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

FAN WANG[1], LEONIDAS J. GUIBAS[2]
Geometric Computing Group, Stanford University
[1] fanw@stanford.edu, [2] guibas@stanford.edu

ABSTRACT

Earth Mover's Distance (EMD) is an intuitive and natural distance metric for comparing two histograms or probability distributions. We propose to jointly optimize the ground distance matrix and the EMD flow-network based on partial ordering of histogram distances in an optimization framework. Two applications in computer vision are used to demonstrate the effectiveness of the algorithm: firstly, we apply the optimized EMD value to face verification, and achieve state-of-the-art performance on public face data sets; secondly, we use the learned EMD flow-network to analyze the internal structure of a set of faces, and consistent paths that demonstrate intuitive transitions on certain facial attributes are found.

1 Supervised Earth Mover's Distance

1.1 Introduction of EMD

Input:
Source histogram $\mathbf{p} \in \mathbb{R}^n$: regarded as piles of earth at various locations (bins). The amount of earth in each pile equals to the value of each corresponding bin.
Destination histogram $\mathbf{q} \in \mathbb{R}^n$: regarded as several holes, the values of which represent their capacities.
Ground distance matrix $D = \{d_{ij}\}$: d_{ij} defines the cost of moving a unit of earth from the i-th bin of $\mathbf{p}$ to the j-th bin of $\mathbf{q}$.
Output:
Flow matrix $F = \{f_{ij}\}$: f_{ij} denotes the amount of earth moved from the i-th bin of $\mathbf{p}$ to the j-th bin of $\mathbf{q}$.
Formulation:

$$\mathrm{EMD}(\mathbf{p},\mathbf{q}) = \min_{\mathbf{f}\in\mathbb{C}(\mathbf{p},\mathbf{q})} \mathbf{d}^T\mathbf{f}, \tag{1}$$

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

1.2 Supervised Earth Mover's Distance with Triplets

Training Data: N triplets of histograms $\{(\mathbf{p}_i,\mathbf{q}_i,\mathbf{r}_i), i = 1,\cdots,N\}$. For each triplet,$\mathrm{EMD}(\mathbf{p}_i,\mathbf{r}_i) \geq \mathrm{EMD}(\mathbf{p}_i,\mathbf{q}_i)$, $\forall i$.
Objective: an optimal ground distance $\mathbf{d}$ so that the resulting EMD values satisfy as many constraints as possible.
Formulation:
Turn each hard constraint into a large-margin constraint:

$$\mathrm{EMD}(\mathbf{p}_i,\mathbf{r}_i) - \mathrm{EMD}(\mathbf{p}_i,\mathbf{q}_i) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i, \tag{2}$$

The overall problem is formulated as:

$$\begin{aligned} \min.\quad & \|\mathbf{d}\|_2^2 + C\cdot\boldsymbol{\xi}^T\mathbf{1} \\ \text{s.t.}\quad & \mathbf{d}^T(M_g - M_f) \geq \mathbf{1}^T - \boldsymbol{\xi}^T,\ \boldsymbol{\xi} \geq 0 \\ & \mathbf{d} \in \mathbb{D} \\ & \mathbf{f}_i = \underset{\mathbf{f}\in\mathbb{C}(\mathbf{p}_i,\mathbf{q}_i)}{\mathrm{argmin}}\ \mathbf{d}^T\mathbf{f},\ \mathbf{g}_i = \underset{\mathbf{g}\in\mathbb{C}(\mathbf{q}_i,\mathbf{r}_i)}{\mathrm{argmin}}\ \mathbf{d}^T\mathbf{g},\ \forall i, \end{aligned} \tag{3}$$

The convex feasible domain for the ground distance is

$$\mathbb{D} = \left\{\mathbf{d} \mid \mathbf{d} = \mathrm{vec}(D), D \in \mathbb{R}^{n\times n},\ D_{ij} \geq 0, D_{ii} = 0, \forall i,j\right\}. \tag{4}$$

1.3 Supervised Earth Mover's Distance In a More General Setting

Training Data: sets of similar pairs $\{(\mathbf{p}_i,\mathbf{q}_i), i = 1,\cdots,N_s\}$ and dissimilar pairs $\{(\mathbf{r}_j,\mathbf{s}_j), j = 1,\cdots,N_d\}$.
Objective: The two sets of distances, $\{\mathrm{EMD}(\mathbf{p}_i,\mathbf{q}_i), i = 1,\cdots,N_s\}$ and $\{\mathrm{EMD}(\mathbf{r}_j,\mathbf{s}_j), j = 1,\cdots,N_d\}$, are separated as much as possible.
Formulation:
The problem is naturally transformed into a max-margin problem:

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

1.4 Solving for Optimal Ground Distance

The formulated problem is bi-convex with respect to the two sets of variables $\{\mathbf{d}\}$ and $\{M_f, M_g\}$:
Given the ground distance $\mathbf{d}$: it can be naturally decoupled into independent standard EMD problems;
Given the flows M_f and M_g, the optimization problem in Eq. 3 can be re-written as:

$$\begin{aligned} \min.\quad & \|\mathbf{d}\|_2^2 + C\cdot\boldsymbol{\xi}^T\mathbf{1} \\ \text{s.t.}\quad & \mathbf{d}^T(M_g - M_f) \geq \mathbf{1}^T - \boldsymbol{\xi}^T,\ \boldsymbol{\xi} \geq 0 \\ & \mathbf{d} \in \mathbb{D}, \end{aligned} \tag{6}$$

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

Given the flows $\mathbf{f}_i$ and $\mathbf{g}_i$, Eq. 5 can also be rewritten as a QP:

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

Finally, the supervised EMD learning problem is solved using alternating optimization:

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. 6 or Eq. 7;
  $\mathbf{d}^k \leftarrow \mathbf{d}^{k-1} + \alpha(\mathbf{d}^k - \mathbf{d}^{k-1})$;
  $k \leftarrow k + 1$;
end while

2 Face Verification Using Supervised EMD

Reference Identities: a set of known identities, each represented by a set of diverse face images of one person.

- The faces of the i-th reference person: matrix X_i.
- The total K identities: $X = [X_1, X_2, \cdots, X_K]$.

A test face $\mathbf{y}$ is reconstructed by X with a L_2 regularization term:

$$\min.\quad \|\mathbf{y} - X\boldsymbol{\alpha}\|_2 + \lambda\|\boldsymbol{\alpha}\|_p, \tag{8}$$

Then the test face is: $\mathbf{y} \approx X\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1;\boldsymbol{\alpha}_2;\ldots;\boldsymbol{\alpha}_K]$.
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:

$$e_i(\mathbf{y}) = \|\mathbf{y} - X_i\boldsymbol{\alpha}_i\|_2. \tag{9}$$

The vector of the reconstruction errors $\mathbf{e}(\mathbf{y}) = [e_1(\mathbf{y}), e_2(\mathbf{y}), \ldots, e_K(\mathbf{y})]$ is transformed to a similarity score vector $\mathbf{s}(\mathbf{y})$:

$$s_i(\mathbf{y}) = \exp(-\frac{1}{2\sigma_i^2}(e_i(\mathbf{y}) - \mu_i)^2), \tag{10}$$

and normalized to have L_1 norm equal to unity, giving our final histogram-like face descriptor as illustrated in the following figure.

The representation is further expanded to multiple local facial parts to make it robust to variations of pose, illumination, expression, etc.

3 EMD Flow for Face Attribute Analysis

Motivation: 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.
Objective: The flow-network is used to analyze face attribute changes within a set of faces of a same person.

Definition 1 *A sequence is a re-ordering of* $\{(i,j) \mid i = 1,2,\ldots,n, j = 1,2,\ldots,n\}$.

Definition 2 *A* Monge *sequence is 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, then a greedy algorithm based on the Monge sequence yields the optimal solution of EMD.
Each flow matrix is decomposed into two parts:

- *Monge flow*: the flows that result from the Monge subsequence.
- *non-Monge flow*: the remaining entries in the flow matrices that cannot be solved using the greedy algorithm.

The distance between two flow-networks is measured by L_2 distance between the *non-Monge* components of the two flows.
Face $A \to C$ differs in the same way as $A \to B$ if the distance between the two flows Flow$(A \to B)$ and Flow$(A \to C)$ is small enough in the way defined above.
A longer path $A \to B \to C \to D \to \cdots$ can be formed by finding more faces along the path repetitively.

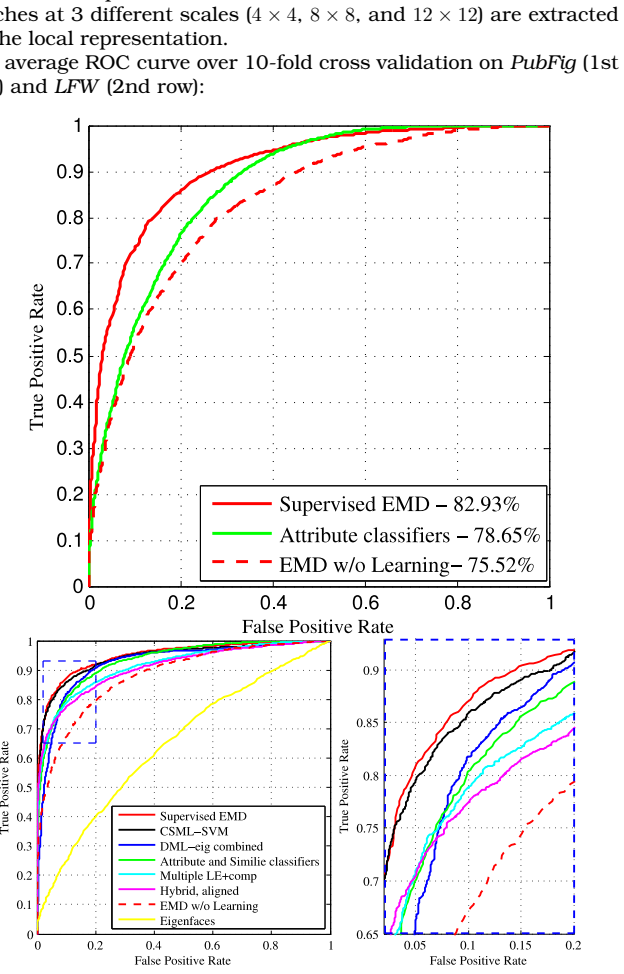
4 Experimental Results

4.1 Face Verification on Standard Face Data Sets

Data sets:

- *Labeled Faces in the Wild (LFW)* (13,233 face images from news photos).
- *PubFig* (58,797 images of 200 public figures or celebrities).

Reference identities: 60 identities of development set from *PubFig*. Seven fiducial points are detected for each face, and the local facial patches at 3 different scales (4×4, 8×8, and 12×12) are extracted as the local representation.
The average ROC curve over 10-fold cross validation on *PubFig* (1st row) and *LFW* (2nd row):



The average accuracy for *LFW*:.

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

4.2 Face Attribute Transition Using EMD

Left: Several paths within face collection of one identity, each reflecting some transition on certain facial attribute.
Right: given a pair of example faces showing expression change from smiling to neutral, the same transition is "transferred" to some other smiling faces, transforming other smiling faces to neutral expression.