Simultaneous Ground Metric Learning and Matrix Factorization with Earth Mover's Distance

Gloria Zen*, Elisa Ricci†‡, Nicu Sebe*
* University of Trento, Trento, Italy
Email: {zen,sebe}@disi.unitn.it
†Fondazione Bruno Kessler, Trento, Italy, ‡University of Perugia, Perugia, Italy
Email: {elricci}@fbk.eu

***Abstract*—Non-negative matrix factorization is widely used in pattern recognition as it has been proved to be an effective method for dimensionality reduction and clustering. We propose a novel approach for matrix factorization which is based on Earth Mover's Distance (EMD) as a measure of reconstruction error. Differently from previous works on EMD matrix decomposition, we consider a semi-supervised learning setting and we also propose to learn the ground distance parameters. While few previous works have addressed the problem of ground distance computation, these methods do not learn simultaneously the optimal metric and the reconstruction matrices. We demonstrate the effectiveness of the proposed approach both on synthetic data experiments and on a real world scenario, *i.e.* addressing the problem of complex video scene analysis in the context of video surveillance applications. Our experiments show that our method allows not only to achieve state-of-the-art performance on video segmentation, but also to learn the relationship among elementary activities which characterize the high level events in the video scene.**

I. Introduction

Non-negative Matrix Factorization (NMF) [1] is a popular tool in pattern recognition, machine learning and computer vision. NMF aims to decompose a matrix finding two non-negative matrices, usually of small size, whose product approximates the original matrix. In practice NMF computes a compressed version of the initial data matrix. Thus, it has been widely used for dimensionality reduction and clustering. While the majority of NMF approaches adopt the L_2 or the Kullback-Leibler distance to measure the reconstruction error, recently a NMF method based on Earth Mover's Distance has been introduced [2]. Sandler *et al.* demonstrate that the EMD matrix decomposition must be preferred when the error mechanism is not modeled well by additive noise but is rather a complex local deformation of the original signal.

A critical aspect, when employing EMD matrix factorization or simply when computing EMD, is represented by the definition of the ground distance parameters. While this is typically done using some *a priori* knowledge, recently some works [3], [4] have demonstrated that learning the ground distance values is greatly beneficial in several applications.

In this paper we follow these recent works and propose a novel approach for weakly supervised EMD-NMF where the matrix decomposition with minimal reconstruction error is computed together with the optimal ground distance parameters solving a simple optimization problem. We further derive an alternate optimization approach to compute its solution efficiently and we show that it reduces to a sequence of convex optimization programs. We demonstrated the effectiveness of the proposed approach on synthetic data experiments and in the context of video scene analysis. Recent works [5], [6] have shown that EMD-NMF methods can be successfully employed to extract automatically complex behaviors from crowded scenes, *e.g.* those depicting public spaces recorded from video surveillance cameras. However, the ground distance values are set according to some heuristics [5] or learnt in advance before clustering is performed [6]. It is intuitive that, assuming to have at disposal some side information, learning them from data while finding the clustering solution is beneficial. In this paper we follow this intuition.

A. Contributions

Up to our knowledge, this is the first paper presenting an approach for weakly-supervised EMD-NMF. The proposed optimization scheme is also new and while previous works have considered EMD-NMF for discovering typical patterns in dynamic scenes, none of them learns the ground distance parameters in a discriminative fashion. Finally, our EMD-NMF algorithm is general and can be used in other applications not limited to the computer vision field, such as data mining and clustering.

II. Related Works

The Non-negative Matrix Factorization algorithm aims to find two non-negative matrices whose product provides a good approximation to an initial matrix. While originally proposed to learn the parts of objects like human faces and text documents [1], in the last decades it has been applied to many other problems, such as action recognition [7], speech denoising [8], analysis of electromyographic signals [9] and blind source separation [10]. Typically NMF approaches adopt a bin-to-bin measure (*e.g.* L_2, Kullback Leibler divergence) to compute the reconstruction error. While this usually implies a simple optimization algorithm, the situations where the original matrix can only be obtained from complex deformations of some elementary signals cannot be modeled. To cope with this, the EMD-NMF algorithm is introduced in [2], where a cross-bin measure, *i.e.* the Earth Mover's Distance, is adopted instead of bin-to-bin distances. Improved performance with respect to traditional NMF are shown in two computer vision tasks, *i.e.* texture descriptor estimation and face recognition. However, in [2] the ground distance values are kept fixed and are not

optimized according to a discriminative criterion as proposed in this paper.

Semi-supervised approaches to NMF have been proposed in [11], [12]. Simultaneous clustering and metric learning has been introduced in [13]. However, up to our knowledge, no previous works have considered these problems in the context of Earth Mover's Distance factorization.

How learning the ground distance parameters affect EMD computation has been investigated in [3], [4]. In [3] an algorithm that learns the ground metric values using a training set of labeled histograms is proposed, overcoming the traditional approach that sets them based on a priori knowledge of the features. Wang *et al.* [4] also uses side information from triplets of samples (*i.e. must link* or *cannot link* constraints) to learn the cross-bin relationships, hence producing more accurate EMD values. However, in [3], [4] the ground metric parameters are learnt in order to simply compute the EMD and not in the context of matrix factorization.

Earth Mover's Distance has been originally proposed in [14] as an effective measure for histogram comparison. While EMD has been used in many applications such as image retrieval or face recognition, recently some approaches for EMD clustering have been introduced. In [15] a two-class clustering problem is formulated as an integer convex optimization problem. In [6] a clustering approach based on a simplified version of EMD is proposed, resulting into a simple linear programming problem. In [5] a EMD-NMF algorithm with sparsity constraints is introduced. However, in all these works the ground distance parameters used in the EMD transportation problems are set a priori without optimizing their values for improved clustering accuracy.

In order to overcome the main limitations of EMD, which are computational complexity and scalability, efficient versions of EMD have been proposed [16], [17], [18]. In some of them [16], the specific situation where the ground distance among histograms' bins is a linear function of the bin position is considered (*e.g.* $d_{ij} = |i-j|$ in EMD-L_1). In applications where different bin positions correspond to sorted elements in space [5] or in time [19] using a linear distance is a natural solution. Conversely, in situations where a histogram's bin corresponds to a word in a specific vocabulary (*e.g.* when the bag of words paradigm is employed), the use of EMD-L_1 implies finding a reasonable words' order, according to which similar words are assigned to neighboring bin positions. The best sorting is typically the one which minimizes some distortion measure over a previously defined ground distance among words. Approaches for sorting have been proposed in literature [6], [20]. However they usually lead to a suboptimal solution, being the problem NP hard. Furthermore, as the initial distances are assigned based on L_1 or L_2 metrics, this may not necessarily reflect the discriminative ability of words. Therefore, by learning ground distances as proposed in this paper we overcome these issues at the expenses of a slightly increased computational cost in EMD calculation.

III. EMD-NMF and Ground Metric Learning

A. Earth Mover's Distance

Given two normalized histograms $\boldsymbol{h}_i, \boldsymbol{h}_j \in I\!R^M$ ($\sum_t h_i^t = \sum_q h_j^q = 1$), the Earth Mover's Distance $\mathcal{D}(\boldsymbol{h}_i, \boldsymbol{h}_j)$ [14] is defined as:

$$\begin{aligned} \mathcal{D}(\boldsymbol{h}_i, \boldsymbol{h}_j) \quad &= \min_{f_{qt} \geq 0} \sum_{t,q=1}^{M} d_{qt} f_{qt} \\ &\text{s.t.} \quad \sum_{q=1}^{M} f_{qt} = h_i^t, \ \forall t \qquad \sum_{t=1}^{M} f_{qt} = h_j^q, \ \forall q \end{aligned} \tag{1}$$

The problem (1) is a transportation problem and the flow variables f_{qt} denotes the amount transported from the q-th supply to the t-th demand. The parameter d_{qt} represents the ground distance between bins q and t. Usually d_{qt} is defined by L_1 or L_2 distance or is determined based on some *a priori* knowledge of the features in the considered application. The problem is a Linear Program (LP) which can be solved efficiently due to the special structure of its sparse constraints [14], [16], [18].

B. EMD-NMF with Ground Metric Learning

We are given a training set $\mathcal{H} = \{\boldsymbol{h}_i\}_{i=1}^N, \boldsymbol{h}_i \in I\!R^M$ of normalized histograms and a small set $\mathcal{H}_s = \{(\boldsymbol{h}_i^s, \boldsymbol{h}_j^s, y_{ij}^s)\}_{i,j=1}^{N_s}$, of pairs of histograms $\boldsymbol{h}_i^s, \boldsymbol{h}_j^s \in I\!R^M$ and associated label $y_{ij}^s \in \{1, -1\}$ indicating if the histograms belong to the same or to a different class. From the set $\mathcal{H}$ we construct the matrix $\mathbf{H} = [\boldsymbol{h}_1 \ \boldsymbol{h}_2 \dots \boldsymbol{h}_N]$, $\mathbf{H} \in I\!R^{M \times N}$. We are interested in decomposing $\mathbf{H}$ finding a set of basis $\mathcal{P} = \{\boldsymbol{p}^1, \ \boldsymbol{p}^2, \ \dots \boldsymbol{p}^K\}$, with $K << N$, $\boldsymbol{p}^k \in I\!R^M$ and a matrix of mixing coefficients $\mathbf{W} = [\boldsymbol{w}_1 \ \boldsymbol{w}_2 \dots \boldsymbol{w}_N]$, $\boldsymbol{w}_i \in I\!R^K$, such that the weighted sum of the computed basis should be as close as possible to the original histograms according to Earth Mover's Distance. We also want to find the optimal ground distance parameters $\boldsymbol{d} \in I\!R^{M \times M}$ imposing that histograms of different classes $(\boldsymbol{h}_i^s, \boldsymbol{h}_j^s)$ in $\mathcal{H}_s$ should be more distant than histograms of the same class $(\boldsymbol{h}_l^s, \boldsymbol{h}_m^s)$. The following optimization problem is formulated:

$$\begin{aligned} \min_{\boldsymbol{p}^k, \mathbf{W}, \boldsymbol{d}} \quad & \|\boldsymbol{d}\|_F^2 + \lambda_1 \sum_{i=1}^{N} \mathcal{D}_{\boldsymbol{d}}(\boldsymbol{h}_i, \sum_{k=1}^{K} w_i^k \boldsymbol{p}^k) + \lambda_2 \sum_{ijlm} \xi_{ijlm} \\ & \mathcal{D}_{\boldsymbol{d}}(\boldsymbol{h}_i^s, \boldsymbol{h}_j^s) - \mathcal{D}_{\boldsymbol{d}}(\boldsymbol{h}_l^s, \boldsymbol{h}_m^s) \geq 1 - \xi_{ijlm} \quad \forall i,j,l,m \\ & \boldsymbol{p}^k \in \mathcal{F}, \mathbf{W} \geq 0, \boldsymbol{d} \in \mathcal{D} \end{aligned} \tag{2}$$

where $\mathcal{F} = \{\boldsymbol{p}^k \in I\!R^M : \sum_q p_q^k = 1, \ p_q^k \geq 0\}$ and $\mathcal{D} = \{\boldsymbol{d} \in I\!R^{M \times M} : d_{qt} \geq 0, \ d_{qt} = d_{tq}, \ d_{qq} = 0\}$. In practice the feasible set $\mathcal{F}$ indicates that the set of the chosen basis vectors should be histograms normalized to unit mass, while the set $\mathcal{D}$ indicates that the matrix of ground distance parameters should be symmetric and with all the elements equal or greater than zeros with the exception of the elements on the diagonal which are equal to zero. This choice corresponds to defining a valid transportation problem in the form (1).

C. Optimization

The optimization problem (2) is non convex. To solve it we adopt an alternate optimization approach and solve separately for $\boldsymbol{p}^k, \mathbf{W}, \boldsymbol{d}$ considering a sequence of convex optimization problems. In practice, at every step the optimal values of the flow vectors in the EMD definition must also be computed. In the following we describe the proposed optimization algorithm.

Initialization. Given $\mathcal{H}$, $\mathcal{H}_s$ initialize $\mathbf{W}$, $\boldsymbol{d}$. The initialization of $\mathbf{W}$ can be done considering a traditional NMF

algorithm [1] modified to handle the required normalizations. The values of the ground distance parameters $\boldsymbol{d}$ are initialized assigning $d_{qt} = 1$ if $q \neq t$, $d_{qt} = 0$ otherwise.

Step 1. Given $\mathcal{H}_s$ and $\boldsymbol{d}$ the several independent optimization problems associated to distance constraints can be solved finding the optimal values of the flow variables vectors $\mathbf{g}^{ij}$, $\mathbf{g}^{lm}$:

$$\mathbf{g}^{ij} = \arg\min_{\mathbf{g}} \mathcal{D}_{\boldsymbol{d}}(\boldsymbol{h}_i^s, \boldsymbol{h}_j^s) \quad \forall i, j \tag{3}$$
$$\mathbf{g}^{lm} = \arg\min_{\mathbf{g}} \mathcal{D}_{\boldsymbol{d}}(\boldsymbol{h}_l^s, \boldsymbol{h}_m^s) \quad \forall l, m$$

where $\boldsymbol{h}_i^s, \boldsymbol{h}_j^s$ are histograms corresponding to the same class while $\boldsymbol{h}_l^s, \boldsymbol{h}_m^s$ are associated to different classes.

Step 2. Given $\boldsymbol{d}$, $\mathbf{W}$ fixed, find $\boldsymbol{p}^k$ and the flow variables $\boldsymbol{f}$. The optimization problem which must be solved is formulated as:

$$\begin{aligned} \min_{p_q^k, f_{q,t}^i \geq 0} \quad & \sum_{i=1}^{N}\sum_{q=1}^{M}\sum_{t=1}^{M} d_{qt} f_{qt}^i \\ \text{s.t.} \quad & \sum_{q=1}^{M} f_{qt}^i = h_i^t, \quad \forall i, \ \forall t \\ & \sum_{t=1}^{M} f_{tq}^i = \sum_{k=1}^{K} w_i^k p_q^k \quad \forall i, \ \forall q \\ & \sum_{q=1}^{M} p_q^k = 1 \ \ \forall k \end{aligned} \tag{4}$$

This is a simple LP which can be solved efficiently with standard solvers.

Step 3. Given $\boldsymbol{d}$, $\boldsymbol{p}^k$ fixed, find $\mathbf{W}$ and the flow variables $\boldsymbol{f}$. The optimization problem which must be solved is:

$$\begin{aligned} \min_{w_i^k, f_{q,t}^i \geq 0} \quad & \sum_{i=1}^{N}\sum_{q=1}^{M}\sum_{t=1}^{M} d_{qt} f_{qt}^i \\ \text{s.t.} \quad & \sum_{q=1}^{M} f_{qt}^i = h_i^t, \quad \forall i, \ \forall t \\ & \sum_{t=1}^{M} f_{tq}^i = \sum_{k=1}^{K} w_i^k p_q^k \quad \forall i, \ \forall q \\ & \sum_{k=1}^{K} w_i^k = 1, \ \ \forall i \end{aligned} \tag{5}$$

As in Step 2, this problem is a linear program which we solve using standard solvers.

Step 4. Given $\boldsymbol{p}^k$, $\mathbf{W}$, $\boldsymbol{f}$, $\boldsymbol{g}$ find the ground distance parameters $\boldsymbol{d}$. The optimization problem which must be solved is a quadratic program (QP), *i.e.* :

$$\begin{aligned} \min_{\boldsymbol{d}, \boldsymbol{\xi} \geq 0} \quad & \|\boldsymbol{d}\|^2 + \lambda_1 \sum_{i=1}^{N}\sum_{q=1}^{M}\sum_{t=1}^{M} d_{qt} f_{qt}^i + \lambda_2 \sum_{ijkl} \xi_{ijkl} \\ \text{s.t.} \quad & \text{vec}(\boldsymbol{d})^T (\mathbf{g}^{ij} - \mathbf{g}^{lk}) \geq 1 - \xi_{ijkl} \\ & d_{qt} = d_{tq}, \quad d_{tt} = 0, \quad \forall q, t = 1, \ldots, M \end{aligned} \tag{6}$$

Algorithm 1 summarizes the main steps of the proposed method.

Algorithm 1 EMD-NMF with ground distance learning

Input: $\mathcal{H} = \{\boldsymbol{h}_i\}_{i=1}^N$, $\mathcal{H}_s = \{(\boldsymbol{h}_i^s, \boldsymbol{h}_j^s, y_{ij}^s)\}_{i,j=1}^{N_s}$.
  Initialize $\mathbf{W}$ with a traditional NMF algorithm [1].
  Normalize $\mathbf{W}$ such that $\sum_k w_i^k = 1, \ \forall i = 1, \ldots N$.
  Initialize $\boldsymbol{d}$ setting $d_{qt} = 1$ if $q \neq t$, $d_{qt} = 0$ otherwise.
  while not converged
    Given $\mathcal{H}_s$ solve the set of transportations problems (3) with respect to $\mathbf{g}^{ij}$,$\mathbf{g}^{lm}$.
    Given $\boldsymbol{d}$, $\mathbf{W}$, solve (4) with respect to $\boldsymbol{p}^k$, $\boldsymbol{f}$.
    Given $\boldsymbol{d}$, $\mathbf{P}$, solve (5) with respect to $\mathbf{W}$, $\boldsymbol{f}$.
    Given $\mathbf{W}$, $\boldsymbol{p}^k$, $\boldsymbol{f}$, $\boldsymbol{g}$ solve (6) with respect to $\boldsymbol{d}, \boldsymbol{\xi}$.
  end
Output: $\mathbf{W}$, $\boldsymbol{p}^k \ \forall k$.

IV. Discovering High-level Activities with Semi-supervised EMD-NMF

In this paper we also propose to apply the proposed matrix factorization approach to the problem of the analysis of dynamic scenes recorded from surveillance cameras. Specifically we show how the semi-supervised EMD-NMF algorithm can be used for extracting high-level activities in complex scenes. Similarly to previous works [21], given a video, we propose to divide it into short clips and we adopt a bag-of-words approach for computing clip histograms.

We first compute level features from the video. Specifically we use a GMM-based background subtraction algorithm [22] to calculate for each pixel the foreground/background information. We also use a KLT tracker [23] to compute the optical flow, which measures the spatial shift between to consecutive frames of selected interest points. Is the spatial shift is less then a threshold T_{of} (*e.g.* 2 pixels) the point is considered static and thus discarded.

We divide the scene of interest in $n_x \times n_y$ patches and for each frame we combine foreground with optical flow information. For each patch the median optical flow direction is computed (in order to filter out noise) and it is quantized according to $N = 8$ directions. Patches with a percentage of foreground pixel major then a threshold T_{fg} (*e.g.* 50%) and no optical flow are considered as static. The active patches in a frame corresponds to elementary activities defined by 3 bins length vectors, which identify the position in the scene (x_c,y_c) and the motion direction ('0' for static, '1-9' for moving).

We collect a set of elementary activities over a sequence of frames, long enough in order to guarantee enough variety of events, and we use a standard k-means algorithm to compute a codebook of n_t words. After the codebook is defined, we extract the elementary activities from our sequence and quantize them according to the computed codebook. The temporal sequence is divided into short video clips and for each clip a histogram of occurred events is collected. The final clip histogram $\boldsymbol{h}_i \in I\!R^{n_t}$ is normalized to sum 1. As further described in the experimental section we manually annotate a small set of pairs of clips if they represent similar or different high level activities (*e.g.* vertical or horizontal traffic flows) in order to build the set $\mathcal{H}_s$. From this, a set of N_q quadruple $\{\boldsymbol{h}_i^s, \boldsymbol{h}_j^s, \boldsymbol{h}_m^s, \boldsymbol{h}_l^s\}$ is then selected in order to be fed to the optimization problem (6). Note that this set is selected randomly, thus probably generating some

	n^o frames	fps	frame size	n^o clips	n^o words
Basket	6000	12	368 × 320	100	16
Junction	12000	25	360 × 288	40	16

TABLE I. DATASETS AND EXPERIMENTAL SETUP.

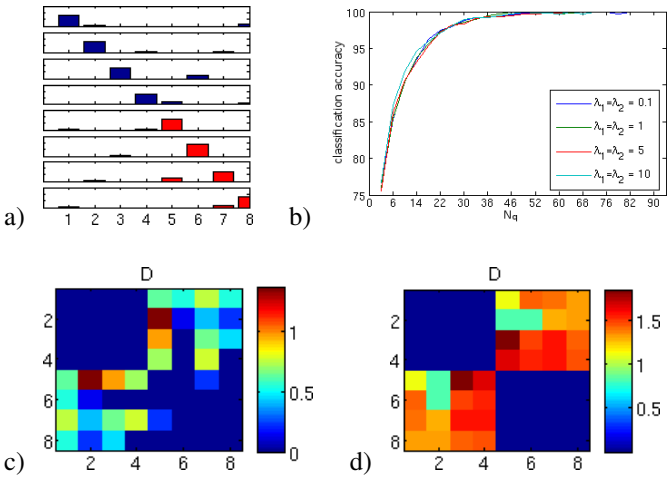


Fig. 1. (a) Synthetic input data, with $m = 8$ and $N = 8$. (b) Performance at varying N_q. Ground distance matrices $\boldsymbol{d}$, learnt (c) with (c) $N_q = 30$ and (d) $N_q = 40$. The final accuracy obtained is respectively equal to (c) 87.5% and (d) 100.0%.

redundant information. Then EMD-NMF is used to compute the prototype vectors $\boldsymbol{p}^k$ representing the salient activities and to learn the distance metric $\boldsymbol{d}$.

V. RESULTS

A. Datasets and Experimental Setup

We tested the effectiveness of our approach on two public datasets, QMUL Junction[1] and APIDIS basket[2]. The Junction dataset depicts a crowded traffic scene, while APIDIS shows a basketball game. The visual vocabularies used in these experiments are the same as in [6], [21]. More details on the datasets and the vocabulary used are reported in Table I. We also show the performance of our method on synthetic data. In our belief, synthetic data can help giving the reader an intuition of the method's working principles, and to understand the effect of varying the parameters values.

B. Synthetic data

Consider synthetic input data as in Fig.1(a). The color identifies to which of the two classes, red or $blue$, the histogram belongs. The performance on classification at varying N_q is shown in Fig.1(b). The final ground distance matrix $\boldsymbol{d}$ obtained for different values of N_q is shown in Fig.1(c,d). It can be easily observed that by increasing the number of quadruples N_q it is possible to get better performance, as well as a more meaningful distance matrix $\boldsymbol{d}$. In Fig.1(d), the learnt ground distances between bins 1-4 are zero, meaning that these are highly correlated. The same consideration holds for bins 5-8. The highest ground distance is between bin 3 and 5, which means that these two bins help to discriminate between the two classes $blue$ and red. Differently from Fig.1(d), in Fig.1(c), the learnt ground distance between bins 3,6 and 4,6 is low, which means that the set of quadruples given to learn $\boldsymbol{p}^k$, $\mathbf{W}$ and $\boldsymbol{d}$ is not representative enough.

[1] http://www.eecs.qmul.ac.uk/~jianli/Junction.html
[2] http://www.apidis.org/Dataset

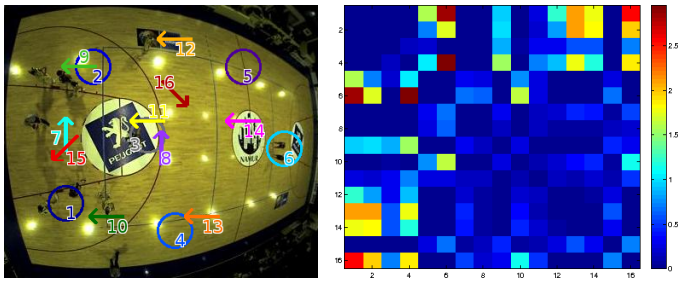


Fig. 2. Basket dataset. (a) visual vocabulary and (b) ground distance matrix $\boldsymbol{d}$ learnt with $N_q = 1000$.

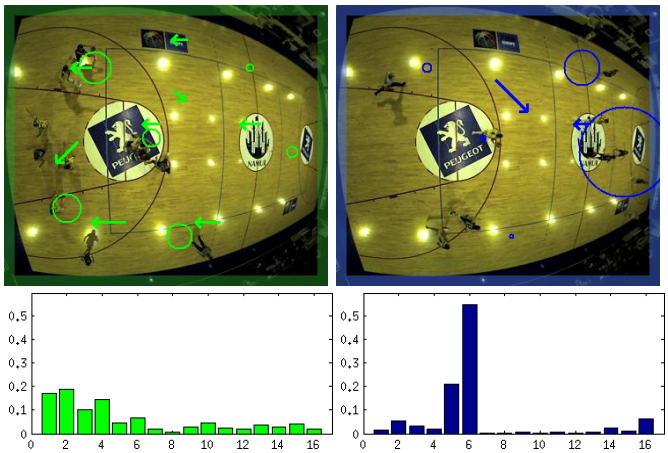


Fig. 3. Basket dataset: prototypes p^k computed based on groundtruth (left) ball in possession of blue team (right) ball in possession of yellow team.

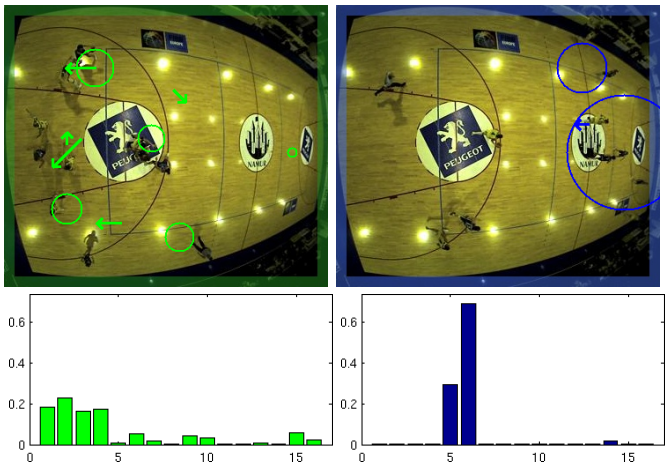


Fig. 4. Basket dataset: prototypes p^k learnt with $N_q = 1000$.

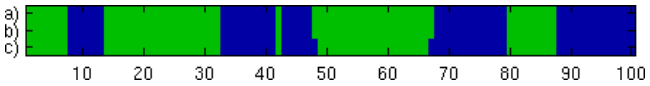
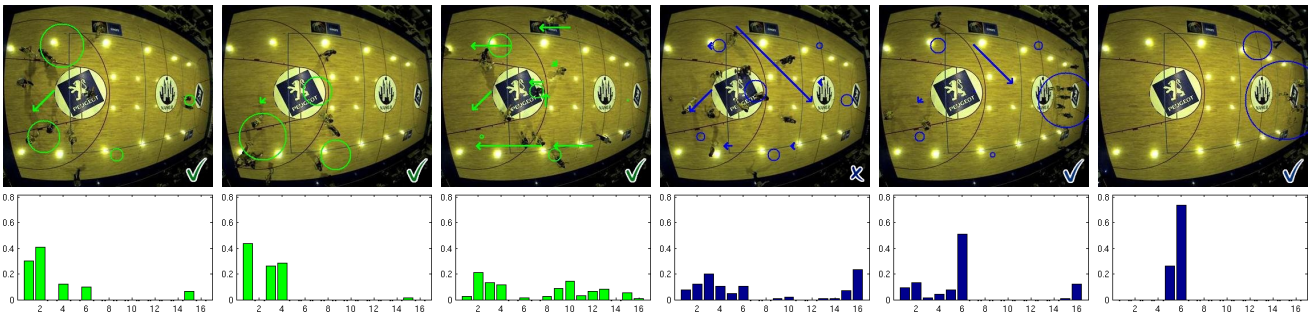


Fig. 5. Basket dataset: temporal segmentation bar obtained with (a) EMD-l1 [6], (b) our method and (c) groundtruth.



(a) (b) (c) (d) (e) (f)

Fig. 6. Basket dataset: sample keyframes with correspondent bag-of-words are shown; the color indicates their ground truth label, respectively (a-c) *green* and (d-f) *blue*. Clips (a-c,e,f) are correctly classified by our method; conversely clip (d) is an example of misclassification, corresponding to clip 68 in the temporal bar of Fig.5. This clip corresponds to the event of '*ball in possession of the yellow team*', but the player configuration is still very similar to the '*ball in possession of the blue team*' event.

C. APIDIS Basket Dataset

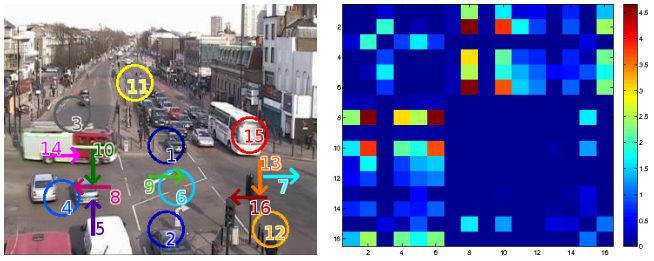
The results obtained on the basket dataset are reported in Fig. 2-6. The two events, *blue* and *green*, to be discovered correspond respectively to i) '*ball in possession of the yellow team*' and ii) '*ball in possession of the blue team*'. Figure 3 shows the prototypes computed based on the groundtruth, *i.e.* obtained by averaging over the clip histograms with the same groundtruth label. In Fig.4 we can observe that the prototypes learnt with $N_q = 1000$ are similar to the ones computed based on the groundtruth. However, the learnt prototypes $\boldsymbol{p^k}$ contains less active bins, *e.g.* the word 16 is missing from the learnt *blue* prototype (Fig.4). This is compensated by the fact that the learnt ground distances between the word 16 and the words $5, 6, 14$ is low, which allows to associate clips with high occurrence of word 16 to the *blue* event (Fig.2(b)). In Fig. 3 we can see that the *green* event is described mostly by the presence of words $1-4$ and 15. However, these words may not show up together at the same time (*e.g.* the occurrence of word 2 is zero for clip (a)), while words typical of one event can occur during a different one (*e.g.* occurrence of word 2 is non-zero for clip (d,e)). The ground distance learnt in this case help to compensate this noisy effect, intrinsic of the nature of the events (*i.e.* the game patterns played by a team will tend to be similar, but not completely identical). by balancing the words distribution among clips, thus emphasizing the occurrence of group of activities w.r.t. the occurrence of a single activity. Figure 5(b) shows the final temporal segmentation obtained with our method, which corresponds to an accuracy of 98%, which is comparable with the results obtained in [21], [6], see Table II and Fig. 5. To be specific, the groundtruth used in [21], [6] was set at frame level, while in this work we converted the groundtruth to clip level, thus all the frames belonging to the same clip has the same label. The changes slightly the result, *i.e.* from 92.25% to 92.0%, and allows an easier verification of the method's accuracy on the classification performance task (see Fig.5 and Tab.II for comparison). It is important to notice that the two misclassification corresponds to transition clips, which collect words occurrence from both the '*green*' and '*blue*' events. For example, in clip 68, whose bag-of-words representation is shown in Fig.6(d), players of the yellow team have just enter in possession of the ball, but they seem to delay the move to the opposite game court, thus making the game configuration more similar to the '*green*' event, *i.e.* when the blue team is on attack. In this case of 'transition clip', the classification error is due to the mixed nature of the clip, rather than to the failure of our method.

	pLSA	pLSA-bin	EMD-L1 [6]	our method
Basket	94.0%	92.0%	98.0%	98.0%

TABLE II. PERFORMANCE ON APIDIS BASKET DATASET.

	std pLSA [24]	hrc pLSA [24]	EMD-L1 [6]	our method
Junction	90.0%	77.5%	92.5%	92.5%

TABLE III. PERFORMANCE ON QMUL JUNCTION DATASET.



(a) (b)

Fig. 7. Junction dataset. (a) visual vocabulary and (b) ground distance matrix $\boldsymbol{d}$ learnt with $N_q = 1000$.

D. QMUL Junction

Similar observations discussed for the Basket dataset can be drawn for the Junction dataset. The results obtained on Junction datasets are reported in Fig. 7-10. The two events, *blue* and *green* to be discovered within this dataset correspond respectively to i) '*vertical traffic flow*' and ii) '*horizontal traffic flow*', where this last on includes alternate '*from left to right*' and '*from right to left*' horizontal flow. As we can see from Fig.3, the most discriminative words for the '*vertical flow*' event are $1, 2, 5, 6$, while words $8, 14, 16$ are mostly characterizing the event '*horizontal flow*'. In Fig.7(b) we can verify that the learnt ground distances among the words inside of each group are low, while among two words of different groups are high. Figure 10 shows the obtained video segmentation results, while a quantitative evaluation is reported in Table III.

VI. CONCLUSIONS AND FUTURE WORKS

We presented EMD-NMF, a semi-supervised method which applied to dynamic scene analysis allows, not only to discriminate among events, but also to learn the relationship of the atomic activities which characterize the event. Differently from dimensionality reduction approaches which map the features vectors to a different space, EMD-NMF allows a more intuitive

Fig. 8. Junction dataset. Prototypes p^k computed based on groundtruth: (left) horizontal and (right) vertical traffic flow.

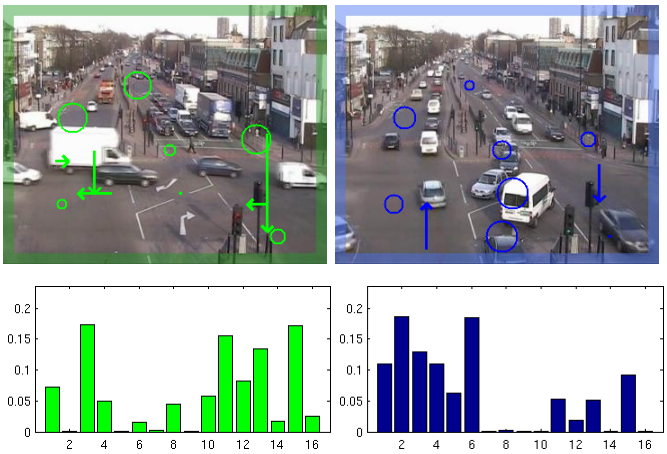


Fig. 9. Junction dataset. Prototypes p^k learnt with $N_q = 1000$: (left) horizontal and (right) vertical traffic flow.

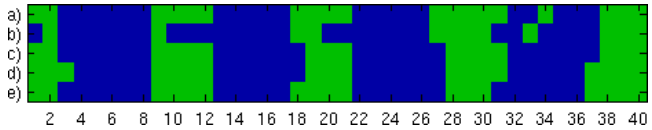


Fig. 10. Junction dataset: temporal segmentation results obtained with (a) standard pLSA [24], (b) hierarchical pLSA [24], (c) EMD-L1 [6], (d) our method and (e) groundtruth. The final accuracy obtained with our method is 92.5%.

interpretation of the learnt relationship. It is worth noting that, in the context of ground distance learning for computationally efficient variations of EMD [16], the problem of learning the ground distance corresponds to finding the best sorting. How to efficiently learn an effective word order would be part of our future works.

ACKNOWLEDGMENT

This research has been funded by the European 7th Framework Program, under grants VENTURI (FP7-288238) and xLiMe (FP7-611346).

REFERENCES

[1] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 1, pp. 788–791, 1999.

[2] R. Sandler and M. Lindenbaum, "Nonnegative matrix factorization with earth mover's distance metric," in *IEEE Conference on Computer Vision (CVPR)*, 2009.

[3] M. Cuturi and D. Avis, "Ground metric learning," *arXiv preprint arXiv:1110.2306*, 2011.

[4] F. Wang and L. J. Guibas, "Supervised earth mover's distance learning and its computer vision applications," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 442–455.

[5] G. Zen, E. Ricci, and N. Sebe, "Exploiting sparse representations for robust analysis of noisy complex video scenes," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 199–213.

[6] E. Ricci, G. Zen, N. Sebe, and S. Messelodi, "A prototype learning framework using emd: Application to complex scenes analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 513–526, 2013.

[7] T. Mauthner, P. M. Roth, and H. Bischof, "Instant action recognition," in *Image Analysis*. Springer, 2009, pp. 1–10.

[8] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4029–4032.

[9] N. Jiang, K. B. Englehart, and P. A. Parker, "Extracting simultaneous and proportional neural control information for multiple-dof prostheses from the surface electromyographic signal," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 1070–1080, 2009.

[10] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley. com, 2009.

[11] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *Signal Processing Letters, IEEE*, vol. 17, no. 1, pp. 4–7, 2010.

[12] F. Wang, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization," in *SIAM International Conference on Data Mining (SDM)*, 2008, pp. 1–12.

[13] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 11.

[14] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision (IJCV)*, vol. 40, no. 2, pp. 99–121, 2000.

[15] J. Wagner and B. Ommer, "Efficiently clustering earth mover distance," in *Asian Conference on Computer Vision (ACCV)*, 2010.

[16] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 840–843, 2006.

[17] O. Pele and M. Werman, "Fast and robust earth movers distances," *IEEE International Conference on Computer Vision (ICCV)*, pp. 460–467, 2009.

[18] S. Shirdhonkar and D. W. Jacobs, "Approximate earth movers distance in linear time," in *IEEE Conference on Computer Vision (CVPR)*, 2008.

[19] D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek, "Unsupervised clustering of multidimensional distributions using earth mover distance," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 636–644.

[20] Y. Zhang, X. Sun, H. Wang, and K. Fu, "High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model," vol. 10, no. 5, 2013, pp. 1055–1059.

[21] G. Zen and E. Ricci, "Earth mover's prototypes: a convex learning approach for discovering activity patterns in dynamic scenes," in *IEEE Conference on Computer Vision (CVPR)*, 2011.

[22] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision (CVPR)*, 1999.

[23] S. Birchfield, "KLT: An implementation of the Kanade-Lucas-Tomasi feature tracker," 2007.

[24] J. Li, S. Gong, and T. Xiang, "Global behaviour inference using probabilistic latent semantic analysis." in *BMVC*, 2008.