The Earth Mover's Distance under Transformation Sets *

Scott Cohen Leonidas Guibas
Computer Science Department
Stanford University, Stanford, CA 94305

Abstract

The Earth Mover's Distance (EMD) is a distance measure between distributions with applications in image retrieval and matching. We consider the problem of computing a transformation of one distribution which minimizes its EMD to another. The applications discussed here include estimation of the size at which a color pattern occurs in an image, lighting-invariant object recognition, and point feature matching in stereo image pairs. We present a monotonically convergent iteration which can be applied to a large class of EMD under transformation problems, although the iteration may converge to only a locally optimal transformation. We also provide algorithms that are guaranteed to compute a globally optimal transformation for a few specific problems, including some EMD under translation problems.

1. Introduction

A major challenge in image retrieval applications is that the images we desire to match can be visually quite different. This can happen even if these images are views of the same scene because of illumination changes, viewpoint motion, occlusions, etc.. Two common approaches to measure image similarity modulo some given factors are: (I) compare invariant image signatures (e.g. [4]), and (II) compare non-invariant signatures with a distance measure that allows for differences due to the given factors (e.g. [6, 12]).

The challenge in approach (I) is to compute invariants that still distinguish images with differences that should be penalized. Using invariants computed over entire images assumes that two images are similar only if all the information in one image matches all the information in the other. Such a complete matching measure is usually not appropriate in an image retrieval system because semantic image similarity often follows from only a partial match. Approach (II) is better in the partial matching case since invariance can be built on top of a distance function which allows for partial matching. Of course, approach (I) can be modified to use invariants computed over parts of images, but this can require quite a lot of space because invariants must be computed for all image regions which might be matched at query time.

A very general distance measure with applications in content-based image retrieval is the Earth Mover's Distance (EMD) between distributions ([10]). The EMD allows for partial matching, and has been successfully used for measuring image similarity with respect to color and texture ([11]). For example, in [11] the color signature of an image is a collection of dominant image colors in CIE-Lab space ([16]), where each color is weighted by the fraction of image pixels classified as that color. Also in [11], the texture signature of a single texture image is a collection of spatial frequencies in log-polar coordinates, where each frequency is weighted by the amount of energy present at that frequency. Experiments in [10] show the superiority of the EMD for color-based image retrieval over many histogram dissimilarity measures, including a common quadratic form distance ([8]).

In this paper, we extend the EMD to allow unpenalized distribution transformations. The goal is to find a transformation of one distribution which minimizes its EMD to another, where a set of allowable transformations is given. Consider, for example, using the EMD to measure object similarity with respect to color. An EMD between color signatures does not account for lighting differences. In [4], the authors show that an illumination change results in a linear transformation of image pixel colors (under certain reasonable assumptions). For the texture signatures mentioned above, a change in texture scale and orientation results in a translation of signature points in log-polar spatial frequency space.

The EMD under transformation ($\mathrm{EMD}_{\mathcal{G}}$) problem is to compute $\min_{g \in \mathcal{G}} \mathrm{EMD}(\mathbf{x}, g(\mathbf{y}))$, where $\mathbf{x} = \{(x_i, w_i)\}_{i=1}^{m}$ and $\mathbf{y} = \{(y_j, u_j)\}_{j=1}^{n}$ are summary distributions (we also call them *signatures*) for the images being compared and $\mathcal{G}$ is a set of transformations. For example, the points x_i and y_j are points in CIE-Lab space and log-polar spatial frequency space in the color and texture cases, respectively. The weights w_i and u_j are the amounts of features x_i and y_j present in the images. The set of allowable transformations

*Authors' email: `(scohen,guibas)@cs.stanford.edu`. See `http://vision.stanford.edu/~scohen` or the CD-ROM proceedings for a color version of this work.

is application-dependent. In the lighting-invariant object recognition application, $\mathcal{G}$ is the set of linear transformations $\{g_L\}$, and $g_L(\mathbf{y}) = \{(Ly_j, u_j)\}$. For texture comparison which is insensitive to differences in scale and orientation, $\mathcal{G}$ is the set of translations $\{g_t\}$, and $g_t(\mathbf{y}) = \{(y_j + t, u_j)\}$. In both these examples, the allowable transformations change the points of a distribution but leave its weights fixed. We shall also consider a set of transformations $\mathcal{G} = \{g_c\}$ in which g_c changes only the weights of a distribution as $g_c(\mathbf{y}) = \{(y_j, cu_j)\}$. This set arises in estimating the size at which color pattern occurs in a color image.

We begin in section 2 with a brief review of the EMD. In section 3, we consider the problem of computing the EMD under various transformation sets. We start in section 3.1 with a discussion of the scale estimation application and the corresponding $\mathrm{EMD}_\mathcal{G}$ problem in which $g \in \mathcal{G}$ changes only distribution weights. This $\mathrm{EMD}_\mathcal{G}$ problem has structure which we exploit to compute a globally optimal transformation. In section 3.2, we consider $\mathrm{EMD}_\mathcal{G}$ problems in which $g \in \mathcal{G}$ changes only distribution points. For such $\mathcal{G}$, we present in section 3.2.1 a very general, monotonically convergent iteration called the *FT iteration*. We apply the FT iteration to the applications of lighting-invariant object recognition and point feature matching in stereo images in sections 3.2.2 and 3.2.3, respectively. The main drawback of the FT iteration is that it may converge to only a locally optimal transformation. In sections 3.2.4 and 3.2.5, we discuss EMD under translation problems which can be solved directly for a globally optimal translation. Finally, section 4 contains some concluding remarks.

2. The Earth Mover's Distance (EMD)

We denote a discrete *distribution* as a set of weighted points $\mathbf{x} = \{(x_i, w_i)\}_{i=1}^m \equiv (X, w) \in \mathbf{D}^{K,m}$, where $X = [\, x_1 \, \cdots \, x_m \,]$, with each $x_i \in \mathbf{R}^K$, $w_i \geq 0$. Here K is the dimension of the ambient space of the points x_i, and m is the number of points. The weight of $\mathbf{x}$ is $w_\Sigma = \sum_{i=1}^m w_i$.

Given two distributions $\mathbf{x} = (X, w) \in \mathbf{D}^{K,m}$ and $\mathbf{y} = (Y, u) \in \mathbf{D}^{K,n}$, a *flow* between $\mathbf{x}$ and $\mathbf{y}$ is any matrix $F = (f_{ij}) \in \mathbf{R}^{m \times n}$. Intuitively, f_{ij} is the amount of weight at x_i which is matched to weight at y_j. F is a *feasible flow* between $\mathbf{x}$ and $\mathbf{y}$ iff (i) $f_{ij} \geq 0$, (ii) $\sum_{j=1}^n f_{ij} \leq w_i$, (iii) $\sum_{i=1}^m f_{ij} \leq u_j$, and (iv) $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min(w_\Sigma, u_\Sigma)$. Constraint (ii) ensures that the weight in $\mathbf{y}$ matched to x_i does not exceed w_i. Similarly, (iii) ensures that the weight in $\mathbf{x}$ matched to y_j does not exceed u_j. Finally, constraint (iv) forces the total amount of weight matched to be equal to the weight of the lighter distribution. In the unequal-weight case $w_\Sigma \neq u_\Sigma$, some weight in the heavier distribution remains unmatched.

Let $\mathcal{F}(\mathbf{x}, \mathbf{y})$ denote the set of all feasible flows between $\mathbf{x}$ and $\mathbf{y}$. The work done by a feasible flow $F \in \mathcal{F}(\mathbf{x}, \mathbf{y})$ in matching $\mathbf{x}$ and $\mathbf{y}$ is given by $\mathrm{WORK}(F, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^m \sum_{j=1}^n f_{ij} d(x_i, y_j)$, where $d(x_i, y_j)$ is the "ground distance" between x_i and y_j. The *Earth Mover's Distance* $\mathrm{EMD}(\mathbf{x}, \mathbf{y})$ is the minimum amount of work to match $\mathbf{x}$ and $\mathbf{y}$, normalized by the weight of the lighter distribution:

$$\mathrm{EMD}(\mathbf{x}, \mathbf{y}) = \frac{\min_{F \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \mathrm{WORK}(F, \mathbf{x}, \mathbf{y})}{\min(w_\Sigma, u_\Sigma)}. \quad (1)$$

In other words, the EMD is the average ground distance that weights travels during an optimal flow. The work minimization problem in (1) is a special type of linear program called the *transportation problem*, and it can be solved efficiently by the transportation simplex algorithm ([5]).

The EMD matches all the weight in the lighter distribution. The *partial Earth Mover's Distance* EMD^γ matches only a given fraction $\gamma \in (0, 1]$ of the weight of the lighter distribution. The constraint (iv) is replaced by $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \gamma \min(w_\Sigma, u_\Sigma)$ to define $\mathcal{F}^\gamma(\mathbf{x}, \mathbf{y})$, and the minimum work is normalized by $\gamma \min(w_\Sigma, u_\Sigma)$.

3. The EMD under Transformation Sets

The *EMD under transformation set* $\mathcal{G}$ is defined as $\mathrm{EMD}_\mathcal{G}(\mathbf{x}, \mathbf{y}) = \min_{g \in \mathcal{G}} \mathrm{EMD}(\mathbf{x}, g(\mathbf{y}))$, where $g(\mathbf{y})$ is the result of applying the transformation $g \in \mathcal{G}$ to the distribution $\mathbf{y}$. In words, we seek a transformation of one distribution which minimizes its EMD to another.[1] The *partial EMD under transformation set* $\mathcal{G}$ is simply $\mathrm{EMD}_\mathcal{G}^\gamma(\mathbf{x}, \mathbf{y}) = \min_{g \in \mathcal{G}} \mathrm{EMD}^\gamma(\mathbf{x}, g(\mathbf{y}))$.

3.1. Example Use in Scale Estimation (only weights change)

In this section, the problem of estimating the size at which a color pattern occurs in an image is phrased and efficiently solved as an $\mathrm{EMD}_\mathcal{G}$ problem. Suppose that a pattern occurs in an image as a fraction $c^* \in (0, 1]$ of the total image area. An instance is shown in Figure 1(a). Let $\mathbf{x}$ and $\mathbf{y} = (Y, u)$ denote unit-weight color signatures of the image and pattern, respectively. A small set of dominant image colors $\{x_i\}_{i=1}^m$ in CIE-Lab space is computed via the color clustering algorithm in [9]. The weight w_i is the fraction of image pixels whose nearest color cluster is x_i. See Figure 1(b),(d). We use $d = L_2$.[2]

Since (Y, c^*u) is lighter than $\mathbf{x}$, the computation $\mathrm{EMD}(\mathbf{x}, (Y, c^*u))$ finds an optimal matching between c^* of the image color weight and the color weight in (Y, c^*u). Consider the ideal case of an exact pattern occurrence in

[1] In some situations, the symmetric definition $\mathrm{EMD}_\mathcal{G}(\mathbf{x}, \mathbf{y}) = \min(\min_{g \in \mathcal{G}} \mathrm{EMD}(\mathbf{x}, g(\mathbf{y})), \min_{g \in \mathcal{G}} \mathrm{EMD}(g(\mathbf{x}), \mathbf{y}))$ may be more appropriate.

[2] Euclidean distance in CIE-Lab space matches perceptual distance between two colors that are not very different ([16]).

the image, with the same color clusters used in $\mathbf{x}$ and $\mathbf{y}$ for the pattern colors. Then the c^* of $\mathbf{x}$'s color weight contributed by the pattern occurrence will match exactly the color weight in (Y, c^*u), and $\mathrm{EMD}(\mathbf{x}, (Y, c^*u)) = 0$. Furthermore, $\mathrm{EMD}(\mathbf{x}, (Y, cu)) = 0$ for $c \in (0, c^*]$ since there is still enough image weight of each pattern color to match all the weight in (Y, cu).

In general, we will prove that $\mathrm{EMD}(\mathbf{x}, (Y, cu))$ decreases as c decreases and eventually becomes constant for $c \in (0, c^0]$, as shown in Figure 1(e). If the graph levels off at a small EMD, then the pattern might occur in the image, and we take c^0 to be the scale estimate. Consider the example in Figure 1. The scale estimate c^0 is such that the amounts of red and yellow in the scaled pattern signature (Y, c^0u) are roughly equal to the amounts of red and yellow in the image, as shown in Figure 1(c). At scale c^0, there is still plenty of image weight to match the other pattern colors in (Y, c^0u). If there were a bit more red and yellow in the image, then the scale estimate c^0 would be a bit too high ($> c^*$).

The main property of our scale estimation method is that in the ideal case it overestimates the scale by the *minimum* amount of background clutter over all pattern colors, where the amount of background clutter for a color is the amount of that color present in the image but not part of the pattern occurrence. In practice, we have observed scale estimates which are a little smaller than predicted by an ideal case analysis. Just one pattern color with a small amount of background clutter in the image is enough to obtain an accurate scale estimate. Note that an accurate scale estimate is computed in Figure 1 even though there is a lot of background clutter for the dark green in the Comet label.

Now consider the function $E(c) = \mathrm{EMD}(\mathbf{x}, (Y, cu))$. The distribution (Y, cu) has total weight $c \leq 1 = w_\Sigma$, so

$$E(c) = \frac{\min_{(f_{ij}) \in \mathcal{F}(\mathbf{x},(Y,cu))} \sum_{i=1}^m \sum_{j=1}^n f_{ij} d(x_i, y_j)}{c},$$

where $(f_{ij}) \in \mathcal{F}(\mathbf{x}, (Y, cu))$ iff $f_{ij} \geq 0$, $\sum_{i=1}^m f_{ij} = cu_j$, and $\sum_{j=1}^n f_{ij} \leq w_i$. Now set $h_{ij} = f_{ij}/c$. Then

$$E(c) = \min_{(h_{ij}) \in \mathcal{F}((X,w/c),\mathbf{y})} \sum_{i=1}^m \sum_{j=1}^n h_{ij} d(x_i, y_j),$$

where $(h_{ij}) \in \mathcal{F}((X, w/c), \mathbf{y})$ iff (A) $h_{ij} \geq 0$, (B) $\sum_{i=1}^m h_{ij} = u_j$, and (C) $\sum_{j=1}^n h_{ij} \leq w_i/c$. Note that

$$\mathcal{F}((X, w/c_1), \mathbf{y}) \subseteq \mathcal{F}((X, w/c_2), \mathbf{y}) \iff c_2 \leq c_1 \quad (2)$$

since the final constraints (C) get weaker (stronger) as c decreases (increases).

Since $E(c)$ is a minimum over $\mathcal{F}((X, w/c), \mathbf{y})$, it follows from (2) that $E(c_1) \geq E(c_2)$ iff $c_1 \geq c_2$. Now consider $Q \subseteq \mathbf{R}^{mn}$ defined by (A) and (B), and $P(c) \subseteq \mathbf{R}^{mn}$ defined by (C), so that $\mathcal{F}((X, w/c), \mathbf{y}) = Q \cap P(c)$. Q is bounded since its constraints imply that $h_{ij} \in [0, u_j]$. The polytope $P(c)$ converges to $\mathbf{R}^{mn}$ as c decreases to zero since $1/c$ increases to ∞. Since Q is bounded, $\exists\, c^0$ for which $Q \subseteq P(c)\ \forall c \leq c^0$. It follows that $\mathcal{F}((X, w/c), \mathbf{y}) = Q$ and $E(c) = E(c^0)$ for $c \leq c^0$.

The scale estimation problem is the $\mathrm{EMD}_\mathcal{G}$ problem $c^0 = \max\ \arg\ \min_{g_c, 0<c\leq 1}\ \mathrm{EMD}(\mathbf{x}, g_c(\mathbf{y}))$, where $g_c(\mathbf{y}) = (Y, cu)$. In practice, we take as the scale estimate the largest c for which there is no real improvement in the EMD when c is decreased. The estimate c^0 can be found efficiently via binary search. See Figure 2. Initially, we assume $c^0 \in [c_{\min}, 1]$. At any step, we have localized $c^0 \in [c_{\text{low}}, c_{\text{high}}]$. Let $c_{\text{mid}} = (c_{\text{low}} + c_{\text{high}})/2$. If $E(c_{\text{mid}}) = E(c_{\text{low}})$, then $c^0 \in [c_{\text{mid}}, c_{\text{high}}]$. Here "=" is approximate with respect to a parameter ε_d. Otherwise, $E(c_{\text{mid}}) > E(c_{\text{low}})$ and $c^0 \in [c_{\text{low}}, c_{\text{mid}}]$. The search stops once $|c_{\text{high}} - c_{\text{low}}| \leq \varepsilon_c$, the required accuracy. In the SEDL image retrieval system, the optimal flow when $c = c^0$ is used to find quickly image regions similar in color signature to the query pattern ([2], pp. 189–200).

Figure 3 shows some results of the scale estimation algorithm for the color pattern problem. The scale estimate is very accurate in the examples shown in Figure 3(a)-(c). In the example shown in Figure 3(d), the scale is overestimated because the pattern occurs twice within the image. Since our method does not use the positions of colors, it cannot tell the difference between one pattern occurrence at scale $c_1 + c_2$ and two pattern occurrences at scales c_1 and c_2. See pp. 85–96 in [2] for more details and examples.

3.2. Point Transformations

In contrast to the previous section, we now consider sets of transformations that modify the points of a distribution but leave its weights fixed. Since $g(\mathbf{y}) = (g(Y), u)$ has the same weights as $\mathbf{y}$, we have $\mathcal{F}(\mathbf{x}, \mathbf{y}) = \mathcal{F}(\mathbf{x}, g(\mathbf{y}))$ and

$$\mathrm{EMD}_\mathcal{G}(\mathbf{x}, \mathbf{y}) = \frac{\min_{g \in \mathcal{G}, F \in \mathcal{F}(\mathbf{x},\mathbf{y})} \mathrm{WORK}(F, \mathbf{x}, g(\mathbf{y}))}{\min(w_\Sigma, u_\Sigma)}. \quad (3)$$

$W(F, g) = \mathrm{WORK}(F, \mathbf{x}, g(\mathbf{y}))$ is linear in F, so the minimum value in (3) occurs at one of the vertices of the convex polytope $\mathcal{F}(\mathbf{x}, \mathbf{y})$. Therefore, we can compute $\mathrm{EMD}_\mathcal{G}(\mathbf{x}, \mathbf{y})$ by solving $\min_{g \in \mathcal{G}} W(F, g)$ for each vertex F of $\mathcal{F}(\mathbf{x}, \mathbf{y})$. Although this strategy is guaranteed to find a globally optimal transformation, it is not practical because the number of vertices of $\mathcal{F}(\mathbf{x}, \mathbf{y})$ is usually very large even for relatively small values of m and n.

Given F or g, we can solve for an optimal value of the other. This leads to an iteration which alternates between finding the best flow for a given transformation, and the best transformation for a given flow. It generates a sequence of (F, g) pairs for which W decreases or remains constant at every step. The details are given in the next section.

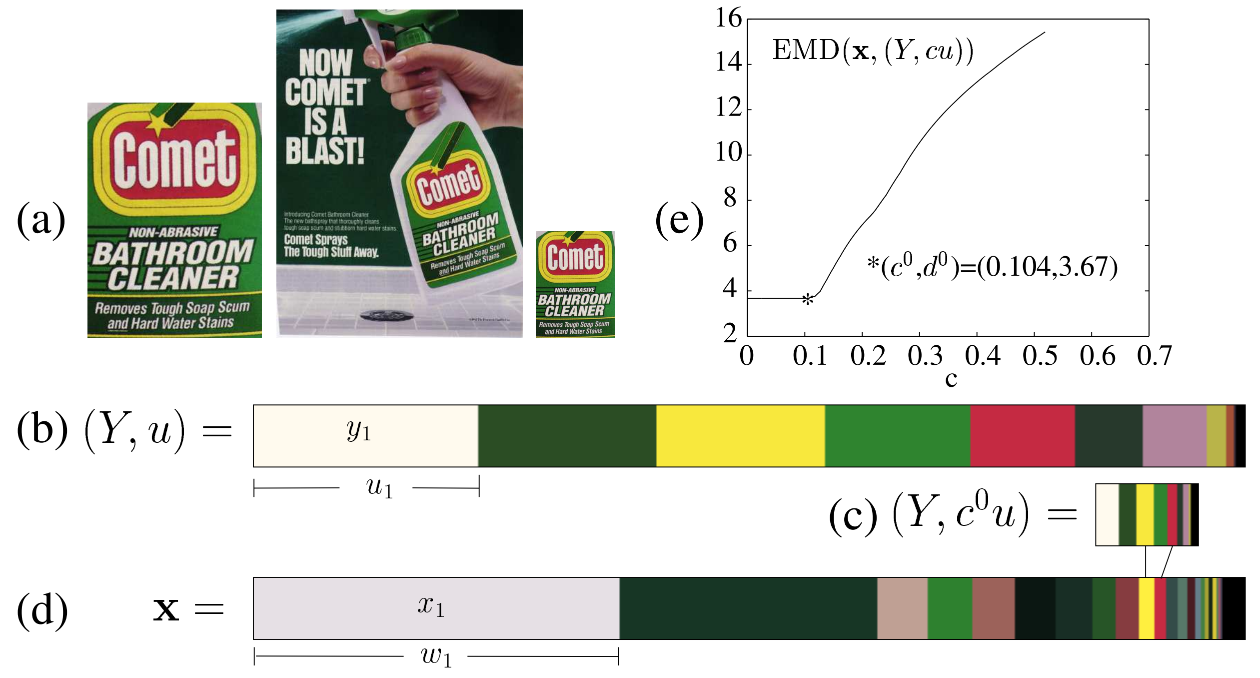

Figure 1. Scale Estimation. (a) pattern, image, and pattern scaled by the scale estimate c^0. (b),(d) pattern, image signatures. (c) pattern signature with weights scaled by c^0. (e) $\mathrm{EMD}(\mathbf{x}, (Y, cu))$ v. c.

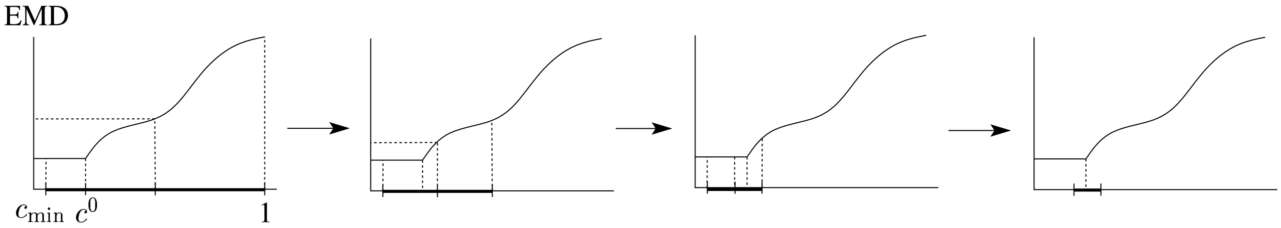

Figure 2. Scale Estimation Algorithm. Binary search narrows the interval in which c^0 must occur.

Figure 3. Scale Estimation Results. See the text for discussion.

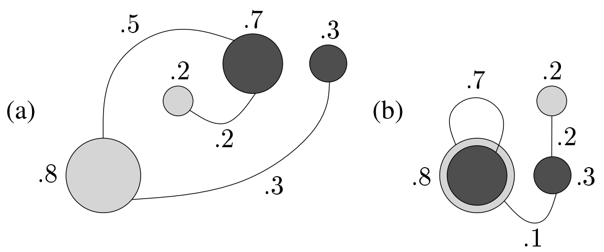

Figure 4. FT Iteration example. See the text.

3.2.1. The FT Iteration

Consider the following iteration that begins with an initial transformation $g^{(0)}$:

$$F^{(k)} = \arg \min_{F \in \mathcal{F}(\mathbf{x},\mathbf{y})} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d(x_i, g^{(k)}(y_j)), \quad (4)$$

$$g^{(k+1)} = \arg \min_{g \in \mathcal{G}} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^{(k)} d(x_i, g(y_j)). \quad (5)$$

The minimization in (4) is the transportation problem. Since this iteration alternates between finding an optimal Flow and an optimal Transformation, we refer to (4) and (5) as the *FT iteration*. It can be applied to equal-weight and unequal-weight distributions.

Figure 4(a) shows an example with a dark and a light distribution that we will match under translation starting with $g^{(0)} = 0$. The best flow $F^{(0)}$ for $g^{(0)}$ is shown by the labelled arcs connecting dark and light weights. This flow matches half (.5) the weight over a large distance. We should expect the best translation for $F^{(0)}$ to move the .7 dark weight closer to the .8 light weight in order to decrease the total amount of work done by $F^{(0)}$. Indeed, $g^{(1)}$ aligns these two weights as shown in Figure 4(b). The best flow $F^{(1)}$ for $g^{(1)}$ matches all of the .7 dark weight to the .8 light weight. No further translation improves the work – $g^{(2)} = g^{(1)}$ and the FT iteration converges.

Define $\text{WORK}^{(k)} = W\left(F^{(k)}, g^{(k)}\right)$. Then (4) and (5) imply $W\left(F^{(k+1)}, g^{(k+1)}\right) \leq W\left(F^{(k)}, g^{(k+1)}\right)$ and $W\left(F^{(k)}, g^{(k+1)}\right) \leq W\left(F^{(k)}, g^{(k)}\right)$, respectively (by definition, $F^{(k+1)}$ is optimal for $g^{(k+1)}$, and $g^{(k+1)}$ is optimal for $F^{(k)}$). It follows that $\text{WORK}^{(k+1)} \leq \text{WORK}^{(k)}$. The decreasing sequence $\text{WORK}^{(k)}$ is bounded below by zero, and hence it converges. There is, however, no guarantee that it converges to the global minimum of $\text{WORK}(F, \mathbf{x}, g(\mathbf{y}))$. In general, the iteration must be repeated with different $g^{(0)}$s in search of a globally optimal transformation.

It is easy to see that transformations which are only locally optimal can occur in unequal-weight cases. If $\mathbf{x}$ is L copies $\mathbf{y} \oplus t_l$ of $\mathbf{y}$, then $\text{EMD}(\mathbf{x}, \mathbf{y} \oplus t_l) = 0$ for $l = 1, \ldots, L$. If the copies of $\mathbf{y}$ in $\mathbf{x}$ are well-separated, then we can produce $\geq L - 1$ only locally optimal translations by slightly perturbing the points in each copy of $\mathbf{y}$. We have observed that only locally optimal transformations can also occur in equal-weight cases ([2], pp. 163–170).

The FT iteration can also be applied with the partial EMD since $\mathcal{F}^{\gamma}(\mathbf{x}, g(\mathbf{y})) = \mathcal{F}^{\gamma}(\mathbf{x}, \mathbf{y})$ if g does not change distribution weights. Furthermore, it can be modified to give a decreasing EMD sequence if a transformation changes points *and* modifies weights by a factor c. Such problems arise, for example, if a distribution point contains the position of an image region with some property, the corresponding weight is the region area, and a similarity transformation of the image plane is allowed. The basic idea is to choose $F^{(k)}$ from an increasing sequence of flow sets $\mathcal{F}^{(k)}$. Then $W\left(F^{(k+1)}, g^{(k+1)}\right) \leq W\left(F^{(k)}, g^{(k+1)}\right)$ since $F^{(k+1)}$ is an optimal flow for $g^{(k+1)}$ chosen from $\mathcal{F}^{(k+1)}$, and $F^{(k)} \in \mathcal{F}^{(k)} \subseteq \mathcal{F}^{(k+1)}$. The change of variables $h_{ij} = f_{ij}/c$ (as used in section 3.1) yields an equivalent EMD problem in which the weight of the lighter distribution is constant throughout the iteration, and hence a decreasing WORK sequence gives a decreasing EMD sequence. See pp. 148–151 in [2] for details.

The FT iteration is similar to the ICP (Iterative Closest Point) iteration ([1]) used to register 3D shapes. The computation of an optimal flow plays the same role as the computation of the closest "model shape" points to the "data shape" points in the ICP iteration. Both these steps determine matches used to compute a transformation that improves the EMD/registration. Another well-known application of the alternation idea is the EM algorithm ([7]) for computing mixture models in statistics.

The FT iteration can be applied whenever the *optimal transformation problem* (5) can be solved. If we let $[a_1 \cdots a_N] = [x_1 \cdots x_1 x_2 \cdots x_2 \cdots x_m \cdots x_m]$, $[b_1 \cdots b_N] = [y_1 \cdots y_n y_1 \cdots y_n \cdots y_1 \cdots y_n]$, and $[c_1 \cdots c_N] = [f_{11} \cdots f_{1n} f_{21} \cdots f_{2n} \cdots f_{m1} \cdots f_{mn}]$, where $N = mn$, then (5) can be rewritten as $\min_{g \in \mathcal{G}} \sum_{k=1}^{N} c_k d(a_k, g(b_k))$. Given a correspondence between point sets, the goal is to find a transformation of the points in one set that minimizes the sum of weighted distances to corresponding points in the other set.

The above problem has been solved with $d = L_2^2$ for translation (straightforward calculus), Euclidean and similarity transformations ([14]), linear transformations ([3]), and affine transformations (easy extension to the linear solution). The optimal translation problems with $d = L_2$ and $d = L_1$ are covered in [15], while the case $d = L_{1,T}$, the L_1 distance in a circular domain with period T (e.g. angles with $T = 2\pi$), is covered on pp. 142–146 in [2]. This last distance arises in the previously discussed texture similarity application in allowing for unpenalized differences in texture orientation (pp. 135–137 in [2], [11]). We show the generality of the FT iteration in the next two sections by applying it for a few different $\mathcal{G}$s and with the partial EMD.

(B)alloon (C)halk (D)ragon (L)emur (P)lant (T)iger (W)aldo

Figure 5. Object Database. For some objects, signatures are computed over only the outlined area.

3.2.2. Lighting-Invariant Object Recognition

For a linear, trichromatic color imaging system with a 3D linear model for the reflectance functions of object surfaces, Healey and Slater ([4]) showed that an illumination change results in a linear transformation of image pixel colors. The following experiment uses a subset of the images in [4]. There are four images of each object, one under nearly white illumination and the other three under yellow, green, and red illumination. See Figure 5 for images of the objects under white light.

Images are indexed by unit-weight color distributions in the RGB color space. Our experiment[3] uses each image as the query, where the desired distance is the EMD under a linear transformation with $d = L_2^2$. To compare a database signature $\mathbf{x}$ to a query signature $\mathbf{y}$, we applied the FT iteration twice: once to transform $\mathbf{y}$ so that it is as close as possible to $\mathbf{x}$, and once to transform $\mathbf{x}$ so that it is as close as possible to $\mathbf{y}$. Both trials were started with $g^{(0)}$ equal to the identity map. The smaller of the results of the two trials is used as the distance between $\mathbf{x}$ and $\mathbf{y}$. Ideally, the closest images to the image of an object are the other three images of the same object.

Figure 6 shows the results of our experiment. These results are excellent, but not perfect as in [4]. It is possible that we are not finding the globally optimal transformation in some comparisons.

3.2.3. Point Feature Matching in Stereo Images

In this section, we use the partial EMD under a transformation set $\mathrm{EMD}_{\mathcal{G}}^{\gamma}$ to compute the best partial matching of two point feature sets extracted from stereo image pairs. The fraction parameter γ compensates for the fact that only some features appear in both images, and the set parameter $\mathcal{G}$ accounts for the appropriate transformation between corresponding features. In our experiments, we extract 50 features of an image using an algorithm due to Shi and Tomasi ([13]). See the first two columns of Figure 7. The points in the distribution summary of an image are its feature locations, and the weight of each point is one. The ground distance is $d = L_2^2$ between image coordinates. We set $\gamma = 0.5$, so only 25 of the 50 features per image will be matched, and use $g^{(0)} = I$, the identity map.

[3] All experiments in this work were done on a 250 MHz SGI Indigo^2.

In the first example, we match features in two images from a motion sequence in which the camera moves approximately horizontally and parallel to the image plane. Figure 7(a) shows the result of applying the FT iteration with $\mathcal{G} = \mathcal{T}$, the group of translations. For this camera motion, all image points translate along the same direction, but the amount of translation for an image point is inversely proportional to the depth of the corresponding scene point. The model of a single translation vector is accurate for a set of features that correspond to scene points with roughly the same depth. In this example, the FT iteration matched features on objects toward the back of the table.

The images in Figure 7(b) are from a motion sequence with a forward camera motion perpendicular to the image plane. Here we apply the FT iteration with $\mathcal{G} = \mathcal{S}$, the set of similarity transformations. In the final example, we match features in images of a toy hotel. The results of the FT iteration with $\mathcal{G} = \mathcal{A}$, the set of affine transformations, are shown in Figure 7(c). In all three cases, it appears that the FT iteration converged to a globally optimal transformation. In many examples, however, running the iteration once leads to only a locally optimal solution. In the next two sections, we consider two equal-weight EMD under translation problems which can be solved directly for a globally optimal translation.

3.2.4. Equal-Weight $\mathrm{EMD}_{\mathcal{T}}$ with $d = L_2^2$

It is easily proven that $\min_t \sum_{i=1}^m \sum_{j=1}^n f_{ij}||x_i - (y_j + t)||_2^2$ occurs at $t^* = (\sum_{i=1}^m \sum_{j=1}^n f_{ij}(x_i - y_j)) / \sum_{i=1}^m \sum_{j=1}^n f_{ij}$. In the equal-weight case, $F \in \mathcal{F}(\mathbf{x}, \mathbf{y})$ requires $\sum_{i=1}^m f_{ij} = u_j$ and $\sum_{j=1}^n f_{ij} = w_i$ since all the weight in both distributions must be matched. Using these facts, $t^* = \overline{\mathbf{x}} - \overline{\mathbf{y}}$, where $\overline{x} = \sum_{i=1}^m w_i x_i / w_\Sigma$ and $\overline{y} = \sum_{j=1}^n u_j y_j / u_\Sigma$ are the centroids of $\mathbf{x}$ and $\mathbf{y}$. The translation that lines up the centroids is optimal for every feasible flow. To compute $\mathrm{EMD}_{\mathcal{T}, L_2^2}(\mathbf{x}, \mathbf{y})$ for equal-weight $\mathbf{x}$ and $\mathbf{y}$, we simply compute $\mathrm{EMD}(\mathbf{x}, \mathbf{y} \oplus (\overline{\mathbf{x}} - \overline{\mathbf{y}}))$.

3.2.5. Equal-Weight $\mathrm{EMD}_{\mathcal{T}}$ in 1D with $d = L_1$

There is a simple solution to computing the EMD between equal-weight distributions in 1D with $d = L_1$ that involves the cumulative distribution functions (CDFs). See Figure 8(a). The CDF for $\mathbf{x}$ starts at 0, increases an amount w_i

	B_W	B_Y	B_G	B_R	C_W	C_Y	C_G	C_R	D_W	D_Y	D_G	D_R	L_W	L_Y	L_G	L_R	P_W	P_Y	P_G	P_R	T_W	T_Y	T_G	T_R	W_W	W_Y	W_G	W_R
W	1	3	2	3	1	4	2	3	1	2	3	2	1	4	6	2	1	4	2	3	1	3	2	2	1	4	3	3
Y	3	1	3	5	4	1	3	2	2	1	$\boxed{2}$	4	4	1	4	4	4	1	3	2	3	1	3	3	2	1	2	2
G	2	2	1	2	2	3	1	4	3	3	1	3	3	2	1	3	2	3	1	5	2	2	1	4	4	2	1	4
R	5	5	4	1	3	2	4	1	4	4	4	1	2	3	2	1	3	2	5	1	4	4	7	1	3	3	4	1
Σ	11	11	10	11	10	10	10	10	10	10	10	10	10	10	13	10	10	10	11	11	10	10	13	10	10	10	10	10

Figure 6. Query Results. The column labels are the query images, and the row labels are the illuminants (W)hite, (Y)ellow, (G)reen, and (R)ed. The boxed entry, for example, indicates that the yellow (Y) dragon image is returned as the second closest image for the green dragon (D_G) query image. The number at the bottom of each column is the total of the ranks in that column, where 10 is the ideal value. The query precision is perfect for 21 of the 28 queries, and the average rank sum is 10.4. One run of the FT iteration required an average of 7.4 steps and 4.6 seconds to converge.

at each point x_i, and eventually becomes w_Σ at the largest point x_m. The CDFs for $\mathbf{x}$ and $\mathbf{y}$ are the bold and regular thickness staircase graphs, respectively. Since $\mathbf{x}$ and $\mathbf{y}$ are equal-weight distributions, the two CDFs become constant at the same value $w_\Sigma = u_\Sigma$. The EMD is equal to the area between the CDFs (shaded) divided by the total weight ([2], pp. 71–80). The corresponding optimal *CDF flow* is indicated with arrows.

The CDF flow is given by $f_{ij}^{\mathrm{CDF}} = |[W_{i-1}, W_i] \cap [U_{j-1}, U_j]|$, where $W_k = \sum_{i=1}^{k} w_i$, $U_l = \sum_{j=1}^{l} u_j$, and $W_0 = U_0 = 0$. Here the points and weights in a distribution are numbered according to increasing position along the real line. The partial sums $U_0, U_1, \ldots, U_n$ are the same for every translated version of $\mathbf{y}$, so the CDF flow is an optimal flow between $\mathbf{x}$ and $\mathbf{y} \oplus t$ for every translation t. See Figure 8(b), where we have re-used the labels y_j instead of using $y_j + t$ in order to save space. To compute the EMD under translation in this case, we simply solve the optimal translation problem for $d = L_1$ ([15]) with $F = F^{\mathrm{CDF}}$.

4. Conclusion

The $\mathrm{EMD}_{\mathcal{G}}$ problem is an example of the common computer vision problem of simultaneously estimating dependent sets of parameters (e.g. shape and motion in structure from motion, or motions and groups in motion mixture models). Avoiding local minima during iterative improvement of the estimation is a challenging problem in general, and the difficulty is magnified in the $\mathrm{EMD}_{\mathcal{G}}$ problem because partial matching is allowed. Some cases with special structure that allow direct computation of a globally optimal transformation were identified. In the absence of such structure, however, an important area for future work is to develop efficient and effective strategies for choosing initial transformations for the FT iteration which are close to a global optimum, particularly in partial matching cases where choosing $g^{(0)}$ based on global statistics such as centroids and principal components will not work.

Acknowledgements

This research was sponsored in part by DARPA under contract DAAH04-94-G-0284. We thank Madirakshi Das for the color ads, David Slater for the object database, Stan Birchfield for his feature extraction code, and Yossi Rubner for his EMD code.

References

[1] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *PAMI*, 14(2):239–256, Feb. 1992.

[2] S. Cohen. Finding color and shape patterns in images. Technical Report STAN-CS-TR-99-1620, Stanford University, May 1999. Available online at `http://vision.stanford.edu/~scohen`.

[3] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1989.

[4] G. Healey and D. Slater. Global color constancy: recognition of objects by use of illumination-invariant properties of color distributions. *JOSA A*, 11(11):3003–3010, 1994.

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

[6] D. P. Huttenlocher et al. Comparing images using the Hausdorff distance. *PAMI*, 15(9):850–863, Sept. 1993.

[7] G. McLachlan and K. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1989.

[8] W. Niblack et al. The QBIC project: querying images by content using color, texture, and shape. In *Proceedings of the SPIE*, volume 1908, pages 173–187, 1993.

[9] Y. Rubner et al. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *ARPA IUW*, pages 661–668, May 1997.

[10] Y. Rubner et al. The earth mover's distance as a metric for image retrieval. Technical Report STAN-CS-TN-98-86, Stanford Computer Science Department, Sept. 1998.

[11] Y. Rubner et al. A metric for distributions with applications to image databases. In *ICCV*, pages 59–66, Jan. 1998.

[12] W. J. Rucklidge. Locating objects using the Hausdorff distance. In *ICCV*, pages 457–464, 1995.

[13] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, June 1994.

[14] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *PAMI*, 13(4):376–380, Apr. 1991.

[15] G. O. Wesolowsky. The Weber problem: History and perspectives. *Location Science*, 1(1):5–23, May 1993.

[16] G. Wyszecki and W. Styles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, 1982.

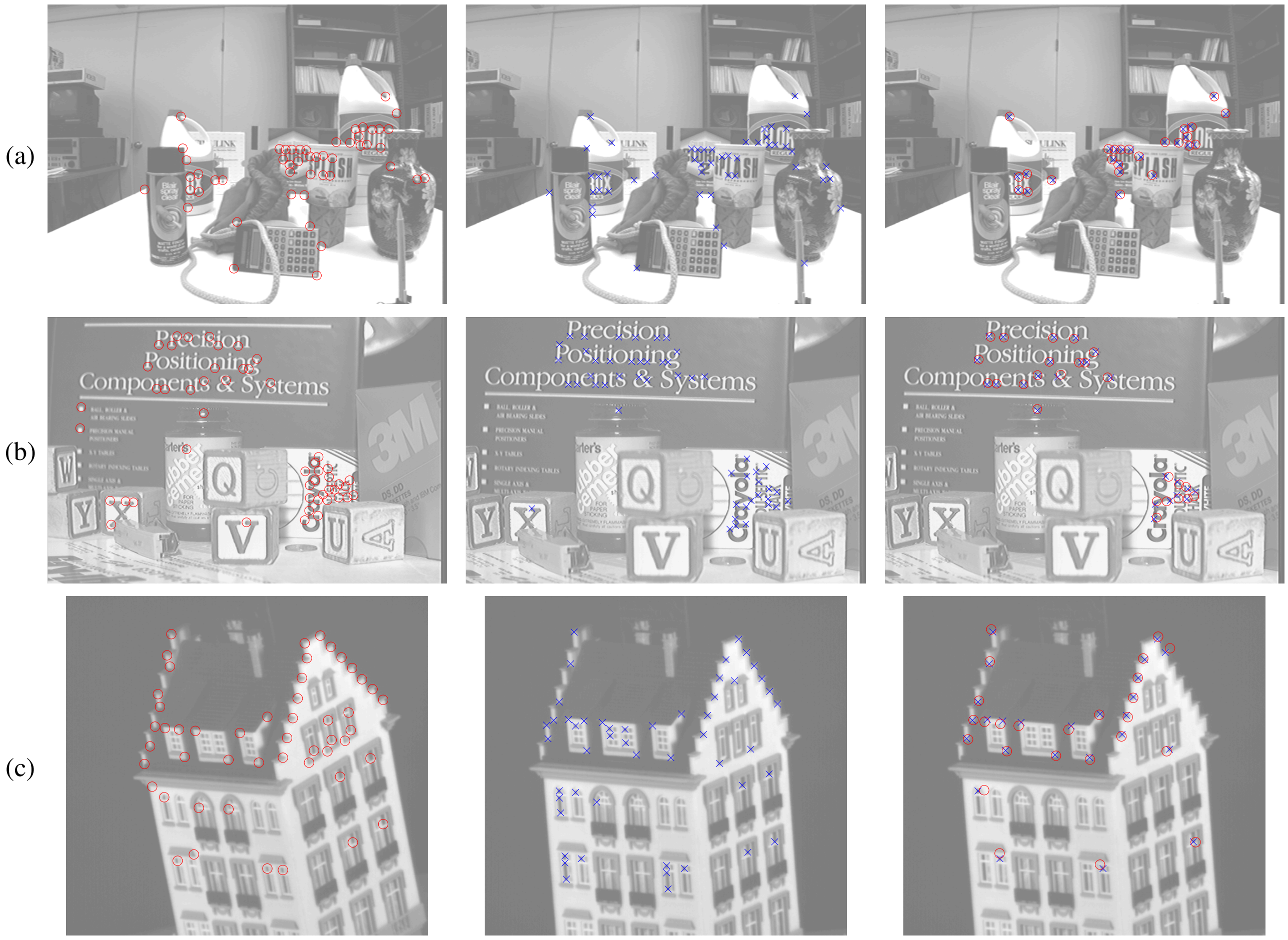

Figure 7. Point Set Matching. See the text. We report the number of steps S and the time T in seconds (s) for the FT iteration to converge. (a) $S = 11$, $T = 1.8$s. (b) $S = 4$, $T = 1.1$s. (c) $S = 8$, $T = 36.2$s.

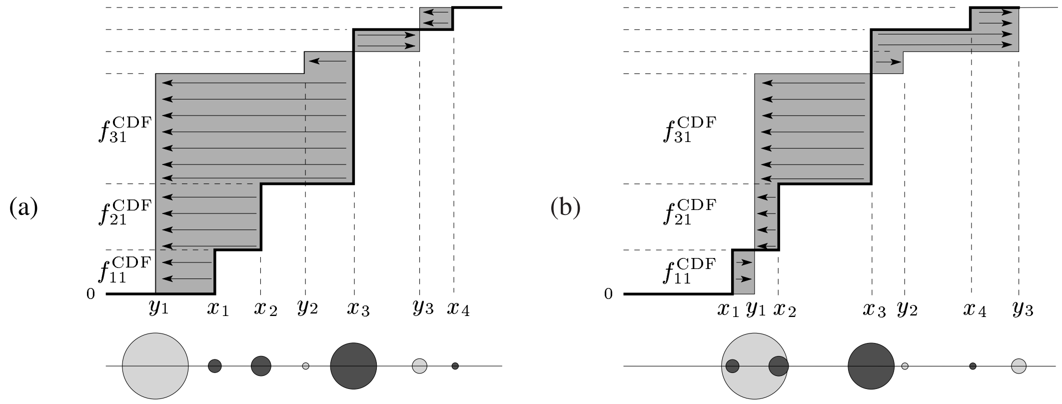

Figure 8. The Equal-Weight EMD under Translation in 1D with $d = L_1$. The same flow F^{CDF} is optimal for (a) x and y, and (b) x and $\mathrm{y} \oplus t$. See the text for details.