Exploiting Sparse Representations for Robust Analysis of Noisy Complex Video Scenes

Gloria Zen[1], Elisa Ricci[2], and Nicu Sebe[1]

[1] DISI, University of Trento
[2] DIEI, University of Perugia
{zen,sebe}@disi.unitn.it, elisa.ricci@diei.unipg.it

Abstract. Recent works have shown that, even with simple low level visual cues, complex behaviors can be extracted automatically from crowded scenes, *e.g.* those depicting public spaces recorded from video surveillance cameras. However, low level features as optical flow or foreground pixels are inherently noisy. In this paper we propose a novel unsupervised learning approach for the analysis of complex scenes which is specifically tailored to cope directly with features' noise and uncertainty. We formalize the task of extracting activity patterns as a matrix factorization problem, considering as reconstruction function the robust Earth Mover's Distance. A constraint of sparsity on the computed basis matrix is imposed, filtering out noise and leading to the identification of the most relevant elementary activities in a typical high level behavior. We further derive an alternate optimization approach to solve the proposed problem efficiently and we show that it is reduced to a sequence of linear programs. Finally, we propose to use short trajectory snippets to account for object motion information, in alternative to the noisy optical flow vectors used in previous works. Experimental results demonstrate that our method yields similar or superior performance to state-of-the arts approaches.

1 Introduction

Developing computational models able to emulate the human ability to interpret complex visual scenes is one of the biggest challenges in computer vision research. Many factors make this task difficult, *e.g.* the fact that humans have *a-priori* knowledge, that they are able to extract and select relevant visual informations, and that they can easily resolve the strong ambiguity typical of many visual scenes (activities associated with the same semantic information can be performed differently while activities having a different interpretation may be acted similarly). The complexity of semantic scene interpretation further increases when many objects are present in a scene.

Despite the many difficulties, several advances have been made in the past few years and many approaches have been proposed to extract semantic behaviors from video scenes. Many of them explicitly focused on analyzing complex and crowded scenes depicting public spaces in video surveillance applications [1–6].

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 199–213, 2012.

Visual analysis of such complex scenes faces two main problems. First of all, as many objects are present, being able to account for spatio-temporal dependencies among them poses a challenge in terms of computational complexity. Secondly, the visual information that can be extracted in this scenario is limited to simple low level cues (*e.g.* background/foreground information, optical flow) as object tracking approaches cannot be employed due to the many objects and to the several occlusions. These features are inevitably uncertain and noisy. Recent works [1–6] have shown how powerful statistical machine learning approaches can be used to implicitly handle noisy, uncertain visual information, leading to excellent results in terms of salient behaviors and anomalous activities identification. However none of the previous works have tackled explicitly this issue, *i.e.* developing methods targeted to noisy data analysis.

The main contribution of this paper is a novel approach for the analysis of complex scenes specifically tailored to cope with the uncertainty and the noise arising in visual modeling of complex dynamic scenes. Similarly to previous works, we follow a non-object centric perspective and compute simple features accounting for motion and foreground/background information. However in this paper, to calculate motion features, we do not rely on noisy optical flow vectors but adopt a representation based on short trajectory snippets. Differently from previous works [1–6], we model the task of extracting salient activities as a matrix factorization problem and we consider as objective function the Earth Mover's Distance (EMD) [7], which is well-known to be a robust metric in case of noisy histogram comparison. To further reduce the influence of noisy data we also constrain the computed vector basis to be sparse. In a surveillance scenario as the one considered here, where scenes have multiple temporal activity patterns happening simultaneously, a sparsification procedure is crucial for semantic scene interpretation purposes, helping to identify the atomic activities which are distinctive of a specific high level behavior. To our knowledge, no previous works have addressed complex video scene analysis under a sparse coding framework. Our method has been tested on several video datasets, all of which are publicly available. The experimental results show that our approach successfully identifies high-level activities and spots anomalous patterns and it is very competitive with respect to state-of-the-art algorithms [2, 3, 5], often outperforming them.

2 Related Works

This paper follows recent works on non-object centric analysis of complex scenes [1–6]. However it departs from many previous works [1–4] as we do not rely on Probabilistic Topic Models for inferring high-level activities. Instead we model the task of discovering spatio-temporal activity patterns as a nonnegative matrix factorization problem. Up to our knowledge, no previous works have considered the problem under this perspective. In this paper we also explicitly force the inferred latent representation to be sparse leading to a novel algorithm for activity localization. This aspect has rarely been addressed in previous works. Exceptions are the algorithms in [8, 9] which are however very different from ours as they are

based on a Probabilistic Topic Model framework. Sparse coding has become very popular in computer vision and has been successfully applied in many problems related to video analysis [10, 11]. Differently from previous works, the bases of our dictionary are computed using EMD as distance function, leading to sparse representations with a nice grouping structure.

The use of trajectory-based representation for embedding motion information into a bag-of-words approach has been successfully used in many previous works [12–14]. However these works have all considered action recognition tasks, while few approaches [15] have focused on showing the effectiveness of such a representation for complex scene analysis from video surveillance data.

The proposed method has some similarity with our previous works on EMD clustering [5, 6], as nonnegative matrix factorization is practically a clustering algorithm. However, with respect to [5, 6], this method is more scalable and produces more interpretable results due to the sparse constraints. Moreover we are not forced to use a dense histogram representation as it is done in [5, 6], avoiding the need of preprocessing steps. Importantly, even without these preprocessing steps, our approach produces better performance in the considered datasets.

Nonnegative matrix factorization (NMF) [16] has been considered in many works in computer vision as well as in other disciplines. NMF provides an elegant framework to achieve sparsity on the basis or coefficient matrices by using the theoretically sound ℓ_1 regularizer or other composite regularizers [17, 18]. Recently Sandler and Lindenbaum [19] proposed a variant of NMF which uses the EMD as objective function. However our work is very different from [19] as it considers specifically the problem of complex scene analysis and integrates effectively a sparse constraint into the Earth Mover's Distance matrix factorization.

3 Earth Mover's Distance

Let $\boldsymbol{h}, \boldsymbol{p}$ be two histograms normalized to unit mass. The Earth Mover's Distance $\mathcal{D}_{EMD}(\boldsymbol{h}, \boldsymbol{p})$ [7] is obtained as the solution of the transportation problem:

$$\min_{f_{qt} \geq 0} \sum_{t,q=1}^{Q} d_{qt} f_{qt} \qquad \text{s.t.} \quad \sum_{q=1}^{Q} f_{qt} = h^t, \ \sum_{t=1}^{Q} f_{qt} = p^q \tag{1}$$

The variable f_{qt} denotes a flow representing the amount transported from the q-th supply to the t-th demand and d_{qt} the ground distance between q and t. Usually d_{qt} is defined by L_1 or L_2 distance. The problem (1) is a Linear Program (LP) which can be efficiently solved due to the special structure of its sparse constraints [7, 20]. However, in the case of high dimensional histograms, solving (1) can be very time consuming due to the large number of flow variables involved. Several methods have been proposed in the past to speed up EMD distance computation [20, 21]. In particular in [20], it is shown how, for histograms normalized to unit mass and EMD with L_1 as ground distance (EMD-L_1), every positive flow between faraway histograms bins can be replaced by a sequence of flows between neighboring bins. This implies that (1) simplifies as:

$$\min_{f_{qt}\geq 0} \sum_q \sum_{t\in\mathcal{N}(q)} f_{q,t} \qquad \text{s.t.} \sum_{t\in\mathcal{N}(q)} f_{q,t} - \sum_{t\in\mathcal{N}(q)} f_{t,q} = h^q - p^q \;\forall q \tag{2}$$

where the index q corresponds to the position of a specific bin of a histogram and its neighborhood $\mathcal{N}(q)$ is represented by the adjacent bins. For example, in case of 2-dimensional histograms if the q-th bin corresponds to the position (i, j) in the grid, its neighborhood is represented by the index set $\mathcal{N}(q) = \{(m, n) : (m, n) \in \mathcal{G}, d_{i,j;m,n} = |i-m|+|j-n| = 1\}$ with $\mathcal{G}$ being the set of indexes corresponding to all the grid nodes. It is worth noticing that using (2) the number of flow variables involved reduces from $O(Q^2)$ in (1) to $O(Q)$, where Q is the number of bins in the original histogram. This is greatly beneficial in terms of computational cost since the number of variables is a dominant factor in the time complexity of all LP algorithms.

4 Mining Sparse Activity Patterns in Complex Scenes

In this Section, the proposed approach for extracting high-level activities in complex scenes is presented. First, the features used to represent the short video clips are described (Subsection 4.1); then, the proposed learning approach is illustrated (Subsection 4.2).

4.1 Computing Clip Histograms

Similarly to previous works [2, 3, 5], we divide a video into short clips and we adopt a bag-of-words approach for computing clip histograms. First, we construct a codebook of trajectory snippets (*trajectons*) as described in [12]. Feature trajectons, *i.e.* sequences of (x_t,y_t) positions over time, are computed by cropping the features trajectories extracted using a KLT tracker [22]. Using a short video segment as training set, a codebook of trajectons is computed by clustering the obtained trajectory snippets into a pre-specified number of clusters n_t. While in general standard k-means can be employed in this phase, in our specific application we manually selected the codebook ensuring that trajectories cover all the space of possible motion orientations. For the dynamic of the scene, in fact, a small codebook defining different motion orientation is more suitable to distinguish between the most relevant activities. This simple codebook corresponds to features more robust to noise than when considering optical flow vectors. In line with [22], we consider trajectory snippets formed by 10 positions in the trajectory. The subsequent phase consists in extracting low level features from the video and quantizing it according to the codebook generated. Specifically for each pixel we compute the foreground/background information using a simple dynamic Gaussian-Mixture background model as background subtraction algorithm [23]. We use KLT to compute trajectory snippets and assign them a label according to the nearest snippets in the codebook. The features extraction process is illustrated in Fig.1. Then we divide the scene of interest in $n_x \times n_y$ patches, in order to take into account the location where the activities take place.

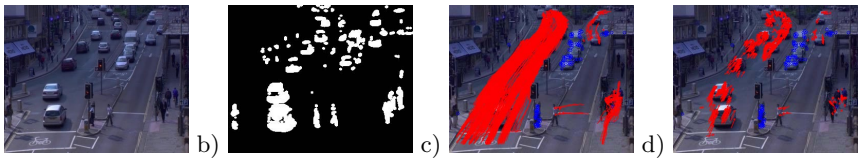

Fig. 1. Low level visual features used in our approach. (a) Original video frame and (b) associated foreground mask. (c) Trajectories (red) extracted with KLT tracker. (d) Trajectory snippets (red) and static pixels (blue) used to construct clip histograms.

We also divide the video into clips. A histogram counting the occurrences of trajectons labels is formed for each clip and each patch. Moreover, for each patch a further bin is used to account for static activities, *i.e.* pixels of foreground that do not belong to trajectons. The clip histogram $\boldsymbol{h}_i \in I\!R^{n_x \times n_y \times n_t}$ is obtained concatenating the patch histograms.

4.2 Discovering Activities with Sparse EMD Matrix Factorization

Given a training set of clip histograms $\mathcal{H} = \{\boldsymbol{h}_1,\ \boldsymbol{h}_2, \ldots \boldsymbol{h}_N\}$, we model the task of discovering high-level activities as the problem of finding a set of basis $\mathcal{P} = \{\boldsymbol{p}^1,\ \boldsymbol{p}^2,\ \ldots \boldsymbol{p}^K\}$, with $K << N$, and a matrix of mixing coefficients $\mathbf{W} = [\boldsymbol{w}_1\ \boldsymbol{w}_2 \ldots \boldsymbol{w}_N]$, $\boldsymbol{w}_i \in I\!R^K$, such that, for each clip, the weighted sum of the computed basis should be as close as possible to the original clip histogram according to the Earth Mover's Distance. More formally the following optimization problem is formulated:

$$\min_{\boldsymbol{p}_k, \mathbf{W} \geq 0} \sum_{i=1}^{N} \mathcal{D}_{EMD}(\boldsymbol{h}_i, \sum_k w_i^k \boldsymbol{p}^k) \tag{3}$$

$$\text{s.t.} \qquad \omega_m \leq \Omega(\boldsymbol{p}^k) \leq \omega_M, \quad \forall\ k = 1 \ldots K \tag{4}$$

The imposed constraints force the computed bases to be sparse. To enforce sparsity, as in previous works on NMF [17, 18], we adopt the following measure:

$$\Omega(\mathbf{x}) = \frac{\sqrt{n} - \|\mathbf{x}\|_1 / \|\mathbf{x}\|_2}{\sqrt{n} - 1} \tag{5}$$

where $\mathbf{x} \in I\!R^n$. In practice the constraint (4) impose a lower and an upper bound (respectively ω_m and ω_M) to the level of sparsity of the computed prototypes.

By replacing the definition of EMD with L_1 ground distance (2) and $\Omega(\cdot)$ into (3), the following optimization problem must be solved:

$$\min_{p_q^k, w_i^k, f_{q,t}^i \geq 0} \sum_{i=1}^{N} \sum_q \sum_{t \in \mathcal{N}(q)} f_{q,t}^i \tag{6}$$

$$\text{s.t.} \qquad \sum_{t \in \mathcal{N}(q)} f_{q,t}^i - \sum_{t \in \mathcal{N}(q)} f_{t,q}^i = h_i^q - \sum_k w_i^k p_q^k, \quad \forall q,\ \forall i$$

$$\sum_k w_i^k = 1, \quad \forall i \qquad \sum_q p_q^k = 1 \quad \forall k$$

$$\|\boldsymbol{p}^k\|_2 \leq \frac{1}{c_M}\mathbf{e}^T\boldsymbol{p}^k \quad \forall k \tag{7}$$

$$\frac{1}{c_m}\mathbf{e}^T\boldsymbol{p}^k \leq \|\boldsymbol{p}^k\|_2 \quad \forall k \tag{8}$$

where $c_M = \sqrt{Q} - \omega_M(\sqrt{Q}-1)$ and $c_m = \sqrt{Q} - \omega_m(\sqrt{Q}-1)$, $Q = n_x \times n_y \times n_t$ and $\mathbf{e} \in I\!R^Q$ is a vector of ones. The normalization constraints impose that each basis vector $\boldsymbol{p}^k$ and each column of the coefficient matrix $\mathbf{W}$ are normalized to sum one. This implies that $\sum_q \sum_k w_i^k p_k^q = 1, \ \forall\ i$, *i.e.* the reconstructed histograms are normalized to unit mass as required by EMD definition (2). The additional constraints (7) and (8) are imposed to force the bases to be sparse vectors.

The optimization problem (6) is not convex. However, to efficiently solve it, in this paper we devise an approximate approach based on an alternate optimization scheme. We first consider (6) when constraints (7) and (8) are not imposed. In this case the problem (6) is still not convex. However if the coefficient matrix $\mathbf{W}$ is fixed, (6) is convex with respect to $p_q^k, f_{q,t}^i$. Similarly, with fixed basis vectors $\boldsymbol{p}^k$, (6) is convex with respect to $w_i^k, f_{q,t}^i$. To solve it, an alternate optimization scheme can be devised where each single optimization problem reduces to a LP. This approach, which turns out to be a special case of the algorithm proposed in [19], can be shown to converge to a local minimum. The convergence proof and further details can be found in the supplementary material.

If the constraints (7) are also considered, the optimization problem (6) can still be solved with an alternate optimization scheme and, in particular, as a sequence of convex optimization problems. Solving with respect to $w_i^k, f_{q,t}^i$ with variables p_q^k fixed is still a LP, while solving with respect to $p_q^k, f_{q,t}^i$ having $\mathbf{W}$ fixed is a Second Order Cone Programming (SOCP) which can be solved efficiently with standard solvers (see supplementary material). However, when the constraints (8) are also considered, solving (6) with respect to $p_q^k, f_{q,t}^i$ and w_i^k fixed is not convex anymore. Therefore, inspired by previous works on NMF [17], we adopt an approximate technique to solve it. The approach is based on the Tangent Plane Constraint (TPC) method [24] and basically consists in approximating the non convex cone constraints by linear constraints and specifically by tangent plane constraints.

The algorithms we develop for solving (6) are shown in Algorithm 1 and Algorithm 2. In particular the alternate optimization approach used to solve (6) is illustrated in Algorithm 1. Step 5 of Algorithm 1 consists in solving (6) subject to (7) and (8) with the TPC method. The TPC method is illustrated in Algorithm 2.

To solve the proposed optimization problem we adopt some practical solutions which reduce the computational cost of our approach and then makes it more appealing to large scale computer vision applications. First of all we note that the convex constraints (7) are not particularly important in order to guarantee sparse solutions. In fact, rather than imposing an upper bound on the maximum level of sparsity, it is much more important to guarantee a minimum level of sparsity. This implies that in practice we can omit convex constraints (7). This is of paramount importance in practical applications since the sequence of SOCP

Algorithm 1. EMD Clustering

Input: Original clips histograms $\mathcal{H} = \{\boldsymbol{h}_1,\ \boldsymbol{h}_2, \dots \boldsymbol{h}_N\}$.
  Initialize $\mathbf{W} = [\boldsymbol{w}_1\ \boldsymbol{w}_2 \dots \boldsymbol{w}_N]$ with positive random values.
  Normalize the columns of $\mathbf{W}$ such that $\sum_k w_i^k = 1,\ \forall i = 1, \dots N$.
  while not converged
    Solve (6) s.t. (7) and (8) with respect to $\boldsymbol{p}^k, \boldsymbol{f}$ using Algorithm 2.
    Solve the LP (6) with respect to $\mathbf{W}, \boldsymbol{f}$.
  end
Output: $\mathbf{W}$, $\boldsymbol{p}^k\ \forall k$.

Algorithm 2. Algorithm for computing sparse prototypes

Input: Original clips histograms $\mathcal{H} = \{\boldsymbol{h}_1,\ \boldsymbol{h}_2, \dots \boldsymbol{h}_N\}$ and coefficent matrix $\mathbf{W}$. The parameters ω_m and ω_M specifying the desired sparsity levels.
  Compute $c_M = \sqrt{Q} - \omega_M(\sqrt{Q} - 1)$ and $c_m = \sqrt{Q} - \omega_m(\sqrt{Q} - 1)$.
  Solve (6) s.t. (7) with respect to $\boldsymbol{p}^k, \boldsymbol{f}$.
  Initialize the index set of violated constraints $\mathcal{V}^0 = \emptyset$.
  Set $t = 0$.
  while not converged
    $\bar{\boldsymbol{p}}^k = \boldsymbol{p}^k, \forall k$.
    Find $\bar{\boldsymbol{p}}^k$ violating (8); update $\mathcal{V}^{t+1} = \mathcal{V}^t \cup \{r : r = 1, \dots, K, \frac{1}{c_m}\mathbf{e}^T\boldsymbol{p}^r \geq \|\boldsymbol{p}^r\|_2\}$
    $\forall r \in \mathcal{V}^{t+1}$ compute the projection $\bar{\pi}_r = \pi(\bar{\boldsymbol{p}}^r)$ as shown in [18].
    $\forall r \in \mathcal{V}^{t+1}$ compute the tangent plane $\boldsymbol{t}_{r,\bar{\pi}}^{t+1}$ to cone (8) in $\bar{\pi}_r$
    Solve (6) s.t. (7) and to the tangent plane constraints $(\boldsymbol{p}^r)^T\boldsymbol{t}_{r,\bar{\pi}}^{t+1} \geq 0,\ \forall r \in \mathcal{V}^{t+1}$ with respect to $\boldsymbol{p}^k, \boldsymbol{f}$.
    $t = t + 1$.
  end
Output: $\boldsymbol{p}^k, \forall k = 1, \dots, K$.

problems in Algorithm 2 (Step 3 and Step 4) reduces to a sequence of efficient LP problems. In alternative, as for the constraints (8), also in case of (7) tangent plane constraints can be devised. Still, the overall optimization problem reduces to a LP. In our experiments we used the former solutions ($\omega_M = 1$).

While the TPC method is guaranteed to converge (*i.e.* Algorithm 2 always converges) [24] the alternate optimization problem in Algorithm 1 is not guaranteed to converge. For this reason in [17], in case of TPC applied to NMF, a more robust but slower approach is proposed. While we also cannot prove the convergence of Algorithm 1 when using TPC method in our experiments we did not observe problems of convergence (Fig.5).

While our approach can be generally applied to several types of histogram data, for computational efficiency reasons in our experiments we consider two-dimensional histograms obtained by reshaping the clip histograms as $\boldsymbol{h}_i \in \mathbb{R}^{n_x \times n_y \times n_t}$. In this way the EMD objective function operates on a grid as neighborhood structure, where neighboring bins in a histogram mostly corresponds to the same features (*e.g.* same trajectons) computed in neighboring patches.

Table 1. Details on the datasets and the experimental setup

	n° frames	fps	video duration	frame size	$n_x \times n_y \times n_t$	Q	clip duration	n° clips
Junction	90000	25	60'	288×360	8×6×9	432	12"	300
Junction2	78000	25	52'	288×360	8×6×9	432	12"	260
Roundabout	93500	25	62'	288×360	12×9×9	972	12"	311

5 Experimental Results

5.1 Experimental Setup

Experiments were conducted on three publicly available datasets collected from researchers of Queen Mary University, namely **Junction**, **Junction2** and **Roundabout**. The videos depict some complex traffic scenes in London and have been extensively used in previous works [1–3, 5, 25]. The ground truth corresponding to activities found by a human annotator are also publicly available[1]. To compare our approach with state-of-the-art methods [3, 5] we also use the code and the results made available by other research groups[2,3]. Our method is implemented in C++ using the publicly available library OpenCV for the video processing and feature extraction parts while MATLAB is employed for Algorithms 1 and 2. The code will be made available to the community[4]. More details about the datasets used and our experimental setup are summarized in Table 1. Video results associated with our approach are provided in the supplementary material.

5.2 Testing the Proposed Approach

The first series of experiments is aimed to demonstrate the ability of the proposed approach to extract high level activities by selecting the most significant elementary features in the scene. Figure 2 depicts the high level activity patterns computed with our approach for the Junction dataset; these three main patterns correspond to vertical traffic flow, horizontal flow from left to right and from right to left. In the same figures, the $n_t + 1$ elementary features are plotted in different colors: green circles correspond to static activities and the other colors identify the n_t different trajectons, whose main direction is indicated by arrows. Also, the intensity of each elementary feature n_t is represented by N_e colored patches that are plotted with a Gaussian distribution around the patch centroid (i, j). The number N_e is proportional to $p_k^{i,j,t}$.

Varying the required minimum sparsity level, and specifically with ω_m close to one, only few elementary features are active in the final prototype representation. Furthermore a grouping effect, which must be ascribed to the use of EMD as objective function, is observed, as elementary features in adjacent regions tend

[1] http://www.eecs.qmul.ac.uk/~jianli/Dataset_List.html
[2] http://disi.unitn.it/~zen/emd.html
[3] http://www.vision.ee.ethz.ch/~calvin/publications.html
[4] http://disi.unitn.it/~zen/sparse_emd.html

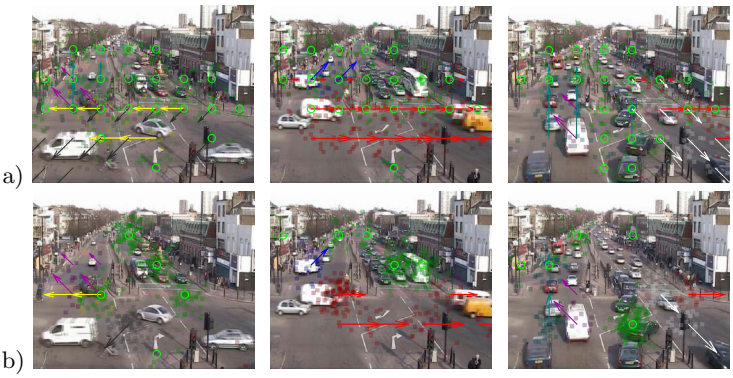

Fig. 2. Junction dataset. High level activities automatically extracted with our approach at different levels of sparsity (a) $\omega_m = 0.0$, (b) $\omega_m = 0.9$.

Table 2. Clustering accuracy at varying sparsity level ω_m

ω_m	0	0.1	0.3	0.5	0.7	0.9
Junction2 (48 clips)	89.58%	89.58%	89.58%	89.58%	**91.67%**	89.58%
Roundabout (60 clips)	88.33%	88.33%	**90.00%**	**90.00%**	**90.00%**	**90.00%**
Junction (39 clips)	**89.74%**	**89.74%**	**89.74%**	**89.74%**	**89.74%**	84.62%

to be active or not active together. The effect of sparse grouping activities can be observed in Figure 4.

Similar observations can be made in the case of the Roundabout dataset (Fig.3) where six main activities are extracted. In details, the yellow and light/dark green activities correspond to the same higher level activity (top-right traffic lights on green) but at different traffic flow intensity. The blue and red activities correspond, respectively, to central-bottom and left traffic lights on green.

Table 2 shows the clustering accuracy obtained by varying ω_m for the three datasets considered. The results correspond to a two clusters groundtruth segmentation. Imposing sparsity constraints in the learning process is important for the semantic interpretation of video contents, and our experiments demonstrate that this is not negatively affecting the accuracy. In some cases, when a high degree of sparsity ($\omega_m = 0.7$) is imposed, the performance can also be better. This can be ascribed to the beneficial effect of sparsity constraints in filtering out noisy features. In few cases instead, for severe level of sparsity ($\omega_m = 0.9$) the accuracy can slighly degrade. This is probably due to the loss of some details that could be useful for some classes' discrimination.

Convergence. As discussed in Section 4.2, Algorithm 1 is not guaranteed to converge when the Tangent Plane Constraint method [17, 24] is adopted. However, in our experimental results we mostly observed a convergent behavior.

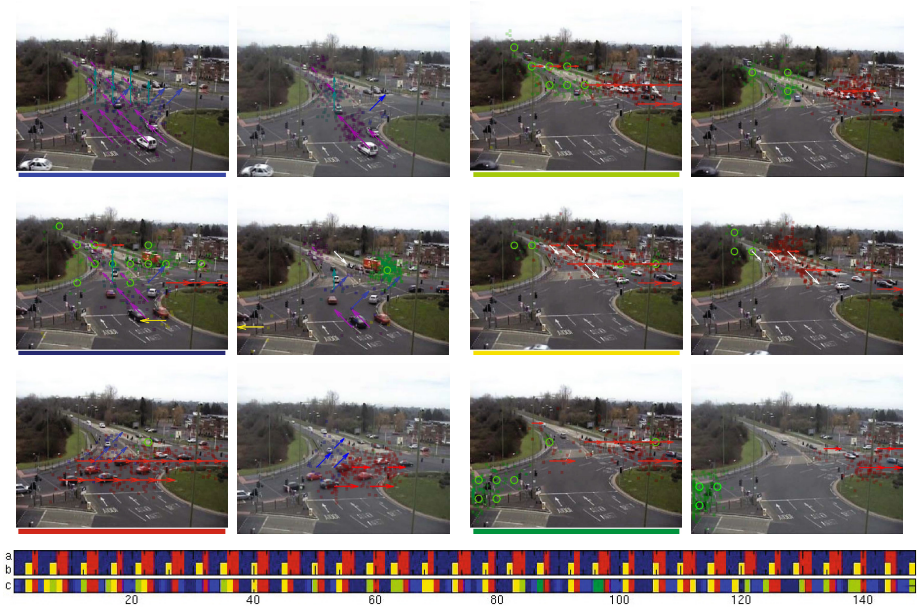

Fig. 3. Roundabout dataset. High level activities automatically extracted with our approach at different levels of sparsity (left) $\omega_m = 0.0$ and (right) $\omega_m = 0.9$; (a) GT [25] (b) GT considering three classes set by the authors and (c) temporal segmentation obtained with our approach.

Figure 5 depicts two examples of convergence for the experiments conducted on the Junction dataset for $\omega_m = 0.5$ and $\omega_m = 0.7$; specifically the value shown is the L_1 distance computed between successive vector bases $\boldsymbol{p}_k^t$ and $\boldsymbol{p}_k^{t-1}$, at each iteration t. Some convergence issues were observed for values of ω_m close to 1. However these situations are of less practical utility as the best clustering accuracy is typically obtained for $\omega_m = 0.7/0.8$.

5.3 Comparison with Previous Works

In this Subsection we report some results aimed at comparing the proposed approach with previous methods [2, 3, 5, 25].

Temporal Segmentation. We first consider the datasets Junction and Roundabout, as for these videos a ground truth annotation with two classes (horizontal and vertical traffic flows) is provided in [25]. However, it is easy to observe that the natural classes of traffic flows are more than two. In particular, for the Roundabout dataset, this is due to the presence of more that two traffic lights regulating the vehicles' flow and to varying traffic flows intensity (*e.g.* traffic light is on green but there are no vehicles in the lane). Therefore Kuettel *et al.* [3] consider a temporal segmentation with $K = 6$. Moreover they use clips of $3s$

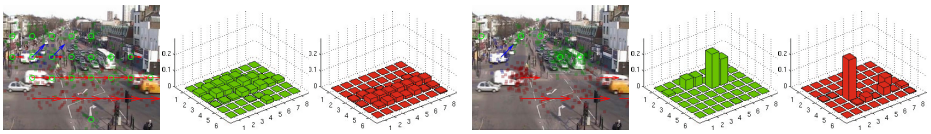

Fig. 4. Effect of combining EMD with sparsity constraints, shown on Junction dataset. Prototype obtained with our method setting (left) $\omega_m = 0.0$ and (right) $\omega_m = 0.9$. The 2D histogram is shown for zero motion and for rightward motion elementary features (drawn respectively in green and red).

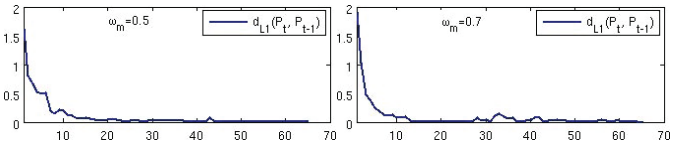

Fig. 5. Junction dataset. Convergence analysis of our approach for different levels of sparsity (left) $\omega_m = 0.5$ and (right) $\omega_m = 0.7$.

length instead of $12s$ as related works. In our experiments, we also show results obtained with $K = 6$. In Fig.6 the results obtained with the different methods are compared. Specifically the segmentation computed with Probabilistic Latent Semantic Analysis (PLSA) e hierarchical PLSA [25], Dependent Dirichlet Process Hidden Markov Model (DDP-HMM) [3], Earth Mover's Prototypes (EMP) [5] and our approach are compared. As shown in the plot, all the approaches obtained consistent results with respect to ground truth annotation. Similar comparative results are also reported for the Junction (Fig.7) and the Junction2 (Fig.8) datasets. For these datasets, a qualitative comparison with the work in [2] is also possible, as our approach is able to extract the same recurrent activities shown in [2]. Note that for the Junction2 dataset only the results provided by [3] are available.

A quantitative comparison between our approach and the methods [3, 5, 25] is also provided in Table 3. Observing the first two rows of the table it is evident that our approach outperforms previous methods in the Roundabout dataset, while it is the second best for Junction. The last row in the table shows the segmentation results for a longer sequence of the Roundabout dataset. In this sequence the best results are obtained by the approach proposed in [5]. However it is worth noting that these results correspond to a different experimental setup, as in [5] all the 148 clips are used as training set, while, similarly to [25] we consider a more challenging task and we train only on 60 clips and use the remaining clips as test set. In these experimental conditions we outperform previous methods.

Comparison with [5]. As observed in Section 2, both our approach and the one presented in [5] use EMD to compute the basis vectors $\boldsymbol{p}^k$ representing the discovered activities. In this section we compare both methods in terms of

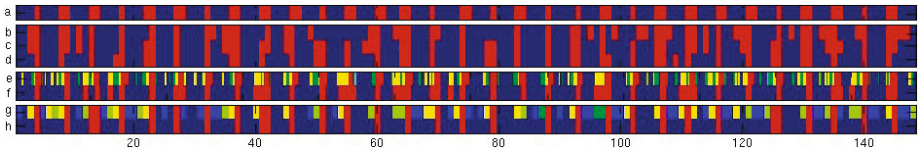

Fig. 6. Roundabout dataset. (a) Ground truth annotation [25] and temporal segmentation results obtained with (b,c) standard and hierarchical pLSA [25], (d) EMP [5], (e,f) HDP-HMM [3] and (g,h) our approach.

Fig. 7. Junction dataset. (a) Ground truth annotation [25] and temporal segmentation results obtained with (b,c) standard and hierarchical pLSA [25], (d) EMP [5], (e,f) DDP-HMM [3] and (g,h) our method. (Left) extracted high-level activities $\omega_m = 0.9$.

computational cost. It worth noting that we could not use our features as the algorithm [5] does not scale to long histograms. Long histograms can be more suitable in case of EMD learning, as the similarity among bins is naturally imposed by the patch division structure. This is different from [5], where an elementary activity order needs to be established manually to create clip histograms. However, to compare our approach with [5], we use a dense histogram representation as described in [5]. We compute one-dimensional histograms where each bin represents an atomic activity (in [5] an atomic activity consists in a specific motion pattern occurring in a specific image region). Atomic activities must be manually sorted. In this paper we consider five different atomic activity orders. Our results are the average of these five runs. Figure 9 shows the results of our comparison. From the plots it is evident that, when histograms dimension Q increases, our approach is much more scalable. On the other hand, as expected, our method has modest performance in terms of accuracy. Our best results are obtained for $Q = 32$ and correspond to an accuracy equal to 75%. On the same data the algorithm in [5] reaches an accuracy of 92%. However, as demonstrated by Table 2, similar performance (91.67%) can be obtained with our approach when a sparse histogram representation is adopted.

Anomaly Detection. In this paragraph we briefly show that our approach can be used to identify anomalous and rare activities. To this aim the mixing coefficients $\mathbf{W}$ can be analyzed. Given a clip histogram $\boldsymbol{h}_i$ and the associated weights $\boldsymbol{w}_i$, we consider the corresponding activity as rare if it cannot be

Table 3. Comparison with previous approaches: clustering accuracy

	std pLSA [25]	hrc pLSA [25]	DDP-HMM [3]	EMP [5]	our approach ($\omega_m = 0.7$)
Roundabout (60 clips)	81.67%	75.00%	85.00%	86.67%	**90.00%**
Junction	89.74%	76.92%	87.18%	**92.31%**	89.74%
Roundabout (148 clips)	84.46%	72.30%	85.14%	**86.40%**	*85.81%*

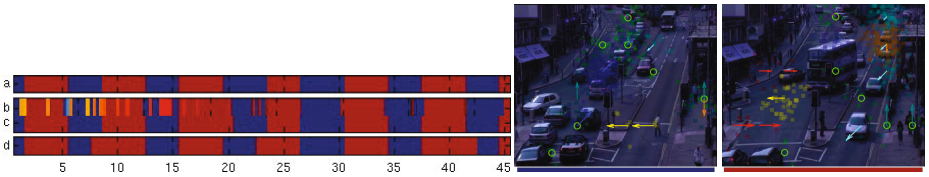

Fig. 8. Junction2 dataset. (a) Ground truth annotation and temporal segmentation results obtained with (b) DDP-HMM [3] and (c) our method. (left) extracted high-level activities with $\omega_m = 0.9$.

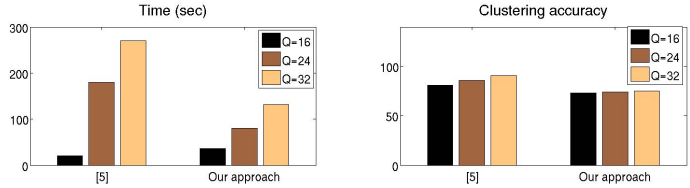

Fig. 9. Junction2 dataset. Comparison with [5] on clustering dense histograms data.

Fig. 10. Junction dataset. (Left) Anomaly score. (Right) Representative frames extracted from the detected anomalous clips: (4,27) interruption of vertical traffic flow due to a fire engine passing, (9) leftward horizontal and (15) vertical flow, both interleaved with rightward horizontal traffic.

"explained" by the computed bases $\boldsymbol{p}_k$. This practically means that none of the w_i^k is close to one, *i.e.* the standard deviation $\sigma_{\boldsymbol{w}_i}$ of the coefficients w_i^k is small. With this intuition, $\sigma_{\boldsymbol{w}_i}$ can be used as anomaly score. The anomaly score computed for the Junction dataset is shown in Fig.10. Negative peaks identifying the anomalous clips are highlighted in green. Clips 4 and 27 correspond to the interruption of traffic flow due to a fire engine passing, Clips 9 is and 15 are anomalous as they are associated, respectively, to leftward horizontal flow and to vertical traffic, but they are also interleaved with rightward horizontal flow. These results are similar to those in previous works [2, 5] and correspond to the anomalies indicated in the ground truth.

6 Conclusions

We introduced a novel approach for the automatic extraction of high-level activities in complex video scenes. Our method combines EMD matrix factorization and sparsity constraints, thus being robust to features' noise and producing as

output a set of sparse bases. This is greatly beneficial for complex scene analysis applications, where multiple activities simultaneously occur in the scene and it is of paramount importance to be able to extract the most relevant elementary activities for automatically inferring high-level behaviors. The proposed approach has been used to find recurrent activities in publicly available video datasets and has been extensively compared with state-of-the-art methods. The application of the proposed matrix factorization algorithm is not limited to video data. Indeed, we believe it will be suitable for many other problems, such as data analysis or human behavior understanding.

References

1. Wang, X., Ma, X., Grimson, W.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 539–555 (2008)
2. Hospedales, T., Gong, S., Xiang, T.: A markov clustering topic model for mining behaviour in video. In: ICCV (2009)
3. Kuettel, D., Breitenstein, M.D., Van Gool, L., Ferrari, V.: What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In: CVPR (2010)
4. Li, J., Gong, S., Xiang, T.: Learning behavioural context. Int. J. of Computer Vision (IJCV) 97, 276–304 (2012)
5. Zen, G., Ricci, E.: Earth mover's prototypes: a convex learning approach for discovering activity patterns in dynamic scenes. In: CVPR (2011)
6. Ricci, E., Zen, G., Sebe, N., Messelodi, S.: A prototype learning framework using EMD: Application to complex scenes analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012) (online)
7. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40, 99–121 (2000)
8. Varadarajan, J., Emonet, R., Odobez, J.M.: A sparsity constraint for topic models - application to temporal activity mining. In: NIPS, Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions (2010)
9. Haines, T., Xiang, T.: Video topic modelling with behavioural segmentation. In: ACM Workshop on Multimodal Pervasive Video Analysis (2010)
10. Zhao, B., Fei-Fei, L., Xing, E.: Online detection of unusual events in videos via dynamic sparse coding. In: CVPR (2011)
11. Lu, Z., Peng, Y.: Latent semantic learning by efficient sparse coding with hypergraph regularization. In: AAAI Conference on Artificial Intelligence (2011)
12. Matikainen, P., Hebert, M., Sukthankar, R.: Action recognition through the motion analysis of tracked features. In: ICCV Workshop on Video-oriented Object and Event Classification (2009)
13. Matikainen, P., Hebert, M., Sukthankar, R.: Representing Pairwise Spatial and Temporal Relations for Action Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 508–521. Springer, Heidelberg (2010)
14. Raptis, M., Soatto, S.: Tracklet Descriptors for Action Modeling and Video Analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 577–590. Springer, Heidelberg (2010)

15. Takahashi, M., Naemura, M., Fujii, M., Satoh, S.: Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. In: Machine Learning for Vision-based Motion Analysis (MLVMA), CVPR Workshop (2011)
16. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
17. Heiler, M., Schnorr, C.: Learning sparse representations by non-negative matrix factorization and sequential cone progamming. Journal of Machine Learning Research 7, 1385–1407 (2006)
18. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469 (2004)
19. Sandler, R., Lindenbaum, M.: Nonnegative matrix factorization with earth mover's distance metric. In: CVPR (2009)
20. Ling, H., Okada, K.: An efficient earth mover's distance algorithm for robust histogram comparison. IEEE Trans. on Pattern Analysis and Machine Intelligence 29, 840–843 (2006)
21. Shirdhonkar, S., Jacobs, D.W.: Approximate earth mover's distance in linear time. In: CVPR (2008)
22. Birchfield, S.: KLT: An implementation of the Kanade-Lucas-Tomasi feature tracker (2007)
23. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR (1999)
24. Tuy, H.: Convex programs with an additional reverse convex constraint. J. of Optim. Theory and Applic. 52, 463–486 (1987)
25. Li, J., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. In: BMVC (2008)