Supervised Earth Mover's Distance Learning and Its Computer Vision Applications

Fan Wang and Leonidas J. Guibas

Stanford University, CA, United States

Abstract. The Earth Mover's Distance (EMD) is an intuitive and natural distance metric for comparing two histograms or probability distributions. It provides a distance value as well as a flow-network indicating how the probability mass is optimally transported between the bins. In traditional EMD, the ground distance between the bins is pre-defined. Instead, we propose to jointly optimize the ground distance matrix and the EMD flow-network based on a partial ordering of histogram distances in an optimization framework. Our method is further extended to accept information from general labeled pairs. The trained ground distance better reflects the cross-bin relationships, hence produces more accurate EMD values and flow-networks. Two computer vision applications are used to demonstrate the effectiveness of the algorithm: first, we apply the optimized EMD value to face verification, and achieve state-of-the-art performance on the PubFig and the LFW data sets; second, the learned EMD flow-network is used to analyze face attribute changes, obtaining consistent paths that demonstrate intuitive transitions on certain facial attributes.

1 Introduction

Histogram-like descriptors such as SIFT [1], shape context [2], and Bag-of-Visual-Words (BoVW) [3] have been successfully applied to various computer vision tasks. To measure distance between two such descriptors, common choices are the L_2 distance, χ^2 distance, and K-L divergence. These distance metrics assume, however, that the histogram bins are perfectly aligned and only account for the difference between corresponding bins. The Earth Mover's Distance (EMD) [4] was proposed to mitigate this assumption. It accounts for cross-bin information, and has been shown to output perceptually natural distances for applications including face recognition [5], visual tracking [6], and shape matching [7].

To calculate the EMD, one needs to specify a ground distance matrix, which defines the distance between each pair of histogram bins. The choice of the ground distance matrix has mostly been empirical, ad hoc, and highly application-dependent. However, inaccurate ground distances can generate sub-optimal EMDs. At the same time, handcrafting a ground distance matrix for each specific application based on domain knowledge is very challenging and hard to generalize. For example, if we describe a face using a histogram-like descriptor based on its affinity or similarity to a set of predefined *reference identities*

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 442–455, 2012.

(Sec.3.1), where each bin corresponds to a collection of face images of one person (Fig.1), it is crucial to define a proper ground distance between the identities in order to compute meaningful EMDs. However, it is difficult to hand-craft a ground distance that agrees with the perceptual difference between these faces. As a result, we propose here to automatically learn an optimal ground distance.

Fig. 1. Histogram-like descriptor for a given face. Learning an accurate ground distance is crucial for good performance of EMD. Face images are from the *PubFig* data set.

In addition to the distance value, the EMD produces a flow-network representing how the mass in the histogram bins is transported. Little attention has been paid to this byproduct, but the flow-network actually contains much more information than the scalar distance value: it reveals how two histograms are cross-matched and how one histogram is transformed into the other. In image retrieval, if each image is represented by a histogram (such as BoVW), there could be many histograms (images) that differ from a given histogram (query image) by the same amount (same EMD value), but each individual flow-network explains how one image is dissimilar to the query in its own distinctive way.

In this paper, we propose to jointly learn the ground distance matrix and the flow-network of the EMD, so that the calculated EMD maximally agrees with the provided training information. The problem is formulated in a bi-convex optimization framework, which can take information from either partial ordering of the histogram distances or labeled pairs of the samples.

Two applications of the proposed method are discussed: one uses the learned EMD for face verification, and the other investigates the flow-networks to discover internal structures within a set of faces. These two applications clearly show the superiority of proposed method:

1. With the supervised EMD learning algorithm, we obtain a ground distance matrix that better reflects the true distance between histogram bins, yielding a more accurate EMD value.
2. In addition, we obtain flow-networks that better reflect the cross-bin matching between two histograms, yielding better description of their differences.

1.1 Related Work

Learning a proper distance metric is important for object recognition, image classification, image retrieval, etc. In distance learning, the training data are usually provided in the form of pairwise constraints: pairs of "similar"samples, and pairs of "dissimilar" ones. The learning algorithm then transforms the data

into a new space so that the distance metric in the new space agrees better with the supervision data. Some approaches learn a global distance function that satisfies all training constraints [8, 9], while others learn a local distance function that only agrees with local training information [10, 11].

Although distance learning methods for L_p distance, Mahalanobis distance, cosine similarity, etc., have been well studied, distance learning for EMD is largely unexplored. Unlike many distance functions, the EMD is the optimal value of a linear programming problem, and does not have an explicit form in general. Therefore, the classic approach of transforming the samples into a new space does not naturally apply to EMD. Cuturi et al. [12] formulated the problem as minimization of the difference between two polyhedral convex functions, but the training data are required to be pairwise similarity values between all training histograms. Wang et al. [13] assumed that the ground distance between histogram bins is Mahalanobis distance, and optimized over the positive semi-definite covariance matrix. Our proposed algorithm works in a more general setting both in terms of the input and the output: the input can be any training pairs of samples that are labeled as "similar" or "dissimilar", and the output is a general ground distance matrix rather than a covariance matrix.

There are many potential applications of the learned EMD. In this paper, we investigate its application in face verification. Distance learning for face verification has been studied in several ways. A mapping function was learned to map input faces into a space in which the L_1 distance approximates the semantic distance [14]. A logistic discriminant approach and a nearest neighbor approach were also proposed to learn the metric from labeled face pairs [15]. A transformation matrix was learned so that cosine similarity between faces performs well in the transformed subspace [16].

Attributes are also powerful tools to describe a face. For example, categorical attributes have been used for face verification [17]. Besides detecting presence or absence of attributes, a ranking function was learned to predict the relative strength of an attribute [18]. These works assigned attributes to each identity, while certain attributes such as pose, expression, etc., are associated with individual face images. In this paper, we use the flow-network to automatically detect the transition "direction" of certain attributes, and organize face images into consistent paths along these directions.

2 Supervised Earth Mover's Distance

The Earth Mover's Distance (EMD) was introduced to the computer vision community as a technique for image retrieval [4] . It is also known as the Mallows distance [19] in statistical literature. In its continuous form, it is a special case of the general Monge-Kantorovich class of transportation metrics [20, 21], also known as Wasserstein distances. We first introduce the basic setups of EMD, and then describe our formulation of supervised EMD learning.

2.1 Earth Mover's Distance

The EMD is a distance measurement between two histograms or distributions, referred to as the *source* and the *destination*, respectively. The source histogram $\mathbf{p} \in \mathbb{R}^n$ is regarded as piles of earth at various locations (bins). The amount of earth in each pile equals to the value of each corresponding bin. The destination histogram $\mathbf{q} \in \mathbb{R}^n$ is regarded as several holes, the values of which represent their capacities. The EMD equals to the minimum amount of effort required to move all the earth from the source to the destination. A ground distance matrix $D = \{d_{ij}\}$ contains elements d_{ij} that defines the cost of moving one unit of earth from the i-th bin of $\mathbf{p}$ to the j-th bin of $\mathbf{q}$.

To solve for EMD, a flow matrix $F = \{f_{ij}\}$ needs to be found with f_{ij} denoting the amount of earth moved from the i-th bin of $\mathbf{p}$ to the j-th bin of $\mathbf{q}$. A convex feasible set for the flow matrix is defined with respect to $\mathbf{p}$ and $\mathbf{q}$ as:

$$\mathbb{C}(\mathbf{p}, \mathbf{q}) = \left\{ \mathbf{f} \mid \mathbf{f} = \mathrm{vec}(F), F \in \mathbb{R}^{n \times n}, F^T \mathbf{1} = \mathbf{q}, F\mathbf{1} = \mathbf{p}, f_{ij} \geq 0, \forall\, i, j \right\}, \quad (1)$$

where we rewrite the flow matrix F into a vector $\mathbf{f}$ for notation simplicity. This feasible set ensures the nonnegativity of the flows, and enforces conservation of the earth amount and the hole capacity. Here we only consider the case when $\mathbf{1}^T\mathbf{p} = \mathbf{1}^T\mathbf{q}$, but the un-normalized case can be dealt with by normalizing the two histograms to have the same L_1 norm, or by moving as much earth as feasible. We also rewrite the ground distance matrix D in a vector form $\mathbf{d} = \mathrm{vec}(D)$. The EMD problem is then formulated as a convex optimization problem:

$$\mathrm{EMD}(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{f} \in \mathbb{C}(\mathbf{p}, \mathbf{q})} \mathbf{d}^T \mathbf{f}, \quad (2)$$

which is a linear programming problem, and more specifically, a transportation problem. Efficient computation of EMD has been studied extensively [22–24].

2.2 Supervised EMD Learning with Triplets

The EMD is highly influenced by the ground distance. However, in traditional EMD, the ground distance is usually predefined as Euclidean distance, city-block distance, etc., which can be inaccurate in many cases. Here we propose to automatically learn the ground distance based on supervised information.

Suppose we have N triplets of histograms $\{(\mathbf{p}_i, \mathbf{q}_i, \mathbf{r}_i), i = 1, \cdots, N\}$. For each triplet, it's given that the distance between $\mathbf{p}_i$ and $\mathbf{r}_i$ is no smaller than that between $\mathbf{p}_i$ and $\mathbf{q}_i$, either by explicitly comparing the pairs, or by knowing that $\mathbf{p}_i$ and $\mathbf{q}_i$ belong to the same class while $\mathbf{p}_i$ and $\mathbf{r}_i$ don't. Therefore,

$$\mathrm{EMD}(\mathbf{p}_i, \mathbf{r}_i) \geq \mathrm{EMD}(\mathbf{p}_i, \mathbf{q}_i), \ \forall i. \quad (3)$$

Intuitively, we'd like to learn a ground distance $\mathbf{d}$ so that the resulting EMD values satisfy as many constraints as possible. However, this combinatorial problem is NP-hard. Instead, we allow slackness in each constraint:

$$\mathrm{EMD}(\mathbf{p}_i, \mathbf{r}_i) - \mathrm{EMD}(\mathbf{p}_i, \mathbf{q}_i) \geq 1 - \xi_i, \ \xi_i \geq 0, \ \forall i, \quad (4)$$

where the non-negative slack variables ξ_i allows violation of constraint i with certain penalty in the objective function. Use $\mathbf{f}_i$ to denote the optimal flow between $\mathbf{p}_i$ and $\mathbf{q}_i$, and $\mathbf{g}_i$ to denote the optimal flow between $\mathbf{p}_i$ and $\mathbf{r}_i$, Eq.4 can be rewritten as $\text{EMD}\left(\mathbf{p}_i, \mathbf{r}_i\right) - \text{EMD}\left(\mathbf{p}_i, \mathbf{q}_i\right) = \mathbf{d}^T(\mathbf{g}_i - \mathbf{f}_i) \geq 1 - \xi_i$. We construct two matrices using the flows from all of the N training triplets: $M_f = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]$ and $M_g = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N]$. The overall problem is finally formulated as:

$$\begin{aligned} \text{min.} \quad & \|\mathbf{d}\|_2^2 + C \cdot \boldsymbol{\xi}^T \mathbf{1} \\ \text{s.t.} \quad & \mathbf{d}^T \left(M_g - M_f\right) \geq \mathbf{1}^T - \boldsymbol{\xi}^T,\ \boldsymbol{\xi} \geq 0,\ \mathbf{d} \in \mathbb{D} \\ & \mathbf{f}_i = \underset{\mathbf{f} \in \mathbb{C}(\mathbf{p}_i, \mathbf{q}_i)}{\text{argmin}}\ \mathbf{d}^T \mathbf{f},\ \mathbf{g}_i = \underset{\mathbf{g} \in \mathbb{C}(\mathbf{q}_i, \mathbf{r}_i)}{\text{argmin}}\ \mathbf{d}^T \mathbf{g},\ \forall i, \end{aligned} \tag{5}$$

where C is a constant that makes a trade-off between the empirical error and the L_2 norm of $\mathbf{d}$, and helps to control overfitting. Vector $\boldsymbol{\xi} \in \mathbb{R}^N$ collects all slack variables. The convex feasible domain for the ground distance is

$$\mathbb{D} = \left\{ \mathbf{d} \mid \mathbf{d} = \text{vec}(D), D \in \mathbb{R}^{n \times n},\ D_{ij} \geq 0, D_{ii} = 0, \forall i, j \right\}. \tag{6}$$

Feasible sets $\mathbb{C}(\mathbf{p}_i, \mathbf{q}_i)$ and $\mathbb{C}(\mathbf{p}_i, \mathbf{r}_i)$ are denoted in the same way as in Eq. 1.

2.3 Supervised EMD in a More General Setting

The method proposed in Sec. 2.2 can be extended to cases where triplet comparisons are not available. Instead, supervision information is provided by sets of similar pairs $\{(\mathbf{p}_i, \mathbf{q}_i), i = 1, \cdots, N_s\}$ and dissimilar pairs $\{(\mathbf{r}_j, \mathbf{s}_j), j = 1, \cdots, N_d\}$. These pairs do not share common samples, therefore cannot form triplets. However, the intuition that similar pairs should have smaller EMD compared to dissimilar pairs still holds. The algorithm finds a ground distance matrix such that the two sets of distances, $\{\text{EMD}\left(\mathbf{p}_i, \mathbf{q}_i\right), i = 1, \cdots, N_s\}$ and $\{\text{EMD}\left(\mathbf{r}_j, \mathbf{s}_j\right), j = 1, \cdots, N_d\}$, are separated as much as possible. The problem is naturally transformed into a max-margin problem:

$$\begin{aligned} \text{min.} \quad & \|\mathbf{d}\|_2^2 + C\left(\boldsymbol{\xi}_f^T \mathbf{1} + \boldsymbol{\xi}_g^T \mathbf{1}\right) \\ \text{s.t.} \quad & \mathbf{d}^T M_f \leq -\mathbf{1}^T + \boldsymbol{\xi}_f^T + t,\ \boldsymbol{\xi}_f \geq 0 \\ & \mathbf{d}^T M_g \geq \mathbf{1}^T - \boldsymbol{\xi}_g^T + t,\ \boldsymbol{\xi}_g \geq 0,\ \mathbf{d} \in \mathbb{D} \\ & \mathbf{f}_i = \underset{\mathbf{f} \in \mathbb{C}(\mathbf{p}_i, \mathbf{q}_i)}{\text{argmin}}\ \mathbf{d}^T \mathbf{f},\ \mathbf{g}_i = \underset{\mathbf{g} \in \mathbb{C}(\mathbf{r}_i, \mathbf{s}_i)}{\text{argmin}}\ \mathbf{d}^T \mathbf{g},\ \forall i, \end{aligned} \tag{7}$$

where the scalar t can be an arbitrary separation threshold of the two sets of EMDs, allowing the linear classifier not to pass the origin. M_f and M_g are two matrices formed by the flows of similar pairs and dissimilar pairs, respectively. The slack variables are collected in $\boldsymbol{\xi}_f \in \mathbb{R}^{N_s}$ and $\boldsymbol{\xi}_g \in \mathbb{R}^{N_d}$ respectively.

2.4 Solving for Optimal Ground Distance

The learning-by-triplet idea has been used for Euclidean distance learning in a convex optimization setting [10]. However, EMD itself is an optimization problem, and the overall problem here is not convex any more. It is instead bi-convex with respect to the two sets of variables $\{\mathbf{d}\}$ and $\{M_f, M_g\}$.

More specifically, take the optimization problem in Eq. 5 as an example, given the ground distance $\mathbf{d}$, the problem can be decoupled into $2N$ independent standard EMD problems, solving for $\mathbf{f}_i$ and $\mathbf{g}_i$, respectively. Given the flows M_f and M_g, the problem can be re-written as:

$$\begin{aligned} \text{min.} \quad & \|\mathbf{d}\|_2^2 + C \cdot \boldsymbol{\xi}^T \mathbf{1} \\ \text{s.t.} \quad & \mathbf{d}^T (M_g - M_f) \geq \mathbf{1}^T - \boldsymbol{\xi}^T,\ \boldsymbol{\xi} \geq 0,\ \mathbf{d} \in \mathbb{D}, \end{aligned} \tag{8}$$

which is a Quadratic Programming (QP) that is similar to the soft-margin SVM.

Similarly, for the the optimization problem in Eq. 7, if $\mathbf{d}$ is fixed, it decouples into $N_s + N_d$ independent linear programming problems. If the flows $\mathbf{f}_i$ and $\mathbf{g}_i$ are fixed, the problem can again be rewritten as a QP:

$$\begin{aligned} \text{min.} \quad & \|\mathbf{d}\|_2^2 + C \left(\boldsymbol{\xi}_f^T \mathbf{1} + \boldsymbol{\xi}_g^T \mathbf{1}\right) \\ \text{s.t.} \quad & \mathbf{d}^T M_f \leq -\mathbf{1}^T + \boldsymbol{\xi}_f^T + t,\ \boldsymbol{\xi}_f \geq 0 \\ & \mathbf{d}^T M_g \geq \mathbf{1}^T - \boldsymbol{\xi}_g^T + t,\ \boldsymbol{\xi}_g \geq 0,\ \mathbf{d} \in \mathbb{D}. \end{aligned} \tag{9}$$

If the flows $\mathbf{f}_i$ and $\mathbf{g}_j$ are regarded as high-dimensional sample points of two classes, the algorithm is essentially looking for a SVM classifier to separate these two classes of samples with the largest margin.

Finally, the supervised EMD learning problem is solved using an alternating optimization framework as below:

Input: Initial estimation of the ground distance matrix $\mathbf{d}^0$ using Euclidean distance or any other suitable metric, threshold ε, and damping factor α.
$k = 0$;
while $\left\|\mathbf{d}^k - \mathbf{d}^{k-1}\right\|_2 \geq \varepsilon$ **do**
  Given ground distance $\mathbf{d}^{k-1}$, solve for the flows M_f^k and M_g^k;
  Given the flows M_f^k and M_g^k, solve for the ground distance $\mathbf{d}^k$ using Eq. 8 or Eq. 9;
  $\mathbf{d}^k \leftarrow \mathbf{d}^{k-1} + \alpha(\mathbf{d}^k - \mathbf{d}^{k-1})$;
  $k \leftarrow k + 1$;
end while

At each iteration, $\mathbf{d}^{k-1}$ from the last update is used as the objective function coefficients to solve for new flows M_f^k and M_g^k through linear programming (Eq.2), whose feasible region is kept unchanged during all iterations. If $\Delta\mathbf{d} = \mathbf{d}^{k-1} - \mathbf{d}^{k-2}$ is sufficiently small, although the optimal value of the LP might change, the optimal solutions $\mathbf{f}_i$ and $\mathbf{g}_i$ will be unchanged [25]. Therefore, $\mathbf{d}$ will not change in the next iteration, meaning the iteration stops and the overall algorithm converges.

The damping factor $0 < \alpha \leq 1$ is introduced to further aid convergence and control the trade-off between convergence speed and stability of the iterations. A small α creates conservative but stable steps in the iterations. With a large α, the iterations are less stable, but have higher chances to jump out of local minimum. α is fixed as 0.1 in our experiments. It's also possible to have a varying α that is

large at the beginning and shrinks with iterations. The algorithm is implemented using `CVX`, a package for specifying and solving convex programs [26].

3 Face Verification Using Supervised EMD

3.1 A Face Descriptor Based on Reference Identities

In this section, we apply the proposed supervised EMD algorithm to face verification, where each face is represented by a histogram-like descriptor as described below.

Often times, we describe someone's facial appearance as "he looks more like John, but not like Jack". What our brain might be doing when recognizing a face is comparing an unknown face with the face templates in our memories. Based on this intuition, we utilize a face descriptor based on the similarity values of the face to some known templates:

We first select a set of known identities, called *reference identities*, each represented by a set of diverse face images of one person. These sets serve as basis to represent a new test face in the global face space. We denote the faces of the i-th reference person as a matrix X_i, with each column being a face of this person. The total K identities in the reference data set are denoted collectively as $X = [X_1, X_2, \cdots, X_K]$. A test face $\mathbf{y}$ is then reconstructed by all the faces in the reference set with a L_p regularization term:

$$\min. \quad \|\mathbf{y} - X\boldsymbol{\alpha}\|_2 + \lambda \|\boldsymbol{\alpha}\|_p , \tag{10}$$

where we chose $p = 2$ since it has been shown to be robust to noise [27].

The solution of the convex optimization problem above gives an encoding of the test face as $\mathbf{y} \approx X\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1; \boldsymbol{\alpha}_2; \ldots; \boldsymbol{\alpha}_K]$ with $\boldsymbol{\alpha}_i$ containing the coefficients associated with the i-th reference identity. It is observed that if $\mathbf{y}$ is from the i-th class, most coefficients in $\boldsymbol{\alpha}_k$ are close to zero for $k \neq i$, and only $\boldsymbol{\alpha}_i$ has significant entries. The reconstruction error using only the coefficients from the i-th identity gives a strong indication of the affinity between the face and the i-th identity. We calculate the reconstruction error as:

$$e_i(\mathbf{y}) = \|\mathbf{y} - X_i\boldsymbol{\alpha}_i\|_2 . \tag{11}$$

The vector of the reconstruction errors, $\mathbf{e}(\mathbf{y}) = [e_1(\mathbf{y}), e_2(\mathbf{y}), \ldots, e_K(\mathbf{y})]$, describes the relationship between the test face and each reference identity, even if $\mathbf{y}$ does not belong to any of these identities.

Finally the error vector $\mathbf{e}(\mathbf{y})$ is transformed to a similarity score vector $\mathbf{s}(\mathbf{y})$ by a Gaussian function $s_i(\mathbf{y}) = \exp(-\frac{1}{2\sigma_i^2}(e_i(\mathbf{y}) - \mu_i)^2)$, and normalized to have L_1 norm equal to unity, giving our final histogram-like face descriptor as illustrated in Fig.1. We chose μ_i and σ_i^2 as the mean and variance of all the reconstruction errors for the i-th identity.

The idea of "reference identity" has been utilized before [17]. The reference identities are pre-selected independent of the test faces. There is no specific restriction on the choices of the reference identities, but efforts are made to ensure that the set contains sufficient variations of gender, race, and various face attributes, to allow for an unbiased face representation.

3.2 Face Verification Framework

In our experiments, we extend the representation in Sec. 3.1 to multiple local facial parts to make the reference-based representation robust to variations of pose, illumination, expression, etc. We detect M fiducial points for each face, representing points-of-interest such as eye corners, nose tip, and mouth corners. The procedure in Sec. 3.1 is repeated for each fiducial point to obtain M histograms for each face. The proposed supervised EMD framework (Sec. 2) is used to learn a ground distance matrix for each of the M fiducial points, using similar and dissimilar pairs from the training data set. In the testing phase, the supervised EMD values between two faces are calculated for each of the M fiducial points, and the M-dimensional distance vector is fed into a pre-trained SVM to decide whether or not the two faces belong to the same person.

4 EMD Flow for Face Attribute Analysis

The scalar value of the supervised EMD gives the dissimilarity between two histograms, but much richer information regarding how the two histograms differ from each other is contained in the flow-network, indicating the optimal transformation between the histograms across all bins. For example, several faces might have exactly the same EMD to an anchor face, but the information about how they differ from the anchor image is contained in the flow-network. Note that the flow-network is a natural byproduct of the EMD calculation, without requiring any extra computation. In this section, we utilize the flow-network to analyze face attribute changes within a set of faces of a same person.

If we define a *sequence* as a re-ordering of $\{(i, j) \mid i, j = 1, 2, \ldots, n\}$, a *Monge sequence* is then defined as a sequence in which for every (i, j) that precedes (i, s) and (r, j), the ground distance matrix D satisfies $d_{ij} + d_{rs} \leq d_{is} + d_{rj}$.

If a Monge sequence exists, a greedy algorithm based on the Monge sequence will yield the optimal solution of EMD [28]. If a full-length Monge sequence doesn't exist, the longest subsequence satisfying the Monge condition is called a Monge subsequence. The entries in the flow matrix corresponding to the elements in the Monge sequence or subsequence are partially determined by the ground distance, thus are insensitive to the actual histograms being compared.

After finding the longest Monge subsequence given the ground distance matrix [28], we decompose the flow matrices into two parts: one containing the flows that result from the Monge subsequence, which is denoted as the *Monge flow*, the other containing the remaining entries in the flow matrices that cannot be solved using the greedy algorithm, denoted as the *non-Monge flow*. To measure distance between two flow-networks, we choose L_2 distance between the non-Monge components of the two flows, because these are the components of the flows that are highly dependent on the histograms involved.

Now that we know how to evaluate similarity between two flows, given two faces A and B where B is the nearest neighbor of A in terms of EMD, we can find a face C that differs from A in the same way as B does. This is specified by requiring the distance between the non-Monge components of the two flows

Flow($A \to B$) and Flow($A \to C$) to be smaller than a certain threshold. We connect the link $A \to B \to C$ to indicate this relationship, and B precedes C because $\mathrm{EMD}(A, B) < \mathrm{EMD}(A, C)$. If this procedure is performed repetitively, a longer path $A \to B \to C \to D \to \cdots$ can be discovered. Since the flows in the path are consistent, we often observe that the path indicates variation from face A along a certain interpretable attribute.

If we repeat the procedure above by setting B to all close neighbors of A, we can find multiple paths to transit A to other faces, each corresponding to a transition along a specific facial attribute. Please see Sec. 5.2 and Fig. 6a for more detailed examples.

5 Experimental Results

5.1 Face Verification on Standard Face Data Sets

We evaluate the proposed algorithm on two standard data sets: *Labeled Faces in the Wild (LFW)* [29] and *PubFig* [17]. These two data sets both contain real-world face images of public figures with a large degree of variations in pose, age, expression, race, illumination, etc.

LFW contains 13,233 face images from news photos. We follow the exact restricted setting as specified in the original paper [29]. The performance is measured through 10-fold cross validation on 6,000 face pairs. Please refer to the LFW paper [29] for more detailed description of the experimental setup.

PubFig contains image URLs collected from the Internet by a face detector. It includes 58,797 images of 200 public figures or celebrities. We downloaded all 50,948 images that were still available online when the experiment was performed (Sep. 2011). The provided face bounding box for each image was also verified by the OpenCV face detector, because some images pointed to by the provided URLs have been resized or even completely changed. Face bounding boxes with severe discrepancies were filtered out. The final data set contains 45,068 face images. The *PubFig* data set was divided into 2 non-overlapping parts: development set and evaluation set, containing images of 60 and 140 individuals, respectively. Among the images we've successfully downloaded and verified, these two sets include 12,603 and 32,465 images, respectively.

Improved Ground Distance Matrix by Learning. In this experiment, we use holistic representation of each face to generate histogram descriptor, i.e., each face is resized to 64×64 and reduced to 500 dimensions by PCA. The reference set were selected as the development set of *PubFig* with 60 identities.

The ground distance should be the distance between identities, which is difficult to define. We would like to investigate how the supervised EMD affects the ground distance matrix. The initial ground distance between the reference identities is defined as the Hausdorff distance between the PCA features of the two corresponding groups of face images. The training data from the face pairs in *PubFig* are used to learn a new ground distance using supervised EMD.

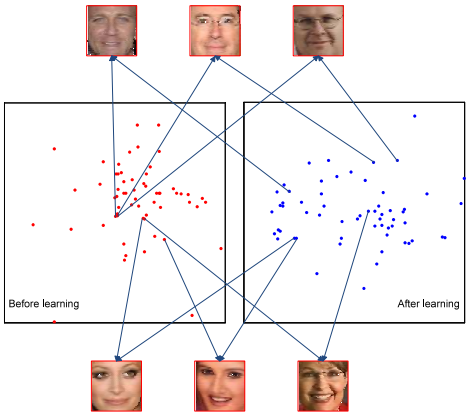


Fig. 2. Ground distance matrix before and after learning, visualized by MDS. The three identities shown on the top were close to each other before learning, but are separated after learning because they do not look alike. Moreover, the first face was moved further away from the other two, because it's more different from the other two, most likely because of the absence of eye glasses. For the three identities at the bottom, the first and the third persons were close to each other before learning, while the second person was some distance away. After learning, the first and the second person was pushed closer because they actually look alike, while the third person was placed further away. Note that only one representative face image is shown for each of the six identities.

We visualize the ground distance matrices before and after training in a 2D space by Multidimensional Scaling (MDS) in Fig. 2, with each dot representing an identity. The figure shows that the ground distance after learning (right, blue dots) makes more sense compared to that before learning (left, red dots). Please refer to the caption of Fig. 2 for more details. Using this improved ground distance, it's reasonable to expect the resulting EMD to improve over the original EMD based on the ad hoc ground distance. We will demonstrate this experimentally in the next section.

Face Verification on Real-World Data Sets. We evaluate the performance of face verification on the two real-world data sets, still using the 60 identities from *PubFig* as reference identities. Local face features are used for robustness as described in Sec. 3.2. Seven fiducial points are detected for each face [30], and the local facial patches at 3 different scales (4×4, 8×8, and 12×12) are extracted as the local representation. Two faces are compared by $7 \times 3 = 21$ local descriptors plus the holistic descriptor, yielding 22 EMD values. These values are then fed in an SVM classifier to determine whether these two faces belong to the same person. During cross-validation, the training set for EMD learning and for learning the SVM classifier are both changed correspondingly in each fold. The test set is excluded from all types of training.

The face verification performance is evaluated by the average ROC curve over 10-fold cross validation on *PubFig* and *LFW*. The performance of our method

before and after EMD training are first compared with Attribute classifiers [17] for *PubFig* data in Fig. 3. Our method is then evaluated on *LFW* In Fig. 5 comparing with the following methods: traditional eigenface [31] utilizes the least information and shows the worst performance; CSML+SVM [16], DML-eig [32] and Hybrid aligned [33] included outside data for alignment or feature extraction, so does our method; Attribute and Similie [17] and Multiple LE [34] had similar framework to ours, although they included more outside data for training. The average accuracy for *LFW* is listed in Fig. 4. Even though only simple histogram-like descriptors are used, the proposed supervised EMD framework still achieves state-of-the-art performance on the two data sets.

The proposed algorithm addresses the application of face verification from a unique angle, without special tuning of parameters or using handcrafted features. Although the results are only slightly better than the state-of-the-art, the tools used here are completely different, thus suggesting opportunities for combined approaches that may perform even better.

5.2 Face Attribute Transitions Using EMD

Using the same setup as in Sec. 5.1, we find paths in face data sets that have consistent attribute changes using the strategies in Sec. 4. Several paths within faces of one identity are shown in Fig. 6a, each indicated by a blue surrounding box. All paths found by analyzing the EMD flows reflect some transition on certain facial attribute. Please see the figure caption for detailed explanations. The transition paths reveal structures present in the underlying manifold of the face space.

Since a flow can represent the changes of face attributes between face pairs, it can be used to find new face pairs that show similar attribute change. For example, in Fig. 6b, given a pair of example faces in the blue box showing expression changing from smiling to neutral, we try to transfer the same change to the smiling faces in the red boxes. To do this, we impose the EMD flow between the given pair to the descriptor of each smiling face in the red boxes. The resulting new descriptor is the used to retrieve a face image whose descriptor is the closest to the new descriptor. The results are shown in the second column.

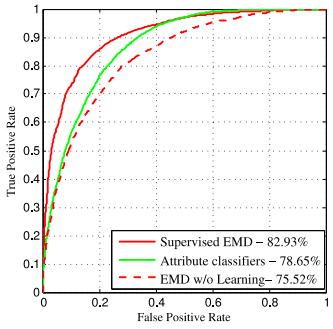
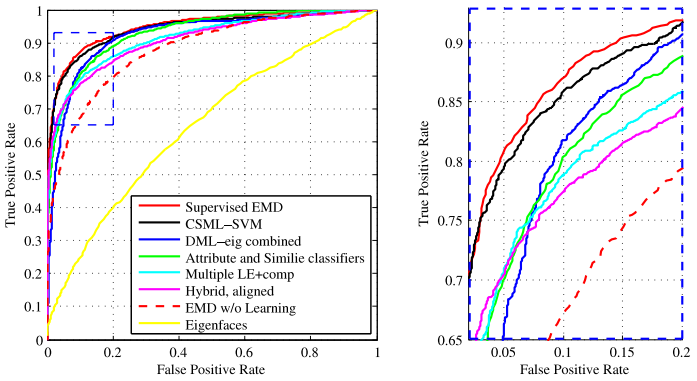


Fig. 3. ROC curve on *PubFig*

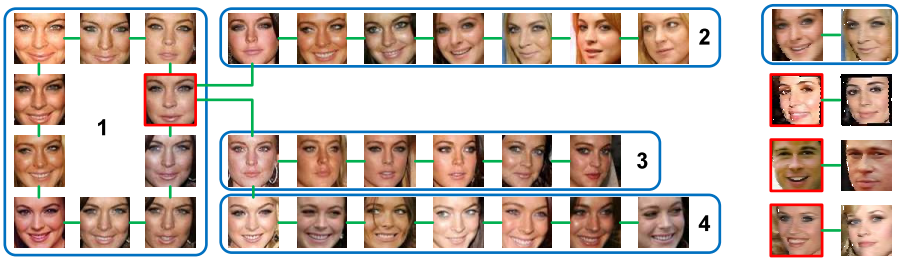
Method	Accuracy ± Std
Supervised EMD	**0.8853 ± 0.0107**
CSML+SVM	0.8800 ± 0.0037
DML-eig combined	0.8565 ± 0.0056
Attribute and Similie classifiers	0.8529 ± 0.0123
Multiple LE+comp	0.8445± 0.0046
Hybrid, aligned	0.8398 ± 0.0035
EMD w/o Learning	0.7977 ± 0.0121
Eigenfaces	0.6002 ± 0.0079

Fig. 4. Performance comparison on *LFW*



(a) ROC curve (b) ROC curve amplified

Fig. 5. ROC curve of face verification on *LFW*. The blue box in (a) is amplified in (b).



(a) (b)

Fig. 6. (a) Attribute paths: The faces of the same identity are organized in several paths corresponding to certain consistent attribute transitions. The face in the red box is selected as the starting point, from which different paths are found to explore the face space. Path 1 contains two branches of faces with gradually changing expressions (both neutral → smile), and the two branches meet each other after some steps to form a circle. Faces in Path 2 show a combination of pose and expression changes. Faces in Path 3 and 4 both show pose changes, but faces in Path 3 all have neutral expression while those in Path 4 are smiling. **(b) Attribute transfer:** Given a pair of faces (in blue box) changing from smiling to neutral, its flow is imposed on other smiling faces (red), and the corresponding neutral faces are found as in the second column.

By applying the flow representing the change of "smiling→neutral", we have successfully transformed other smiling faces to neutral ones. Note that before transferring, the given flow needs to be normalized to have uniform L_1 norm on each row to ensure relative independence to the source histogram.

6 Discussion and Conclusion

In this paper, we have presented an algorithm that jointly learns a ground distance matrix and a flow-network for the Earth Mover's Distance. Learning a

better EMD is a fundamental problem in measuring distance between distributions or histograms. The effectiveness of the optimized EMD distance is demonstrated by face verification results on two standard data sets. The proposed EMD learning framework can also be directly applied to measuring distances for other histogram-like features such as color histogram, histogram of gradient (HoG), SIFT, Bag-of-Words, etc.

EMD yields not only a distance value but also a transformation plan. We demonstrate that the flow-network contains valuable information about the "direction" of facial attribute transition, which cannot be easily achieved by classic distance metrics. The "directions" of the flows in image space can be further applied in many other contexts, such as image understanding and retrieval.

Acknowledgement. The authors would like to acknowledge NSF grants CNS 0832820 and IIS 1016324, a research grant from Google, Inc., and the valuable comments and suggestions from anonymous reviews and area chairs.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
2. Belongie, S., Malik, J., Puzicha, J.: Shape Context: a new descriptor for shape matching and object recognition. In: NIPS (2001)
3. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
4. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
5. Xu, D., Yan, S., Luo, J.: Face recognition using spatially constrained Earth Mover's Distance. IEEE Transactions on Image Processing 17(11), 2256–2260 (2008)
6. Zhao, Q., Yang, Z., Tao, H.: Differential Earth Mover's Distance with its applications to visual tracking. IEEE TPAMI 32, 274–287 (2008)
7. Grauman, K., Darrell, T.: Fast contour matching using approximate Earth Mover's Distance. In: CVPR, pp. 220–227 (2004)
8. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: NIPS (2003)
9. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: ICML (2003)
10. Frome, A., Singer, Y., Malik, J.: Image retrieval and classification using local distance functions. In: NIPS (2006)
11. Domeniconi, C., Gunopulos, D.: Adaptive nearest neighbor classification using Support Vector Machines. In: NIPS (2002)
12. Cuturi, M., Avis, D.: Ground Metric Learning. arXiv:1110.2306v1 (2011)
13. Wang, X.L., Liu, Y., Zha, H.: Learning robust cross-bin similarities for the bag-of-features model. Technical report, Key Labs of Machine Perception, Peking University, China (2009)
14. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR, pp. 539–546 (2005)
15. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: ICCV, pp. 498–505 (2009)

16. Nguyen, H.V., Bai, L.: Cosine Similarity Metric Learning for Face Verification. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 709–720. Springer, Heidelberg (2011)
17. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
18. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (2011)
19. Levina, E., Bickel, P.: The Earth Mover's Distance is the Mallows distance: some insights from statistics. In: ICCV, pp. 251–256. IEEE Computer Society (2001)
20. Villani, C.: Topics in Optimal Transportation. American Mathematical Society (2003)
21. Villani, C.: Optimal transport, Old and New. Grundlehren der Mathematischen Wissenschaften, vol. 338. Springer (2009)
22. Pele, O., Werman, M.: Fast and robust Earth Mover's Distances. In: ICCV, pp. 460–467 (2009)
23. Pele, O., Werman, M.: A Linear Time Histogram Metric for Improved SIFT Matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 495–508. Springer, Heidelberg (2008)
24. Andoni, A., Ba, K.D., Indyk, P., Woodruff, D.: Efficient sketches for Earth-Mover Distance, with applications. In: IEEE FOCS, pp. 324–330 (October 2009)
25. Chvátal, V.: Linear Programming. W. H. Freeman (1983)
26. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 1.21 (April 2011), `http://cvxr.com/cvx`
27. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: ICCV (2011)
28. Alon, N., Cosares, S., Hochbaum, D.S., Shamir, R.: An algorithm for the detection and construction of Monge sequences. Linear Algebra and Its Application, 669–680 (1989)
29. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)
30. Sivic, J., Everingham, M., Zisserman, A.: "who are you?" - learning person specific classifiers from video. In: CVPR (2009)
31. Turk, M.A., Pentland, A.P.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
32. Ying, Y., Li, P.: Distance metric learning with eigenvalue optimization. JMLR (Special Topics on Kernel and Metric Learning) 13, 1–26 (2012)
33. Taigman, Y., Wolf, L., Hassner, T.: Multiple one-shots for utilizing class label information. In: BMVC (2009)
34. Cao, Z., Yin, Q., Tang, X., Sun, J.: Face recognition with learning-based descriptor. In: CVPR (2010)