow from P to R, and h_{ik} is some legal flow from P to R,

$$\begin{aligned}
\text{EMD}\,(P, R) &\le \sum_{i,k} h_{ik} d\,(\mathbf{p}_i, \mathbf{r}_k) \\
&= \sum_{i,j,k} b_{ijk} d\,(\mathbf{p}_i, \mathbf{r}_k) \le \sum_{i,j,k} b_{ijk} d\,(\mathbf{p}_i, \mathbf{q}_j) \\
&\quad + \sum_{i,j,k} b_{ijk} d\,(\mathbf{q}_j, \mathbf{r}_k) \quad (d(\cdot,\cdot) \text{ is metric}) \\
&= \sum_{i,j} f_{ij} d\,(\mathbf{p}_i, \mathbf{q}_j) + \sum_{j,k} g_{jk} d\,(\mathbf{q}_j, \mathbf{r}_k) \\
&= \text{EMD}\,(P, Q) + \text{EMD}\,(Q, R).
\end{aligned}$$

B. Lower Bound Proof

Here we show that when the ground distance is induced by the norm $\|\cdot\|$, the distance between the centroids of two signatures is a lower bound on the EMD between them. Let $\mathbf{p}_i$ and $\mathbf{q}_j$ be the coordinates of cluster i in the first signature, and cluster j in the second signature respectively. Then, using the notation of Eqs. (1)–(4),

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} &= \sum_{i=1}^{m}\sum_{j=1}^{n} \|\mathbf{p}_i - \mathbf{q}_j\| f_{ij} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \| f_{ij}(\mathbf{p}_i - \mathbf{q}_j)\| \quad (f_{ij} \ge 0) \\
&\ge \left\| \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(\mathbf{p}_i - \mathbf{q}_j) \right\| \\
&= \left\| \sum_{i=1}^{m} \left(\sum_{j=1}^{n} f_{ij} \right) \mathbf{p}_i \right. \\
&\qquad \left. - \sum_{j=1}^{n} \left(\sum_{i=1}^{m} f_{ij} \right) \mathbf{q}_j \right\| \\
&= \left\| \sum_{i=1}^{m} w_{\mathbf{p}_i} \mathbf{p}_i - \sum_{j=1}^{n} w_{\mathbf{q}_j} \mathbf{q}_j \right\| \\
&= \|\bar{P} - \bar{Q}\|,
\end{aligned}$$

where $\bar{P}$ and $\bar{Q}$ are the centers of mass of P and Q respectively.

C. Proof that $\rho(x, y) = 1 - e^{-\alpha\|x-y\|}$ is Metric

Here we show that the ground distance that we use in Section 5 is indeed a metric. Clearly positive definiteness and symmetry hold given that $\alpha \ge 0$. We now prove that also the triangle inequality holds.

Given that $\|x - y\| + \|y - z\| \ge \|x - z\|$, we can write

$$\begin{aligned}
0 &\le \rho(x, y)\rho(y, z) \\
&= \left(1 - e^{-\alpha\|x-y\|}\right)\left(1 - e^{-\alpha\|y-z\|}\right) \\
&= 1 - e^{-\alpha\|x-y\|} - e^{-\alpha\|y-z\|} + e^{-\alpha(\|x-y\|+\|y-z\|)} \\
&\le 1 - e^{-\alpha\|x-y\|} - e^{-\alpha\|y-z\|} + e^{-\alpha(\|x-z\|)} \\
&= \left(1 - e^{-\alpha\|x-y\|}\right) + \left(1 - e^{-\alpha\|y-z\|}\right) \\
&\quad - \left(1 - e^{-\alpha(\|x-z\|)}\right) \\
&= \rho(x, y) + \rho(y, z) - \rho(x, z)\ ,
\end{aligned}$$

and hence, $\rho(x, y) + \rho(y, z) \ge \rho(x, z)$.

Acknowledgment

This work was supported by DARPA grant DAAH04-94-G-0284, NSF grant IRI-9712833, and a grant from the Charles Lee Powell Foundation.

References

Ahuja, R.K., Magnanti, T.L., and Orlin, J.B. 1993. *Network Flows*. Prentice Hall: Englewood Cliffs, NJ.

Belongie, S., Carson, C., Greenspan, H., and Malik, J. 1998. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In *IEEE International Conference on Computer Vision*, Bombay, India. pp. 675–682.

Bentley, J.L. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18:509–517.

Bigün, J. and du Buf, J.M. 1994. N-folded symmetries by complex moments in Gabor space and their application to unsupervised texture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):80–87.

Bovik, A.C., Clark, M., and Geisler, W.S. 1990. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(12):55–73.

Bozkaya, T. and Ozsoyoglu, M. 1997. Distance-based indexing for high-dimensional metric spaces. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 26(2):357–368.

Brodatz, P. 1966. *Textures: A Photographic Album for Artists and Designers*: Dover: New York, NY.

Clarkson, K.L. 1997. Nearest neighbor queries in metric spaces. In *ACM Symposium on the Theory of Computing*, pp. 609–617.

Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons: New York, NY, USA.

Das, M., Riseman, E.M., and Draper, B.A. 1997. FOCUS: Searching for multi-colored objects in a diverse image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 756–761.

Daugman, J.D. 1998. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36:1169–1179.

Duda, R.O. and Hart, P.E. 1973. *Pattern Classification and Scene Analysis*. Wiley: New York.

Farrokhnia, F. and Jain, A.K. 1991. A multi-channel filtering approach to texture segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 364–370.

Field, D.J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394.

Gabor, D. 1946. Theory of communication. *The Journal of the Institute of Electrical Engineers, Part III*, 93(21):429–457.

Hafner, J., Sawhney, H.S., Equitz, W., Flickner, M., and Niblack, W. 1995. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–735.

Hillier, F.S. and Lieberman, G.J. 1990. *Introduction to Mathematical Programming*. McGraw-Hill, New York, NY.

Hitchcock, F.L. 1941. The distribution of a product from several sources to numerous localities. *J. Math. Phys.*, 20:224–230.

Karmarkar, N. 1984. A new polynomial-time algorithm for linear programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, Washington, DC. pp. 302–311.

Klee, V. and Minty, G. 1972. How good is the simplex algorithm. In *Inequalities*, Vol. III, O. Shisha (Ed.). Academic Press: New York, NY, pp. 159–175.

Kullback, S. 1968. *Information Theory and Statistics* Dover: New York, NY.

Liu, F. and Picard, R.W. 1996. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733.

Manjunath, B.S. and Ma, W.Y. 1996. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842.

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

Niblack, W., Barber, R., Equitz, W., Flickner, M.D., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G., and Heights, Y. 1993. Querying images by content, using color, texture, and shape. In *SPIE Conference on Storage and Retrieval for Image and Video Databases*, Vol. 1908, pp. 173–187.

Peleg, S., Werman, M., and Rom, H. 1989. A unified approach to the change of resolution: Space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:739–742.

Poynton, C. 1996. *A Technical Introduction to Digital Video*. John Wiley and Sons, New York, NY.

Puzicha, J., Hofmann, T., and Buhmann, J. 1997. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 267–272.

Rachev, S.T. 1984. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications*, XXIX(4):647–676.

Russell, E.J. 1969. Extension of Dantzig's algorithm to finding an initial near-optimal basis for the transportation problem. *Operations Research*, 17:187–191.

Shen, H.C. and Wong, A.K.C. 1983. Generalized texture representation and metric. *Computer, Vision, Graphics, and Image Processing*, 23:187–206.

Smith, J.R. 1997. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression*. Ph.D. Thesis, Columbia University.

Stolfi, J. 1994. Personal communication.

Stricker, M. and Orengo, M. 1995. Similarity of color images. In *SPIE Conference on Storage and Retrieval for Image and Video Databases III*, Vol. 2420, pp. 381–392.

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

Tversky, A. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

Werman, M., Peleg, S., and Rosenfeld, A. 1985. A distance metric for multi-dimensional histograms. *Computer, Vision, Graphics, and Image Processing*, 32:328–336.

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

Zikan, K. 1990. *The Theory and Applications of Algebraic Metric Spaces*. Ph.D. Thesis, Stanford University.