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
[rubner,tomasi,guibas]@cs.stanford.edu

Abstract

We introduce a metric between two distributions that we call the *Earth Mover's Distance* (EMD). The EMD is based on the minimal cost that must be paid to transform one distribution into the other, in a precise sense. We show that the EMD has attractive properties for content-based image retrieval. The most important one, as we show, is that it matches perceptual similarity better than other distances used for image retrieval. 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.

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, often 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 new descriptions, the dominant clusters are extracted from the original distribution 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. Our measures of distance encompass both cases, since we define them for signatures, and signatures subsume histograms.

Given two distributions, represented as either histograms or signatures, 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 measure 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 [30].

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. Also, metric distances lead to more efficient data structures and search algorithms [4, 6]. Practically, it is crucial 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 [28] of the non-metric nature of perceptual distances.

We introduce a distance between two signatures that we call the *Earth Mover's Distance* [1] (EMD) . This is a useful and flexible metric distance, based on the minimal cost that must be paid to transform one signature into the other, in a sense that will be made precise in section 4. The EMD is based on a solution to the *transportation problem* [13] from linear optimization, for which efficient algorithms are available. The EMD 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 the rigid binning of the latter. It allows for partial matching, and it can be applied to signatures with different sizes. When used to compare distributions that have the same overall mass, the EMD is a true metric.

Although the EMD is a general method for matching multidimensional distributions, in this paper we focus on applications 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. For color we compare the results of image retrieval using the EMD with the results obtained with other metrics. For texture, a similar comparison is hard to make since other method cannot handle the high-dimensional space that we use to represent texture. Instead we 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. Texture features are commonly defined by energy distributions in the spatial frequency domain [9, 2, 16]. Image database are indexed by these histograms, 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 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

[1] The name Earth Mover's Distance was suggested by Stolfi

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).

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 differences. 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 [27]. Other common usages are L_2 and L_∞. In [26] 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 [27] is attractive because of its ability to handle partial matches when the areas of the two histograms are different. It is shown in [27] 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 [14] is defined as:

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

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 [7]. 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 [20]. 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 $m_i = \frac{h_i+k_i}{2}$. 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. The major drawback of these measures 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 figure 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 into the dissimilarity measure results in perceptually more meaningful dissimilarity measures.

Quadratic-form distance: this distance was suggested for color based retrieval in [17]:

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

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 [17] 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 true distance, it can be shown that with this choice of $\mathbf{A}$ the quadratic-form is indeed a distance.

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 figure 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 [26], 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 [24, 29] 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 difference

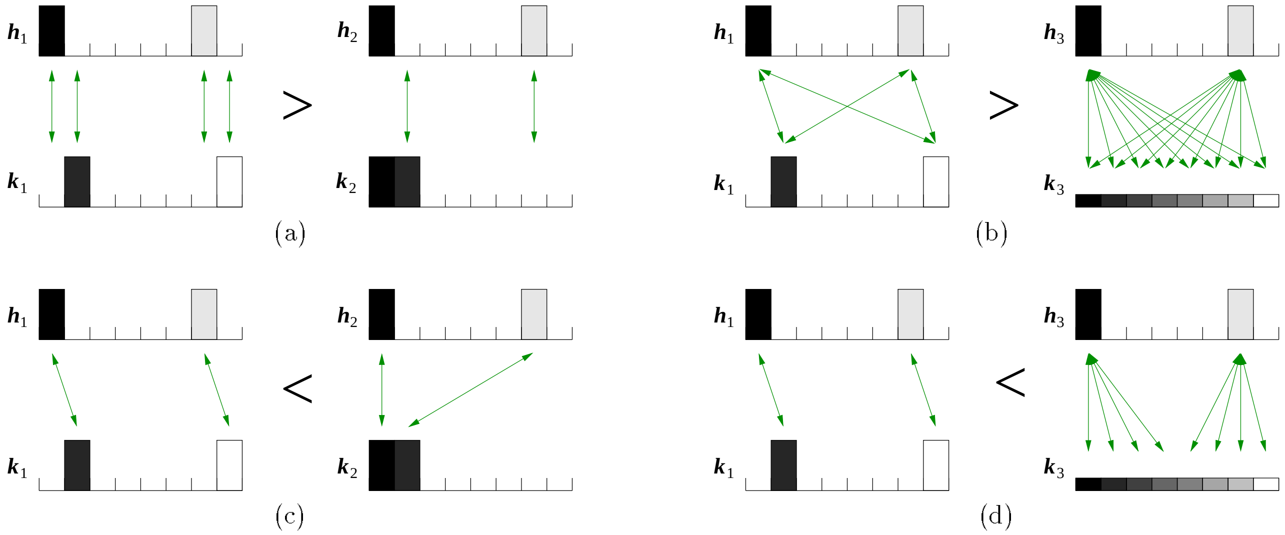


Figure 1: Examples where the L_1 distance (as a representative of bin-by-bin dissimilarity measures) and the quadratic-form distance do not match perceptual dissimilarity. Assuming that histograms have unit mass, (a) $d_{L_1}(\mathbf{h}_1, \mathbf{k}_1) = 2$, $d_{L_1}(\mathbf{h}_2, \mathbf{k}_2) = 1$. (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.

that the match distance cannot handle partial matches. The match distance does not extend to higher dimensions because the relation $\mathbf{j} \leq \mathbf{i}$ 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.

Parameter-based distances: These methods first compute a small set of parameters from the histograms, either explicitly or implicitly, and then compare these parameters. For instance, in [26] the distance between distributions is computed as the sum of the weighted distances of the distributions' first three moments. It is unclear how to tune the weights of the different moments. Moreover, the resulting measure is not a metric distance. In [15], textures are compared based on measures of their periodicity, directionality, and randomness, while in [16] 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 [25, 20].

An example of color-based image retrieval using different dissimilarity measures is shown in figure 2. 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. The details about the extraction of the color histograms are given in section 5.1. Although, in general, semantic interpretation should not be used to judge color similarity of images, notice that the different methods returned different number of red cars. A more thorough comparison of the different dissimilarity measures is given in section 5.1.

Figure 2: 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.

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. Similar problems arise even when adaptive histograms are used. In brief, because histograms are fixed-size structures, they cannot achieve a good balance between expressiveness and efficiency.

A *signature* $\{\mathbf{s}_j = (\mathbf{m}_j, w_j)\}$, 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_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. The size of the clusters in the feature space should be limited and not exceed the extent of what is perceived as the same, or very similar, feature.

Since the definition of cluster is open, a histogram $\{h_\mathbf{i}\}$ can be viewed as a signature $\{\mathbf{s}_j = (\mathbf{m}_j, w_j)\}$ 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 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 the difference between those colors. 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 [18, 31], defined as the minimum cost of matching the bins of two histograms. In this section we take this approach and extend it to derive the *Earth Mover's Distance* (EMD) as a general metric between signatures for image retrieval.

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. Examples of ground distances are given in section 5.

Computing the EMD is based on a solution to the well-known *transportation problem* [13] a.k.a. the Monge-Kantorovich problem [21]. 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 $P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\}$ be the first signature with m clusters, where p_i is the cluster representative and w_{p_i} is the weight of the cluster; $Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$ the second signature with n clusters; and $\mathbf{D} = [d_{ij}]$ the ground distance matrix where d_{ij} is the ground distance between clusters p_i and q_j.

We want to find a flow $\mathbf{F} = [f_{ij}]$, with f_{ij} the flow between p_i and 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:

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

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \qquad 1 \leq i \leq m \tag{2}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j} \qquad 1 \leq j \leq n \tag{3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) , \tag{4}$$

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 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. In addition, 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. Fortunately, efficient algorithms for the transportation problem are available. We used the transportation-simplex method [12], 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 [23].

To measure the time-performance of our EMD implementation, we generated random signatures of sizes that range from 1 to 100. For each size we generated 100 pairs of random signatures and computed the average CPU time to compute the EMD between the pairs. The results are shown in figure 3. This experiment was done on a SGI Indigo 2 with a 195MHz CPU. For non-random signatures, the running times are usually lower because a small fraction of the clusters in a signature typically captures most of the signatures' weight.

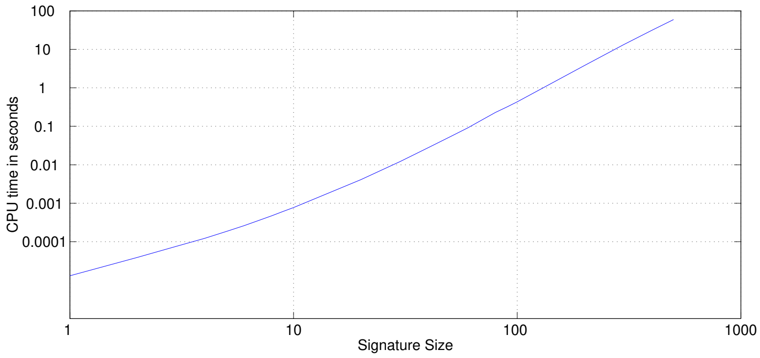


Figure 3: A log-log plot of the average computation time for random signatures as a function of signature size.

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

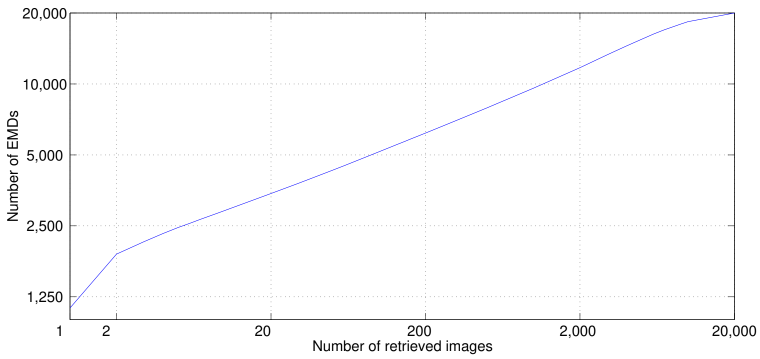


Figure 4: A log-log plot of the number of EMDs as a function of the number of images retrieved. The database contains 20,000 images.

An efficiency 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.

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 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.

5.1 Color Distributions

For the computation of the earth mover's distance between color images, we use Euclidean distance in the CIE-Lab color space [30] as the underlying ground distance between individual colors. This color space is expressly designed so that short Euclidean distances correlate strongly with human color discrimination performance.

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 [19]. 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 simple *k-d* tree [1] 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 (p, w_p) to the signature representation of the image where p is the average color of the cluster, and the corresponding weight w_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.

We implemented a color-based image retrieval system that uses the EMD on color signatures [2]. Although our system works remarkably well on color signatures, in order to compare the EMD to the histogram dissimilarity measures described in section 2 we also computed histograms for all the images in the database. We ran two experiments, one on color histograms with coarse binning, and one with fine binning. In the first experiment, 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 is comparable to that contained in the histograms.

---

[2] A demo of our color-based image retrieval can be found at `http://vision.stanford.edu/~rubner/demo`

In our second experiment, 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. Nevertheless, we claim that even with less information, signatures result in better retrieval precision than histograms.

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. Our goal is rather to compare the EMD to the other dissimilarity measures described in section 2. For that purpose we conducted a few experiments where we created a common ground truth on which we measured the performance of the different methods.

In our first experiment, we manually identified 75 images of red cars in the database and marked them as relevant. From this set we chose the ten "good" images that are shown at the top of figure 5. 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. An example of such a query is shown in figure 2. The average number of relevant images for the different dissimilarity measures as a function of the number of images retrieved is shown in figure 5(bottom) and 5(middle) 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.

In this experiment, 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.

The second experiment is an example where both the colors of the objects and the colors of the backgrounds are similar for all the relevant images. This experiment was done with the relevant images being a set of 157 images of brown horses in green fields. Again 10 "good" images of horses (figure 6, top) were used for the query, with the results shown in figure 6(middle) and 6(bottom) for the coarse and fine histograms respectively. Here again the EMD that was computed on the signatures performed 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. 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 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 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

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

The parameter λ defines the importance of the color positions relative to the color values. Figure 7 shows the effect of position information where the leftmost image of a skier was used as the query. Part (a) shows the

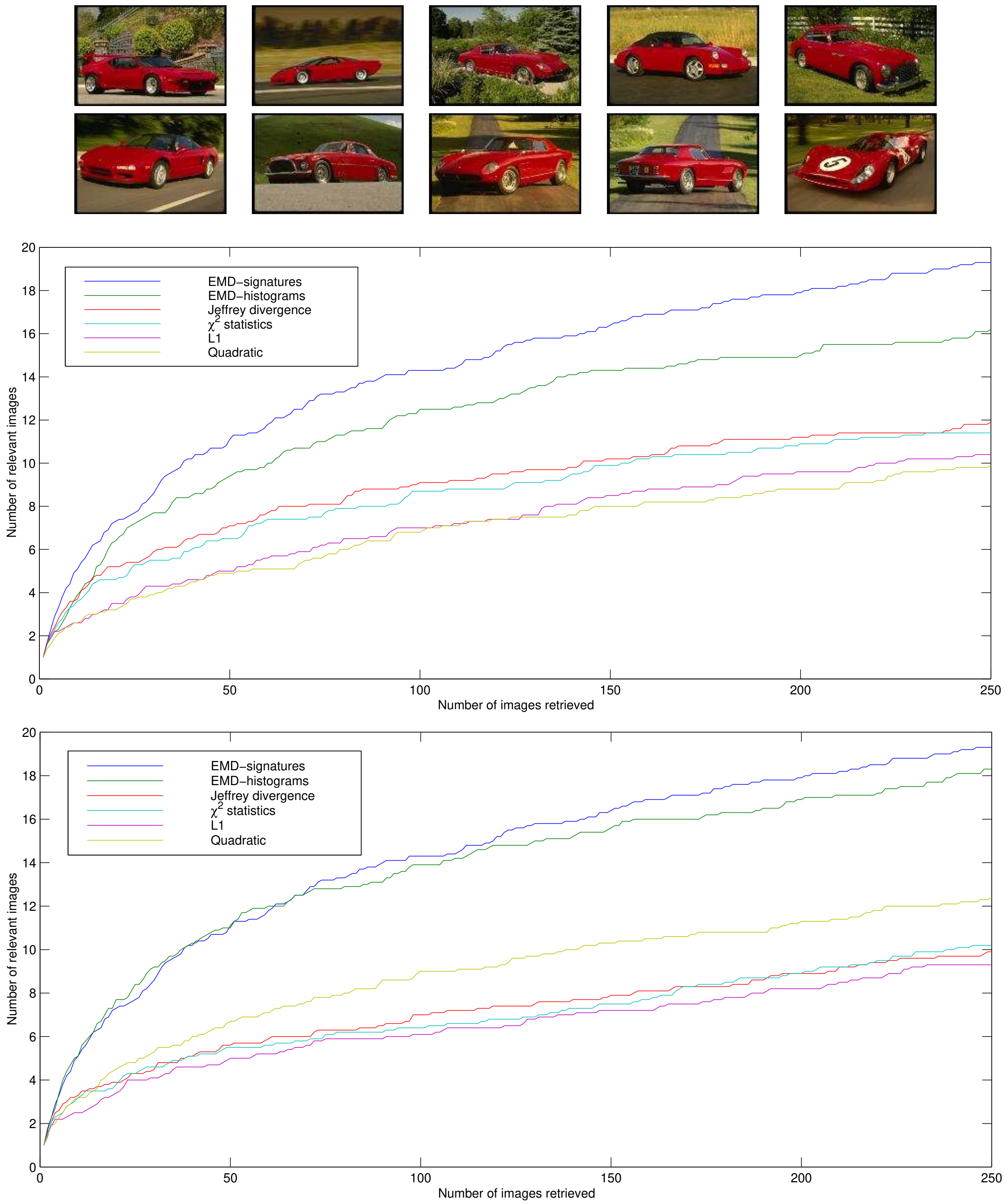
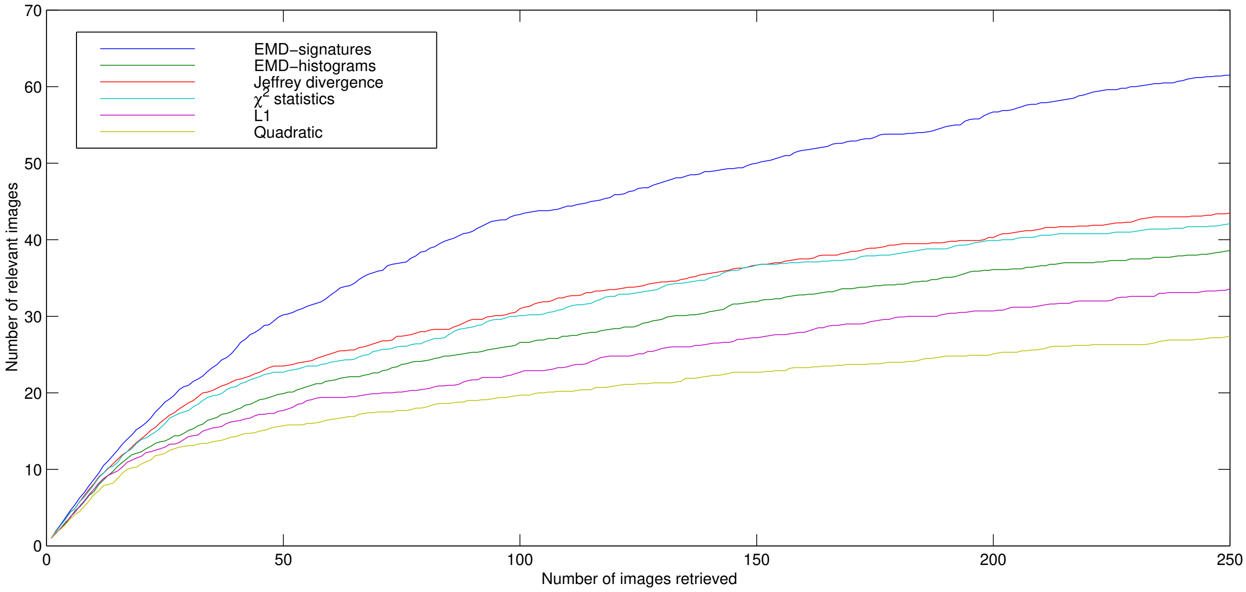


Figure 5: Ten images of red cars (top) and the average number of relevant images, for the different dissimilarity measures, that were returned by using the ten images as the queries for the histograms (middle) and fine (bottom) histograms. The results obtained by using signatures is also shown in the two graphs for reference.



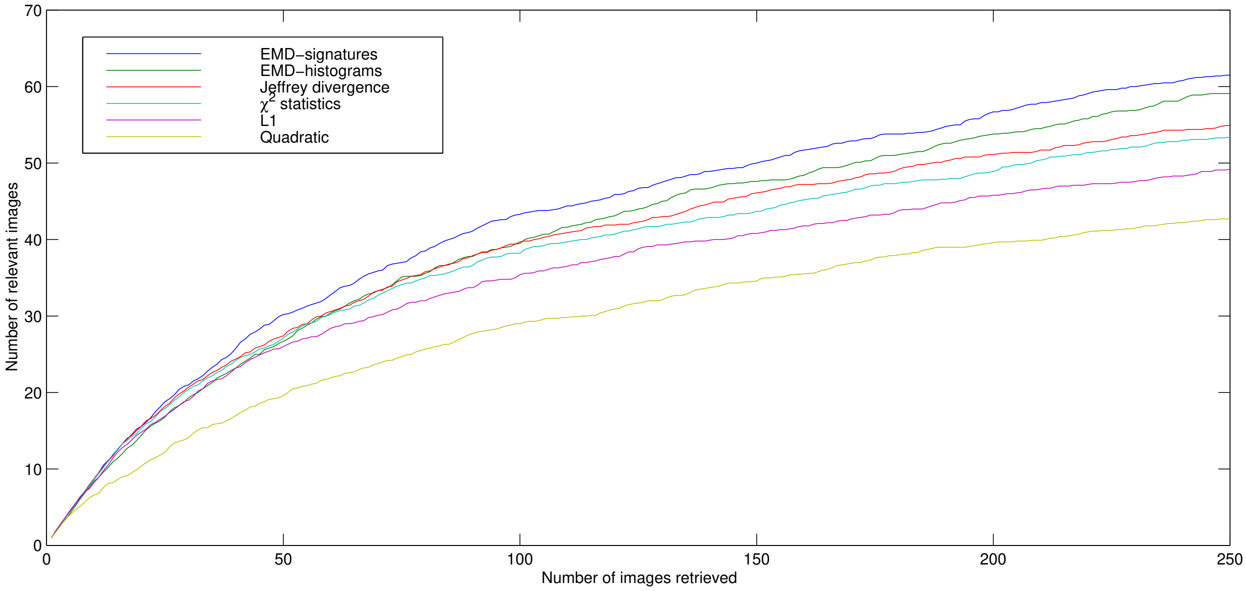


Figure 6: Ten images of horses (top) and the average number of relevant images, for the different dissimilarity measures, that were returned by using the ten images as the queries for the coarse (middle) and fine (bottom) histograms. The results obtained by using signatures is also shown in the two graphs for reference.

6 best matches when position information was ignored ($\lambda = 0$). Part (b) uses position information ($\lambda = 0.5$). Exact color matches are somewhat compromised in order to get more similar positional layouts.

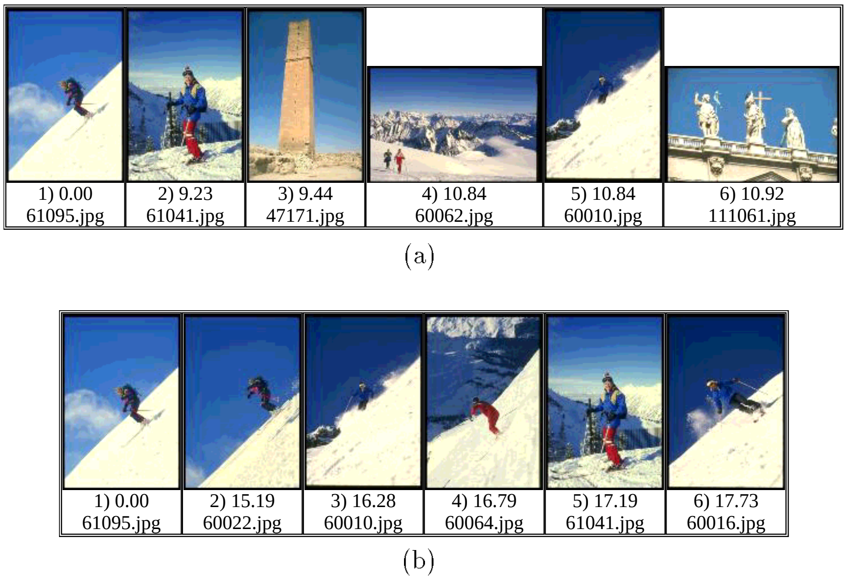


Figure 7: Using the leftmost image of a skier as the query. The six best matches without position information (a) and with position information (b).

5.3 Texture

While color is a purely point wise 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 (e.g. [3, 9, 16]) because they are optimally localized in both the spatial and frequency domains [11]. 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 [8].

In this paper we used the Gabor filter dictionary that was derived in [16] 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 be able to treat all the Gabor responses from the different scales in a similar way, we need to appropriately normalize the vector. Unlike [16], who normalizes each feature in the vector by the standard deviation of the respective feature over the entire database, we normalize the feature by the radial frequency f of the corresponding Gabor filter. This follows [10] who shows that the magnitude of the power spectra of natural images, falls as $1/f$, suggesting that cells in the visual pathway are likely to follow this pattern. Empirically, the two methods yield similar normalization factors. 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*. Examples of texture features are given in Figure 8.

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. The resulting set of cluster centers together with the cluster weights is the *texture signature*.

We constructed a database of 1744 texture patches, by dividing each of 109 textures from the Brodatz

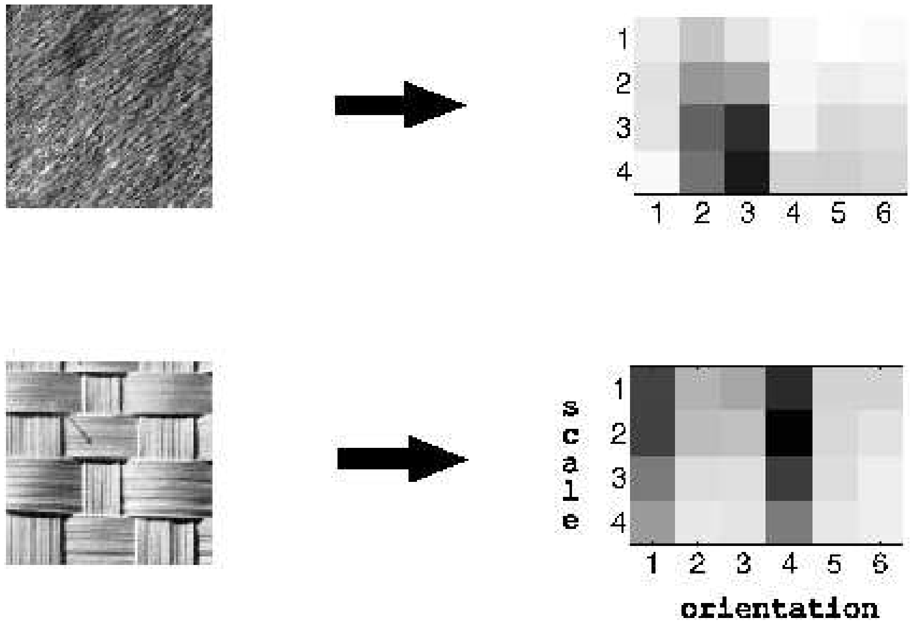


Figure 8: Texture features. *(left)* Texture patches from the Brodatz album [5]. *(right)* The average Gabor responses over the whole texture patch. The Gabor filter bank consists of four scales and six orientations. The top texture (Fur, D93) has one dominant orientation fine scales. The bottom texture (Hand-woven oriental rattan, D64) has two dominant orientations, mostly at coarse scales.

album [5][3], into 4 by 4 non-overlapping patches. Every patch is 128 by 128 pixels. After the clustering process, the average size of the texture signatures was 12 clusters.

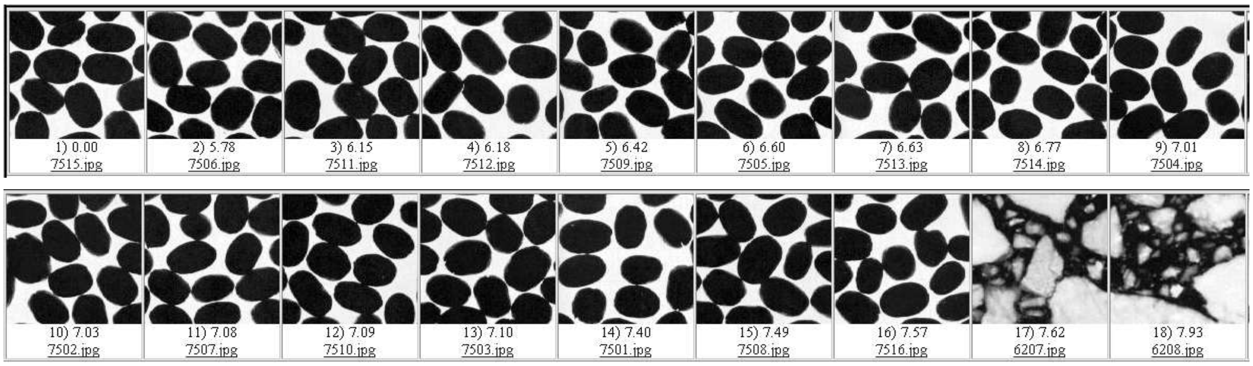
In order to use the EMD on our texture signatures, we have to define a ground distance. A natural ground distance is to consider the texture features as a distribution in a two-dimensional space: 4 scales by 6 orientations, and define a low-level EMD that will serve as the ground distance of the high-level one. This low-level EMD is discussed in [22]. For image retrieval, this two-level EMD approach is too slow, so we use the L_1-distance between the texture features as an approximation for the low-level EMD. This is reasonable, because the quantization of the two-dimensional space is very coarse (4 scales and 6 orientations): The fact that the L_1-distance does not consider neighboring bins is justified here, since neighboring bins are not too similar - they are one octave apart in scale and 30 degrees apart in orientation.

Having defined texture signatures and a ground distance between them, we can now use the EMD to retrieve images with textures. Figure 9 shows two examples of retrieving texture patches by using a given texture as the query. Notice that the full distribution in the Gabor space is used. Other methods for texture similarity that are based on histograms work effectively only on the marginals [16, 20]. This difference makes it hard to compare the EMD with the other similarity measures described in section 2.

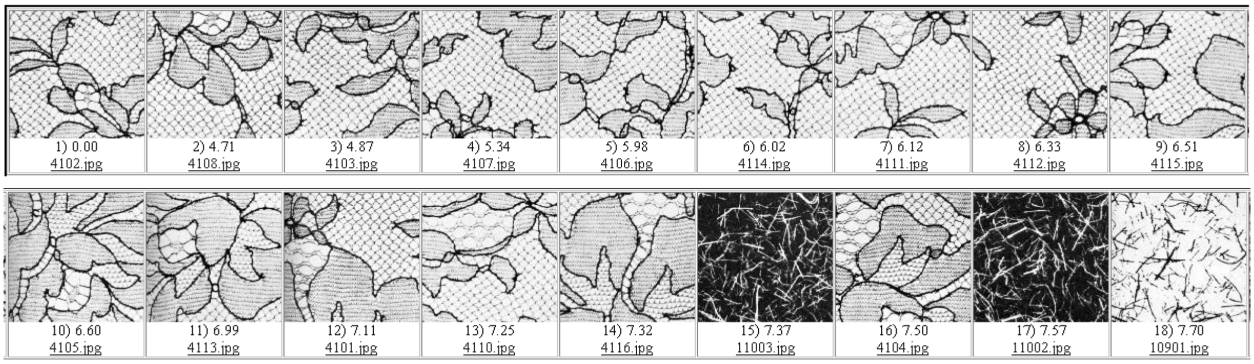
An important advantage of the EMD over other measures for texture similarity is its ability to handle images that contain more than one texture without first segmenting the images as needed when using other measures. Using the EMD for partial matches can find images that contain specific textures. Figure 10 shows 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 11 shows an example of texture-based image retrieval on natural images. We created a database of 250 images of animals from the Corel Photo Collection with zebras in 25 of the images. From the image of zebra in part (a) we cropped a block with the zebra's stripe pattern as shown in part (b), and asked for

[3] The full Brodatz album consists of 112 textures. We used only the 109 textures that were used by [16] in order to be able to compare results from both methods.



(a)



(b)

Figure 9: Texture queries. The first image in each of the two parts was used as the query. (a) Coffee beans (D75). All the 16 patches from the same texture were returned first. (b) Lace (D41). Here, 15 out of the 16 texture patches are in the top 18 matches.

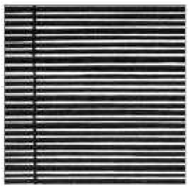
images with at least 20% of this texture. The 8 best matches are shown in part (c) ranked by their similarity to the query.

6 Conclusions

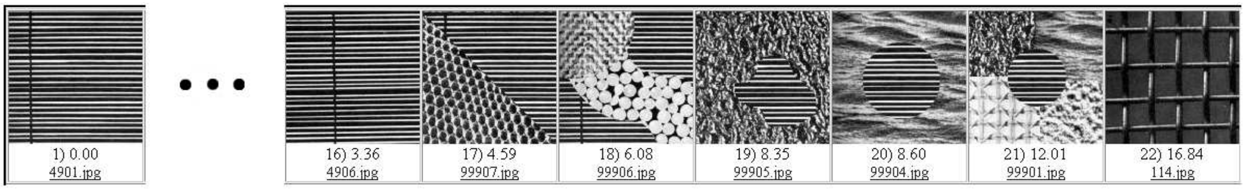
The earth mover's distance is a general and flexible metric and has desirable properties for image retrieval. It allows for partial matches, and it can be applied to variable-length representations of distributions. It can be computed efficiently, and lower bounds are readily available for it. Because of these advantages, both conceptual and computational, 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 fundamental element of image retrieval systems. Comparisons with other dissimilarity measures show that the EMD matches perceptual dissimilarity better.

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.



(a)



(b)

Figure 10: Texture partial query. The query was 20% of the texture in part (a) and 80% "don't care". (b) The best matches: the 16 patches from the same texture followed by all the compositions that contain some part of the queried texture. No segmentation was performed.

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

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

[2] J. Bigün and J. M. du Buf. 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, 1994.

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

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

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

[6] Kenneth L. Clarkson. Nearest neighbor queries in metric spaces. In *ACM Symposium on the Theory of Computing*, pages 609–617, 1997.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

[8] J. D. Daugman. 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, 1988.

[9] F. Farrokhnia and A. K. Jain. A multi-channel filtering approach to texture segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 364–370, June 1991.

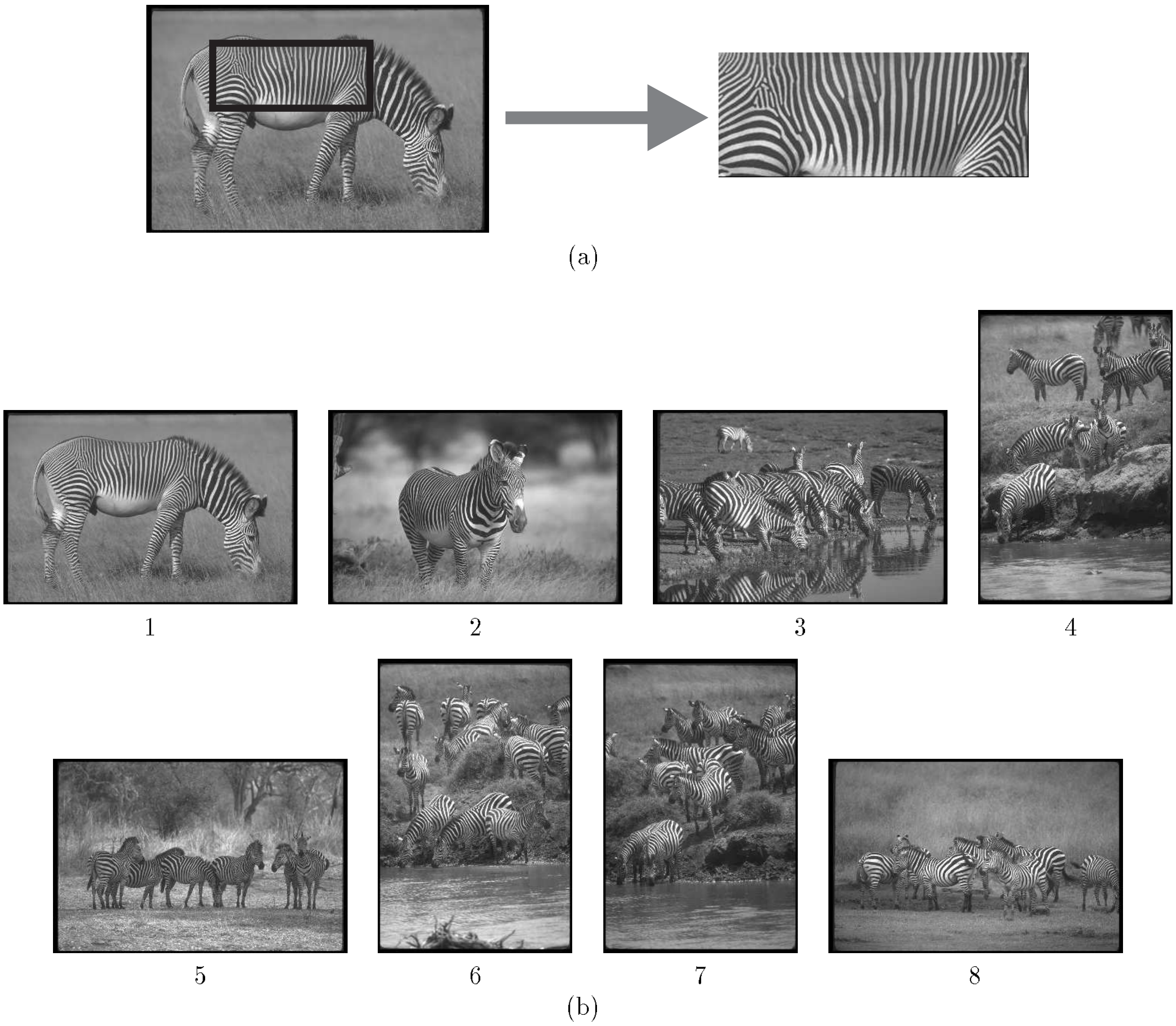


Figure 11: Looking for zebras. (a) An image of a zebra and a block of zebra stripes extracted from it. (b) The eight best matches to a query asking for images with at least 20% of the texture in (a).

[10] David J. Field. 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, December 1987.

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

[12] F. S. Hillier and G. J. Liberman. *Introduction to Mathematical Programming.* McGraw-Hill, 1990.

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

[14] S. Kullback. *Information Theory and Statistics.* Dover, New York, NY, 1968.

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

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

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

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

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

[20] J. Puzicha, T. Hofmann, and J. M. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 1997.

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

[22] Y. Rubner and C. Tomasi. Texture metrics. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, San Diego, CA, October 1998. IEEE Systems, Man and Cybernetics Society. To appear.

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

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

[25] J. R. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression.* PhD thesis, Columbia University, 1997.

[26] M. Stricker and M. Orengo. Similarity of color images. In *SPIE Conference on Storage and Retrieval for Image and Video Databases III*, volume 2420, pages 381–392, February 1995.

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

[28] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

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

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

[31] K. Zikan. *The Theory and Applications of Algebraic Metric Spaces.* PhD thesis, Stanford University, 1990.

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 p_i to q_j and from q_j to r_k defines a flow which we call b_{ijk} where i, j and k correspond to p_i, q_j and 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

$$h_{ik} \triangleq \sum_j b_{ijk}$$

to be a flow from p_i to 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

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

and

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

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

$$\begin{aligned}
\mathrm{EMD}(P,R) &\leq \sum_{i,k} h_{ik} d(p_i, r_k) \\
&= \sum_{i,j,k} b_{ijk} d(p_i, r_k) \\
&\leq \sum_{i,j,k} b_{ijk} d(p_i, q_j) + \sum_{i,j,k} b_{ijk} d(q_j, r_k) \qquad \text{(because } d(\cdot,\cdot) \text{ is metric)} \\
&= \sum_{i,j} f_{ij} d(p_i, q_j) + \sum_{j,k} g_{jk} d(q_j, r_k) \\
&= \mathrm{EMD}(P,Q) + \mathrm{EMD}(Q,R) \ .
\end{aligned}$$

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 p_i and 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 equations (1)-(4),

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} &= \sum_{i=1}^{m} \sum_{j=1}^{n} \|p_i - q_j\| f_{ij} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \|f_{ij}(p_i - q_j)\| \qquad (f_{ij} \geq 0) \\
&\geq \Big\| \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(p_i - q_j) \Big\| \\
&= \Big\| \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} f_{ij} \Big) p_i - \sum_{j=1}^{n} \Big(\sum_{i=1}^{m} f_{ij} \Big) q_j \Big\| \\
&= \Big\| \sum_{i=1}^{m} w_{p_i} p_i - \sum_{j=1}^{n} w_{q_j} q_j \Big\| \\
&= \|\bar{P} - \bar{Q}\| \ ,
\end{aligned}$$

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