A Prototype Learning Framework using EMD: Application to Complex Scenes Analysis

Elisa Ricci, Gloria Zen, Nicu Sebe and Stefano Messelodi

Abstract—In the last decades, many efforts have been devoted to develop methods for automatic scene understanding in the context of video surveillance applications. This paper presents a novel non-object centric approach for complex scene analysis. Similarly to previous methods, we use low-level cues to individuate atomic activities and create clip histograms. Differently from recent works, the task of discovering high-level activity patterns is formulated as a convex prototype learning problem. This problem results into a simple linear program that can be solved efficiently with standard solvers. The main advantage of our approach is that, using as objective function the Earth Mover's Distance (EMD), the similarity among elementary activities is taken into account in the learning phase. To improve scalability we also consider some variants of EMD adopting L_1 as ground distance for one and two dimensional, linear and circular histograms. In these cases only the similarity between neighboring atomic activities, corresponding to adjacent histogram bins, is taken into account. Therefore we also propose an automatic strategy for sorting atomic activities. Experimental results on publicly available datasets show that our method compares favorably with state-of-the-art approaches, often outperforming them.

Index Terms—Video surveillance, Complex scene analysis, Earth Mover's Distance, Parametric Linear Programming.

1 Introduction

In the last few years the large deployment of distributed visual surveillance systems in public spaces has increased the demand for sophisticated tools performing the automatic analysis of long video streams. There is an increasing need for developing approaches which are able to extract typical and anomalous patterns in complex and crowded scenes. These scenarios are particularly challenging due to the presence of many occluded targets and to the need of complex models taking into account the spatial and temporal correlations between objects. Recently, unsupervised non-object centric approaches for dynamic scene understanding have gained popularity [6], [7], [43]. They have shown to be a reliable alternative to traditional visual surveillance approaches based on an object centric perspective [9], [10], *i.e.* relying on the classical detection/tracking scheme. These methods use low level features (*e.g.* position, size and motion of small blobs) to individuate elementary activities. Then by analyzing the co-occurrences of atomic activities, high level patterns are discovered.

The most recent and successful approaches for complex scene analysis are based on Probabilistic Topic Models (PTMs) [6], [7], [43]. These methods have shown to be very effective for discovering spatio-temporal patterns as well as for inferring behaviors'correlation over time and space. Their main limitation lies on the use of the standard word-document paradigm for representing atomic activity occurrences into clips. In this way the dependencies among atomic activities are not considered in the learning process.

To overcome this drawback in this paper we propose a different approach. We show that the problem of discovering high-level activity patterns in dynamic scenes can be modeled as a simple and convex optimization problem, *i.e.* a Linear Program (LP). At the core of our approach there is the idea that choosing as objective function a cross-bin distance, *i.e.* the EMD, rather than a bin-to-bin one, dependencies among atomic activities can be easily encoded in the learning process. To analyze long video sequences, in this paper we also consider some efficient variations of EMD which use L_1 norm and its variants for ground distance definition. In these cases, the flow network involved in the computation of the EMD is simplified and a words'order needs to be defined as only the similarity among adjacent bins is considered. To compute automatically the order of atomic activities a novel strategy based on simulated annealing is proposed. Interestingly, our approach permits to perform a multiscale analysis of the scene by varying a single parameter. We also show that anomalous patterns can be detected by comparing activity patterns at multiple scales and we propose a novel Multiscale Anomaly Score (MAS). Our approach is extensively evaluated on five datasets, four of which are publicly available.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 gives an overview of the proposed approach for extracting spatio-temporal patterns in complex scenes. In Sec-

- *Elisa Ricci is with the Department of Electrical and Information Engineering, University of Perugia, Perugia, 06125, Italy. E-mail: elisa.ricci@diei.unipg.it*
- *Gloria Zen and Nicu Sebe are with Department of Information Engineering and Computer Science, University of Trento, Trento, Italy.*
- *Stefano Messelodi is with Fondazione Bruno Kessler, Trento, Italy.*

tion 4 the Earth Mover's prototype learning algorithm is presented. Our approach for ordering atomic activities is also discussed. In Section 5 we show that, under some assumptions, all possible prototypes can be computed with improved efficiency exploiting the theory of Parametric LP. Results and conclusions are presented respectively in Sections 6 and 7.

2 Related Works

The approaches for complex scenes analysis without object tracking/detection have recently gained an increasing popularity [6], [7], [8], [13], [14]. Most of these methods adopt a probabilistic framework: a word-document paradigm is employed to represent the co-occurrences of atomic events and sophisticated PTMs are used to extract salient activities (topics). These approaches, specifically developed for unsupervised scene analysis, have several advantages over standard clustering techniques (*e.g. k-means*), such as a greater flexibility to model complex tasks and the ability to infer spatio-temporal dependencies among discovered activities. The approach we propose is significantly different from PTMs-based methods. In this paper the task of discovering high-level activity patterns is formulated as a Parametric LP. This permits not only to avoid the typical local minima problems but, more interestingly, to efficiently compute, under special conditions, the so-called regularization path associated to the LP. This means that we can explore the most k relevant activities for all possible values of k at roughly the same time as for one fixed value $k = \hat{k}$. In other words a multiscale video scene analysis arises naturally using our approach.

A large number of works in video analysis adopts a bag-of-words representation, not only in the context of complex scene analysis [15], [38] but also for related tasks such as human action recognition [2], [3]. This representation, while being very powerful, ignores the spatio-temporal arrangement of elementary features. Differently our approach explicitely focuses on exploiting atomic activity dependencies.

The most similar work to ours in the context of video scene understanding is perhaps [15]. In [15] a multiscale analysis is also proposed and diffusion maps are used in a preprocessing step before clustering. Differently our multi-resolution analysis is obtained during the clustering phase and it is also used for individuating unusual behaviors. Being able to detect anomalous patterns is of fundamental importance not only in visual surveillance applications [4], [5], but in many other contexts (see [16] for a review). Our MAS is related to previous nonparametric outlier mining techniques where the global and local density of the data are used to define the so-called outlier factors [17], [18]. However, MAS is novel since it is specifically tailored to the proposed clustering algorithm, aiming to quantify how the clusters size changes at subsequent scales. Previous approaches [4], [5] do not exploit multiscale segmentation levels for detecting unusual behaviours.

Our approach draws its inspiration from sparse signal approximation algorithms such as the fused lasso [19]. However, to the best of our knowledge we are the first to adopt a similar strategy for mining complex video scenes and to show that parametric LP can be a useful tool for multiscale analysis. To compute the entire solution path we resort on the approach described in [20]. However, our clustering algorithms are novel with respect to sparse signal approximation methods in [20]. In particular EMD has never been used in this context. This choice is motivated by the fact that with noisy histogram data the EMD is a better metric with respect to bin-to-bin distances.

Our work is related to [21] where EMD is used in the objective function of an optimization problem. However, in [21] the authors focused on Nonnegative Matrix Factorization. Finally recent clustering methods [23], [41], [42] are also closely related to our approach. In [41], [42] two algorithms for clustering with EMD are also presented, while in [23] the link between sensitivity analysis in LP and multiscale clustering is exploited. However, these works, not developed in the context of dynamic scene analysis, rely on optimization problems which are significantly different from ours.

3 Discovering Spatio-Temporal Patterns in Dynamic Scenes

This Section gives an overview of the proposed approach for discovering high-level activity patterns in dynamic scenes.

In the first phase (Fig.1) low level features are extracted from the video, *i.e.* for each pixels the foreground/background information and the optical flow are computed. As background subtraction algorithm we use a simple dynamic Gaussian-Mixture background model [24]. Then for each pixel of foreground we also compute the optical flow vector using the Lucas-Kanade algorithm. By thresholding the magnitude of the flow vector foreground pixels are divided into static and moving pixels. For moving pixels we also quantize the optical flow into $n_\theta = 8$ directions. Then we divide the scene into $p \times q$ patches. For each patch we build a patch descriptor vector $\boldsymbol{v} = [x\ y\ f_g\ \bar{d}_{of}\ \bar{m}_{of}]$ where (x, y) denotes the coordinates of the patch center in the image plane, f_g is the percentage of foreground pixels in the patch, $\bar{d}_{of}$ is the mode of the optical flow orientations distribution and $\bar{m}_{of}$ is the average magnitude of optical flow vectors with direction $\bar{d}_{of}$. For patches of static pixels we set $\bar{d}_{of} = \bar{m}_{of} = 0$. To limit the influence of noise in low level features extraction we discard patches with few

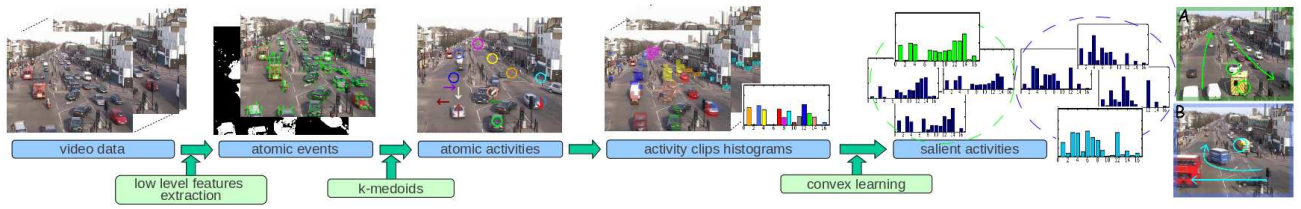


Fig. 1. Flowchart of the proposed approach

pixels of foreground, *i.e.* such that $f_g \leq T_{fg}$. We define an *atomic event* as a valid patch descriptor $\boldsymbol{v}$.

In the second phase a codebook of *atomic activities* is constructed. To this aim we define the following distance function between two atomic events $\boldsymbol{v}_q = [x^q\ y^q\ \bar{q}^q_{of}\ \bar{m}^q_{of}]$ and $\boldsymbol{v}_t = [x^t\ y^t\ \bar{d}^t_{of}\ \bar{m}^t_{of}]$ as

$$\delta_{qt} = \alpha\Delta p + (1-\alpha)(\Delta m + \Delta\theta) \tag{1}$$

where:

$$\Delta p = \sqrt{(x^t - x^q)^2 + (y^t - y^q)^2}$$
$$\Delta m = |\bar{m}^t_{of} - \bar{m}^q_{of}|$$
$$\Delta\theta = \begin{cases} 0 & \text{if } \bar{m}^q_{of} = 0 \vee \bar{m}^t_{of} = 0 \\ \min(|\bar{d}^t_{of} - \bar{d}^q_{of}|, n_\theta - |\bar{d}^t_{of} - \bar{d}^q_{of}|) & otherwise \end{cases}$$

In practice the parameter α in (1) controls the relative importance of position and motion information. In our experiments we set $\alpha = 0.5$. Then we group atomic events using K-medoids clustering. Each cluster represents an atomic activity. Subsequently we divide the video into short video clips and for each clip c we construct an *activity histogram* $\boldsymbol{h}_c$ representing the distribution of atomic activities. In the last phase the video clips are grouped according to their similarity. We propose a novel algorithm which, given a training set of clips histograms, outputs a small set of histograms constituting a synthetic representation of the original data. These histogram prototypes represent the *salient activities* occurring in the scene.

4 Earth Mover's Prototypes

In this Section we first review some basic concepts about EMD and its variations, then we present our Earth Mover's prototypes learning approach.

4.1 Earth Mover's Distance

The EMD [25] $\mathcal{D}_E(\boldsymbol{h}, \boldsymbol{p})$ between two histograms $\boldsymbol{h}, \boldsymbol{p}$ normalized to unit mass is obtained as the solution of the following transportation problem:

$$\min_{f_{qt}\geq 0} \sum_{t,q=1}^{D} d_{qt} f_{qt} \quad \text{s.t.} \quad \sum_{q=1}^{D} f_{qt} = h^t,\ \sum_{t=1}^{D} f_{qt} = p^q \tag{2}$$

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. Figure 2.a depicts the flow network associated to EMD. The problem (2) is a LP which can be solved efficiently due to the special structure of its sparse constraints [25], [26]. However, in the case of high dimensional histograms solving (2) can be very time consuming due to the large number of flow variables involved.

4.2 Linear, Circular and Thresholded EMD-L_1

Several methods have been proposed in the past to speed up the EMD distance computation. In [26], it is observed that, for histograms normalized to unit mass and L_1 ground distance (*i.e.* $d_{qt} = |q - t|$), every positive flow between faraway histograms bins can be replaced by a sequence of flows between neighbor bins. This implies that for *unidimensional histograms* (*i.e.* $\boldsymbol{h}, \boldsymbol{p} \in I\!R^D$), (2) can be simplified:

$$\begin{aligned} \min \quad & \sum_{q=1}^{D-1} f_{q,q+1} + \sum_{q=2}^{D} f_{q,q-1} \\ \text{s.t.} \quad & f_{q,q+1} - f_{q+1,q} + f_{q,q-1} - f_{q-1,q} = b^q\ \forall q, q = 1 \dots D \\ & f_{q,q+1}, f_{q,q-1} \geq 0 \end{aligned} \tag{3}$$

with $b^q = h^q - p^q$. The number of flow variables reduces from $O(D^2)$ in (2) to $O(D)$. 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. Moreover, the number of equality constraints is reduced by half and all the ground distances involved in the EMD-L_1 are ones. This is practically useful saving multiplications during computation. Eqn. (3) considers unidimensional histograms but the EMD-L_1 can be defined also for higher dimensional cases [39]. For example for *two-dimensional histograms* (*i.e.* $\boldsymbol{h}, \boldsymbol{p} \in I\!R^{D_1 \times D_2}, D_1 D_2 = D$) the only difference is that the neighborhood structure is not a line but a grid. The resulting optimization problem is:

$$\begin{aligned} & \min_{f_{m,n;q,t}\geq 0} \sum_{q,t} \sum_{m,n\in\mathcal{N}(q,t)} f_{q,t;m,n} \\ & \text{s.t.} \quad \sum_{m,n\in\mathcal{N}(q,t)} f_{q,t;m,n} - \sum_{m,n\in\mathcal{N}(q,t)} f_{m,n;q,t} = b^{q,t}\ \forall q, t \end{aligned} \tag{4}$$

where $b^{q,t} = h^{q,t} - p^{q,t}$, the indices q, t correspond to the position of a bin while its neighborhood $\mathcal{N}(q,t)$ is represented by the four adjacent bins (see Fig.2.c). In [27], [28] other computationally efficient variations of EMD have been proposed. In [28] the EMD with thresholded L_1 ground distance (*i.e.* $d_{qt} = \min(|q - t|, 2)$) is considered for robust comparison of noisy histograms. The adoption of the threshold implies the introduction of a transhipment vertex, slightly increasing the number of flow variables [28]. However, it has been shown that saturated distances are beneficial in terms of accuracy results in several applications. In [27] the same authors proposed a *circular histogram*

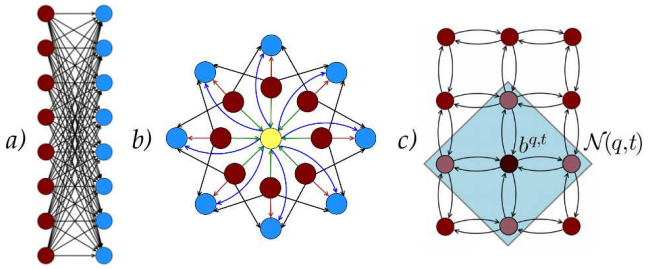


Fig. 2. The flow networks associated to **(a)** EMD, **(b)** EMD with thresholded L_1 ground distance for circular histograms. In **(b)** the yellow node is the transhipment vertex. Ingoing edge (green) cost is the threshold (2 in this case) and outgoing edge (blue) cost is 0. Red edges have cost 0. Black edges are 1-cost edges. **(c)** The two dimensional grid associated to (4).

representation. In this case a different ground distance is needed, *i.e.* $d_{qt} = \min(\min(|q-t|, D-|q-t|), 2)$. With thresholded ground distance and circular histograms, (2) assumes the form:

$$\begin{aligned} \min \quad & \sum_{q=1}^{D} f_{q,q+1} + \sum_{q=1}^{D} f_{q,q-1} + 2\sum_{q=1}^{D} f_{q,D+1} \\ \text{s.t.} \quad & f_{q,q+1} - f_{q+1,q} + f_{q,q-1} - f_{q-1,q} + f_{q,D+1} = h^q - p^q \\ & f_{q,q+1}, f_{q,q-1}, f_{q,D+1} \geq 0 \end{aligned} \tag{5}$$

where the flow variables $f_{q,D+1}$ correspond to the links connecting sources to the transhipment vertex. Figure 2.b depicts the associated flow network. In practice with respect to (3) in (5) also flows between sources and the transhipment vertex are considered. However, the number of flow variables is still $O(D)$.

4.3 Convex Optimization for Prototypes Learning

Given a set of histograms $\mathcal{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\}$, the task of prototype learning is the problem of computing a set $\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\}$, such that the following two requirements are jointly satisfied:

- each prototype $\boldsymbol{p}_i$ must be as much similar as possible to the associated histogram $\boldsymbol{h}_i$
- the set of prototypes is a sparse representation of the original dataset $\mathcal{H}$ (*i.e.* the number of different prototypes must be small)

The prototype learning problem can be formalized as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{h}_i, \boldsymbol{p}_i) + \lambda \sum_{i \neq j,\ i,j=1}^{N} \eta_{ij} \max_{q=1\ldots D} |p_i^q - p_j^q| \\ \text{s.t.} \quad & \boldsymbol{p}_i \geq 0, \qquad \sum_t p_i^t = 1 \quad \forall i = 1 \ldots N \end{aligned} \tag{6}$$

where the constraints ensure that the computed prototypes are histograms normalized to unit mass. The objective function consists of two terms. The loss function $\mathcal{L}(\cdot)$ penalizes the difference between the original histograms and the associated prototypes. In this paper we focus on the specific form of (6) when $\mathcal{L}(\cdot)$ is a convex function. The second term is meant to minimize the number of different prototypes. In fact the adoption of the $L_1 - L_\infty$ norm induces sparsity, thus producing a small number of prototypes. The set of binary coefficients $\eta_{ij} \in \{0, 1\}$ indicates the pairs of histograms which must be merged. In the absence of prior knowledge, for each histogram $\boldsymbol{h}_i$ a set of N_P nearest neighbors can be identified and the associated η_{ij} set to 1 if $\boldsymbol{h}_j$ is a neighbor of $\boldsymbol{h}_i$. In alternative temporal dependencies can be encoded into η_{ij}: for example if histograms represent temporally adjacent clips it is reasonable to set $\eta_{ij} = 1$ if $i = j - 1, j = 2 \ldots N$, $\eta_{ij} = 0$ otherwise. The relative importance of loss and regularization is controlled by the positive coefficient λ. When $\lambda = 0$ all prototypes $\boldsymbol{p}_i$ must be equal to their corresponding histograms $\boldsymbol{h}_i$ while for $\lambda \to \infty$ all prototypes should be equal to each others. For $0 \leq \lambda < \infty$ a number of prototypes k between N and 1 can be obtained. In truth, for large values of λ and few prototypes the L_1 norm also induces the prototypes to be quite similar to each other. In practice as λ decreases the effect of the loss function is stronger and the computed prototypes are quite different.

4.4 Learning Prototypes with EMD

In this paper we present a specific formulation of (6) where the EMD is adopted as loss function:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} \mathcal{D}_E(\boldsymbol{h}_i, \boldsymbol{p}_i) + \lambda \sum_{i \neq j} \eta_{ij} \max_{q=1\ldots D} |p_i^q - p_j^q| \\ \text{s.t.} \quad & \boldsymbol{p}_i \geq 0, \qquad \sum_t p_i^t = 1 \quad \forall i = 1 \ldots N \end{aligned} \tag{7}$$

Therefore to compute the prototypes we introduce (2) into (7) and we get the following LP:

$$\begin{aligned} \min_{p_i^q, f_{qt}^i, \zeta_{ij} \geq 0} \quad & \sum_{i=1}^{N} \sum_{t,q=1}^{D} d_{qt} f_{qt}^i + \lambda \sum_{i \neq j} \eta_{ij} \zeta_{ij} \\ \text{s.t.} \quad & -\zeta_{ij} \leq p_i^q - p_j^q \leq \zeta_{ij}, \ \forall q, \forall i, j \ i \neq j \\ & \sum_{q=1}^{D} f_{qt}^i = h_i^t, \ \forall t \quad \sum_{t=1}^{D} f_{qt}^i = p_i^q, \ \forall q, \forall i \end{aligned} \tag{8}$$

Note that the constraints $\sum_t p_i^t = 1$ are removed since they are automatically satisfied as the original histograms are normalized. It is worth noting that at the coordinate level we adopt the L_∞ norm rather than the L_1 norm. This does not promote sparsity but produces the effects that all coordinates of a prototype go to zero together and significantly reduces the computational cost of solving (8) limiting the number of slack variables ζ_{ij}.

Regarding the ground distance d_{qt} definition, we use the fact that each histogram bin corresponds to an atomic activity q, which is represented by the associated centroid $\boldsymbol{c}_q = [x^q \ y^q \ \bar{q}_{of}^q \ \bar{m}_{of}^q]$ computed by K-medoids in the first phase of our approach. Therefore we define the ground distance between two atomic activities $\boldsymbol{c}_q = [x^q \ y^q \ \bar{d}_{of}^q \ \bar{m}_{of}^q]$ and $\boldsymbol{c}_t = [x^t \ y^t \ \bar{d}_{of}^t \ \bar{m}_{of}^t]$ as follows:

$$d_{qt} = \alpha \Delta p + \beta(\Delta m + \Delta\theta) + (1 - \alpha - \beta)(1 - \Delta T_C) \tag{9}$$

where the terms $\Delta p, \Delta m$ and $\Delta\theta$ are defined as in (1). The last term ΔT_C takes into account the temporal correlation between atomic activities: starting from a training set of activity histograms $\{h_1, \ldots, h_{N_c}\}$, where N_c is a fixed number of clips, we consider, for each pair $\boldsymbol{c}_q, \boldsymbol{c}_t$ of atomic activities, the vectors $H_q = (h_1^q, \ldots, h_{N_c}^q)$ and $H_t = (h_1^t, \ldots, h_{N_c}^t)$ and set ΔT_C equal to the correlation coefficient between H_q and H_t. In (9) the ground distance depends on two parameters, α and β which control the relative importance of position, motion and temporal correlation.

4.5 Speeding up Prototype Learning

For large N and D solving (8) is still time consuming even for today's sophisticated LP solvers. The computational cost is especially high due to the large number of flow variables f_{qt}^i. Actually we do not specifically need them since we are only interested in computing the prototypes $\boldsymbol{p}_i$. Therefore to speed up calculations we also propose to modify (8) as follows.

We consider the special case of EMD with L_1 distance over bins as ground distance. In our specific application the idea is that similar atomic activities should correspond to neighboring bins in activity histograms. To this aim the atomic activities are sorted according to the associated location and motion information (see subsection 4.6). With this premises, we propose to simplify (8) using (3). So substituting the definition of EMD-L_1 (3) into (7) we get:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N}\left(\sum_{q=1}^{D-1} f_{q,q+1}^i + \sum_{q=2}^{D} f_{q,q-1}^i\right) + \lambda \sum_{i\neq j} \eta_{ij}\zeta_{ij} \qquad (10) \\ \text{s.t.} \quad & -\zeta_{ij} \leq p_i^q - p_j^q \leq \zeta_{ij},\ \forall q,\ \forall i,j,\ i \neq j \\ & f_{q,q+1}^i - f_{q+1,q}^i + f_{q,q-1}^i - f_{q-1,q}^i = h_i^q - p_i^q,\ \forall q,\ \forall i \\ & p_i^q, f_{q,q+1}^i, f_{q,q-1}^i, \zeta_{ij} \geq 0 \end{aligned}$$

The resulting optimization problem is a LP with $n_{var} = n_f + n_p + n_\zeta = 2N(D-1) + ND + \frac{1}{2}NN_P$ variables if we adopt the nearest neighbor approach for setting the coefficients $\eta_{ij} = 1$. In this case for large datasets and small histograms ($N \gg D$) the computational cost of (10) is dominated by the number of slack variables. However, by considering a small number of neighbors N_P, (10) can be solved efficiently even for large datasets. Analogously a prototype learning approach can be devised for two dimensional histograms by considering the EMD-L_1 definition (4). Similarly for circular histograms and EMD with thresholded L_1 ground distance, the prototype learning algorithm can be obtained by inserting (5) in (7):

$$\begin{aligned} \min \quad & \sum_{i,q} f_{q,q+1}^i + \sum_{i,q} f_{q,q-1}^i + 2\sum_{i,q} f_{q,D+1}^i + \lambda \sum_{i\neq j} \eta_{ij}\zeta_{ij} \\ \text{s.t.} \quad & -\zeta_{ij} \leq p_i^q - p_j^q \leq \zeta_{ij},\ \forall q,\ \forall i,j,\ i \neq j \qquad (11) \\ & f_{q,q+1}^i - f_{q+1,q}^i + f_{q,q-1}^i - f_{q-1,q}^i + f_{q,D+1}^i = h_i^q - p_i^q \\ & p_i^q, f_{q,q+1}^i, f_{q,q-1}^i, f_{q,D+1}^i, \zeta_{ij} \geq 0 \end{aligned}$$

The resulting optimization problem is a LP with $n_{var} = 4ND + \frac{1}{2}NN_P$.

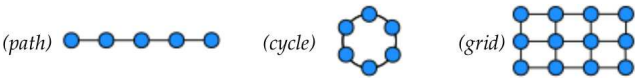


Fig. 3. Structures used to arrange atomic activities.

4.6 Ordering Atomic Activities

Elementary activities are not independent and it is desirable to take into account their similarity when learning activity prototypes. A straightforward way to impose this is to encode atomic activities similarity in the ground distance definition (9). This means considering similarity among all possible pairs of atomic activities and a high computational cost of solving (8) even for problems with a small N. A similar requirement can be imposed also in the case of the more efficient EMD variants based on L_1. In this case considering atomic activities similarity means sorting them according to a prespecified criterion. The idea is that, when constructing clip histograms, neighboring activities correspond to similar ones.

To this aim we propose to find the best arrangement of the atomic activities into appropriate graph structures in order to minimize the distortion between the ground distances d_{qt} and the distances $\mathcal{D}_g$ of the nodes q and t within the graph (*i.e.* the length of the shortest path connecting them). As discussed at the beginning of this section, in this work we consider the three following graph structures: *path* graph, *cycle* graph and *square grid* graph (corresponding respectively to 1D, circular and 2D histograms, see Fig.3), where the number of nodes is equal to the number of atomic activities. The distortion is defined as follows:

$$\sum_{q=1}^{D}\sum_{t=q+1}^{D} \left(d_{qt} - \mathcal{D}_g(\sigma(\boldsymbol{c}_q) - \sigma(\boldsymbol{c}_t))\right)^2 \qquad (12)$$

which has to be minimized with respect to $\sigma()$, a one-to-one function mapping atomic activities to nodes of the graph. The minimization is achieved by Algorithm 1 which implements a simulated annealing approach. The temperature T_0 is set to a value such that a given fraction (about 0.75) of the moves would be initially accepted. The values of N_{iter} and η used in the experiments are 10000 and 0.99, respectively.

4.7 Learning Prototypes with bin-to-bin Distances

To demonstrate the advantages of considering cross-bin similarities when learning prototypes, we briefly discuss the form that (6) assumes when bin-to-bin distances are used as metrics and some related approaches in the literature. For example when the L_1 norm is chosen as loss function, (6) assumes the form:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N}\sum_{q=1}^{D} |h_i^q - p_i^q| + \lambda \sum_{i\neq j} \eta_{ij} \max_{q=1\ldots D} |p_i^q - p_j^q| \qquad (13) \\ \text{s.t.} \quad & \boldsymbol{p}_i \geq 0, \qquad \sum_t p_i^t = 1 \quad \forall i = 1 \ldots N \end{aligned}$$

The resulting optimization is still a LP (as in the case of EMD) and can be solved efficiently with standard

Algorithm 1 Sorting atomic activities

Input: atomic activities $\boldsymbol{C} = \{\boldsymbol{c}_1, \dots \boldsymbol{c}_D\}$, graph $G(V, E)$, with $V = \{v_1, \dots, v_D\}$
$T \leftarrow T_0$ *initialize temperature*
$\sigma(\boldsymbol{c}_i) \leftarrow v_i, i = 1 \dots D$ *initialize* $\sigma()$
$\mathcal{D}_0 \leftarrow$ initial distortion *equation (12)*
$M \leftarrow 1$ *counter of accepted moves*
while $M > 0$
  $M \leftarrow 0$ *reset the counter*
  repeat N_{iter} times *generate move hypothesis*
    $\boldsymbol{c}_k \leftarrow$ randomly selected atomic activity
    $v_j \leftarrow$ randomly selected node from $V \setminus \{\sigma(\boldsymbol{c}_k)\}$
    $\boldsymbol{c}_i \leftarrow \sigma^{-1}(v_j)$
    $\sigma(\boldsymbol{c}_i) \leftarrow \sigma(\boldsymbol{c}_k)$
    $\sigma(\boldsymbol{c}_k) \leftarrow v_j$
    $\mathcal{D}_n \leftarrow$ compute distortion
    $\Delta\mathcal{D} \leftarrow \mathcal{D}_n - \mathcal{D}_0$;
    Accept move with probability $\min(e^{-\Delta\mathcal{D}/T}, 1)$
    if move accepted **then**
      $M \leftarrow M + 1$; $\mathcal{D}_0 \leftarrow \mathcal{D}_n$
  $T \leftarrow \eta * T$ *decrease the temperature* ($\eta < 1$)
end
Output: function $\sigma()$

solvers once slack variables have been introduced. The proposed approach (6) can also be used with Kullback-Leibler distance as loss functions:

$$\min \quad \sum_{i=1}^{N}\sum_{q=1}^{D} h_i^q \log \frac{h_i^q}{p_i^q} + \lambda \sum_{i \neq j} \eta_{ij} \max_{q=1\dots D} |p_i^q - p_j^q|$$

Similarly to L_1 and KL, also the L_2 norm can be used in the loss function in (6). In particular, if a sum of L_1 norms rather than a combination of L_1-L_∞ is used as regularization term and no constraints are imposed on the prototypes $\boldsymbol{p}_i$, the following optimization problem is obtained:

$$\min \quad \sum_{i=1}^{N} ||\boldsymbol{h}_i - \boldsymbol{p}_i||^2 + \lambda \sum_{i \neq j} \eta_{ij} \sum_{q} |p_i^q - p_j^q| \qquad (14)$$

The special case where $\eta_{ij} = 1$ if $i = j - 1$ and $\eta_{ij} = 0$ otherwise leads to the well known "total variation denoising" procedure [32] or to a special case of the fused lasso [19]. However, it is worth nothing that in our case the choice of using a L_1-L_∞ norm rather than a sum of L_1 is motivated by computational efficiency reasons. In fact since our optimization problem is an LP and we solve it with standard solvers, the number of slack variables is kept limited. In all these cases, only bin-to-bin comparisons are allowed. Indeed the experimental results presented in the Section 6 demonstrate that bin-to-bin distances are less effective than EMD when learning prototypes for dynamic scene understanding.

4.8 Multiscale Anomaly Score

A crucial property of (6) is that the sparsity achieved is controlled by a single parameter, *i.e.* the regularization constant λ. In other words, for λ varying between ∞ and 0, a different number of prototypes between 1 and N can be obtained. In the case of automatic scene understanding, this corresponds to discover different salient activities at multiple scales. For example, for traffic scene analysis, for large values of λ we can obtain a very rough description of the activities differentiating among clips with moving vehicles or clips corresponding to vehicles stopped at the traffic lights. As λ decreases we gradually enhance the level of details of the analysis differentiating among vehicles flows of different intensity.

Instead of finding the value of λ which provides the optimal prototypes we propose to exploit the solutions of (6) for different values of λ. More formally, given a set of N histograms $\boldsymbol{h}_i$ we first introduce the following characterization of sets of fused histograms as they are generated by our algorithms.

Definition 1. (Sets of Fused Histograms) *Let $\lambda = \bar{\lambda}$ and $\mathcal{H}_\ell^{\bar{\lambda}}$ be a set of histograms with $\ell = 1, \dots, N(\bar{\lambda})$ where $N(\bar{\lambda})$ is the number of different prototypes obtained for $\lambda = \bar{\lambda}$. Then a valid set of fused histograms $\mathcal{H}_\ell^{\bar{\lambda}}$ satisfies the following properties:*

- $\cup_{\ell=1}^{N(\bar{\lambda})} \mathcal{H}_\ell^{\bar{\lambda}} = \mathcal{H}$
- $\mathcal{H}_\ell^{\bar{\lambda}} \cap \mathcal{H}_m^{\bar{\lambda}} = \emptyset, \forall \ell \neq m.$
- $\forall\ \boldsymbol{h}_\ell, \boldsymbol{h}_m \in \mathcal{H}_k^{\bar{\lambda}}$ *we have* $p_\ell^q = p_m^q\ \forall\ q = 1 \dots D$
- $\forall\ \boldsymbol{h}_\ell \in \mathcal{H}_\ell^{\bar{\lambda}}$ *and* $\boldsymbol{h}_m \in \mathcal{H}_m^{\bar{\lambda}}\ \exists q: \quad p_\ell^q \neq p_m^q$

In a nutshell a set of fused histograms corresponds to histograms associated to the same prototype. Different sets of histograms are generated for different values of λ. Comparing clustering results at multiple scales (*i.e.* comparing sets of fused histograms for different values of λ) we can detect unusual behaviors corresponding to atypical histograms. To this aim we define for each $\boldsymbol{h}_\ell$ an associated anomaly score. The general idea behind this score is to monitor how the clusters size changes for decreasing values of λ. From $\lambda = \infty$ (where all the histograms are represented by a single prototype) to $\lambda = 0$ where each histogram corresponds to a different prototype, the anomaly score of $\boldsymbol{h}_k$ can be computed as the sum of the ratios of the size of the clusters containing $\boldsymbol{h}_\ell$ at two subsequent scales. Analyzing multiple levels we can distinguish between cases where a cluster with a single histogram is merged at higher level with a small cluster and situations where it belongs to a big cluster: in the first case its anomaly score is higher. Formally:

Definition 2. (MAS) *Let $\boldsymbol{h}_\ell \in \mathcal{H}_\ell^{\lambda_i}$ and $\boldsymbol{h}_\ell \in \mathcal{H}_{\ell'}^{\lambda_{i-1}}$ with $\lambda_{i-1} > \lambda_i$. We define the* **Multiscale Anomaly Score (MAS)** *of the histogram $\boldsymbol{h}_\ell$ as:*

$$MAS = 1 - \frac{1}{NL} \sum_{i=2}^{L} \frac{|\mathcal{H}_{\ell'}^{\lambda_{i-1}}|}{|\mathcal{H}_\ell^{\lambda_i}|}$$

In practice the most anomalous clips tend to get a higher MAS. Let us consider the case of a cluster made by a single clip. In this case the ratio in the MAS definition is very low (actually zero) until the clip is merged into a large cluster. The later it is merged, the smaller the ratio value is, thus the higher the MAS is.

Note that while large values of L may lead to more accurate estimates of MAS, this also increases the computational cost since (8), (10) and (11) must be solved L times. However, in the following we show how in the special case of temporal segmentation a multiscale analysis can be obtained with computational cost comparable with that of solving (8), (10) or (11) for a single value of λ. As a final remark we should say that we experimentally observed that if two histograms are fused for a certain value of $\lambda = \bar{\lambda}$ (*i.e.* they belong to the same fused set) they will not necessarily remain fused for any $\lambda \geq \bar{\lambda}$. However, we found that for moderately large values of L, this does not decrease the accuracy of MAS analysis.

5 Multiscale Analysis in One Shot

In this Section we focus our attention on linear histograms and on the temporal segmentation approach *i.e.* we consider $\eta_{ij} = 1$ for $i = j-1, j = 2 \dots N$ and $\eta_{ij} = 0$ otherwise. In particular we consider (8) and (13). We show that since (8) and (13) are parametric LP, an algorithm based on a variant of the revised simplex method can be developed to compute all possible sets of histogram prototypes for increasing values of λ.

5.1 Preliminaries: LP and Parametric LP

Given a matrix $\mathbf{A} \in I\!R^{n \times m}$ and the vectors $\mathbf{c} \in I\!R^m$, $\mathbf{b} \in I\!R^n$ a LP in standard form [29], [30] is given by:

$$\min_{\mathbf{x} \geq 0} \quad \mathbf{c}'\mathbf{x} \qquad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{15}$$

If the matrix $\mathbf{A}$ is of full rank n and the polyhedron $\mathcal{P} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$ is bounded and non-empty, the LP has a bounded optimal solution. Let $\mathcal{B} \in \mathcal{I} = \{1, \dots, m\}$ be an ordered set of n column indexes. Let $\mathbf{A}_\mathcal{B}$ be the $n \times n$ sub-matrix of $\mathbf{A}$ whose i-th column is $\mathbf{A}_i$. The set $\mathcal{B}$ is called a *feasible basis* if $\mathbf{A}_\mathcal{B}$ is of full-rank and $\mathbf{A}_\mathcal{B}^{-1}\mathbf{b} \geq 0$. Since $\mathbf{A}$ is of full rank and the linear program is feasible, a feasible basis always exists. A column $\mathbf{A}_i$ with $i \in \mathcal{B}$ is called a basic column, otherwise it is called a non-basic column and belongs to the set $\mathcal{N} = \mathcal{I} - \mathcal{B}$. A *basic feasible solution* (bfs) $\mathbf{x}_\mathcal{B}$ of the LP corresponding to a feasible basis $\mathcal{B}$ is obtained by $\mathbf{x}_\mathcal{B} = \mathbf{A}_\mathcal{B}^{-1}\mathbf{b}$ and $\mathbf{x}_\mathcal{N} = \mathbf{0}$. A bfs is *optimal* if it corresponds to a solution of the LP. There is a bijection between bfs and vertices of $\mathcal{P}$. The *simplex* method systematically explores the extreme points (bfs) of $\mathcal{P}$, *i.e.* starting from an initial extreme point, until an optimal extreme point is found.

A parametric LP problem has the form:

$$\min_{\mathbf{x} \geq 0} \quad (\mathbf{c} + \lambda\mathbf{a})'\mathbf{x} \qquad \text{s.t.} \ \mathbf{A}\mathbf{x} = \mathbf{b} \tag{16}$$

with $\mathbf{a} \in I\!R^m$ and $\lambda \in I\!R$. In [20] Yao and Lee showed that many algorithms in machine learning and specifically the family of regularization problems with piecewise linear loss and L_1 penalties (such as L_1 SVM) can be written in the form of (16) and a variant of the simplex method can be used for solving (16) for all possible values of λ simultaneously.

5.2 Multiscale Analysis

Let $\boldsymbol{p}, \boldsymbol{\delta}_+, \boldsymbol{\delta}_- \in I\!R^{ND}$, $\boldsymbol{\zeta} \in I\!R^{N-1}$ and $\mathbf{H} \in I\!R^{ND}$ be the vector obtained concatenating the histograms in the training set (*i.e.* $\mathbf{H}' = (\boldsymbol{h}'_1, \dots, \boldsymbol{h}'_N)$). We first define the following matrices: the block diagonal matrix $\mathbf{D} \in I\!R^{(N-1)D \times (N-1)}$, $\mathbf{D} = \text{diag}(\mathbf{-1})$ and $\mathbf{-1} \in I\!R^D$ and the block Toeplitz matrix $\mathbf{\Sigma} \in I\!R^{(N-1)D \times ND}$,

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{I} & \mathbf{-I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{-I} & \dots & \mathbf{0} \\ \vdots & \ddots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} & \mathbf{-I} \end{pmatrix},$$

with $\mathbf{I}$, $\mathbf{0}$ and $-\mathbf{I} \in I\!R^{D \times D}$.

Proposition 1. *The following elements:*

$$\mathbf{x} = \begin{pmatrix} \boldsymbol{f}' & \boldsymbol{\zeta}' & \boldsymbol{p}' & \boldsymbol{\delta}_+{}' & \boldsymbol{\delta}_-{}' \end{pmatrix}$$
$$\mathbf{a}' = \begin{pmatrix} \boldsymbol{\omega}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \end{pmatrix}$$
$$\mathbf{c}' = \begin{pmatrix} \mathbf{0}' & \mathbf{1}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' \end{pmatrix}$$
$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{D} & \mathbf{\Sigma} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & -\mathbf{\Sigma} & \mathbf{0} & \mathbf{I} \\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{G} & \mathbf{0} & \mathbf{-I} & \mathbf{0} & \mathbf{0} \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{H} \\ \mathbf{0} \end{pmatrix}$$

with $\boldsymbol{f} \in I\!R^{ND^2}$, $\boldsymbol{\omega} \in I\!R^{ND^2}$, $\boldsymbol{\omega} = (\mathbf{d} \dots \mathbf{d})$, $\mathbf{d} \in I\!R^{D^2}$, $\mathbf{d} = (d_{11}, \dots d_{1D} d_{21} \dots d_{DD})$, $\mathbf{F}, \mathbf{G} \in I\!R^{ND \times ND^2}$ being two block diagonal matrices, $\mathbf{F} = \text{diag}(\mathbf{Q}), \mathbf{G} = \text{diag}(\mathbf{T}), \mathbf{Q}, \mathbf{T} \in I\!R^{D \times D^2}$, $\mathbf{Q} = \text{diag}(\mathbf{1}')$, $\mathbf{1}' \in I\!R^D$

$$\mathbf{T} = \begin{pmatrix} \mathbf{e'}_1 & \mathbf{e'}_1 & \dots & \mathbf{e'}_1 \\ \mathbf{e'}_2 & \mathbf{e'}_2 & \dots & \mathbf{e'}_2 \\ \vdots & \ddots & & \vdots \\ \mathbf{e'}_D & \mathbf{e'}_D & \dots & \mathbf{e'}_D \end{pmatrix}$$

with $\mathbf{e'}_i \in I\!R^D$ *is a vector of all 0 and 1 in the i-th position, define (8) in the standard form (16) of a parametric LP.*

Given a parametric LP problem in standard form all possible solutions $\bar{\mathbf{x}}$ for different values of λ can be computed. For this purpose in this paper we use a variation of the algorithm proposed in [20] by considering a different variant of the simplex methods rather than the tableau simplex *i.e.* the revised simplex method with the lexico-min rule since it offers computational advantages for sparse LPs and avoid situations of degeneracy. According to this, the basic column to exit the current basis $\mathcal{B}$ is selected according to the lexico-min rule: the column which exits the basis is $\mathbf{A}_\ell$, where ℓ is the index of the lexicographically smallest row $\mathbf{A}^i/u_i$, $u_i > 0$, $\mathbf{u} = \mathbf{A}_\mathcal{B}^{-1}\mathbf{A}_j$ and $\mathbf{A}^i$ denotes the i-th row of $\mathbf{A}_\mathcal{B}$. The index ℓ always exists, since otherwise $u_i \leq 0$ for all i and the problem is unbounded. The resulting algorithm is presented in Algorithm 2.

The main difference and the main issue when running Algorithm 2 is how to individuate an optimal bfs $\mathcal{B}_0$. This can be obtained using any feasible basic index set $\bar{\mathcal{B}}_0$ and running the standard simplex algorithm for the associated LP problem *i.e.* for $\mathbf{a} = \mathbf{0}$. The following

Algorithm 2 Mutiscale analysis in one-shot

Input: $\mathbf{H} = (\boldsymbol{h}'_1, \ldots, \boldsymbol{h}'_N)'$, $i = 0$.
Set (8) (or (13)) in standard form (16) according to Proposition 1 (Proposition 3).
Find an optimal bfs $\mathcal{B}_0$ for $\lambda_0 = \infty$ following Proposition 2 (Proposition 4).
while $\lambda_i \geq 0$
  Compute $\mathbf{x}^i$, with $\mathbf{x}^i_{\mathcal{B}_i} = \mathbf{A}^{-1}_{\mathcal{B}_i}\mathbf{b}$ and $\mathbf{x}^i_{\mathcal{N}_i} = \mathbf{0}$.
  $\bar{c}_j = c_j - \mathbf{c}_{\mathcal{B}_i}\mathbf{A}^{-1}_{\mathcal{B}_i}\mathbf{A}_j$ with $j \in \mathcal{N}_i$.
  $\bar{a}_j = a_j - \mathbf{a}_{\mathcal{B}_i}\mathbf{A}^{-1}_{\mathcal{B}_i}\mathbf{A}_j$ with $j \in \mathcal{N}_i$.
  $m = \arg\max_j\{-\frac{\bar{c}_j}{\bar{a}_j} : \bar{a}_j > 0\}$ (*entry index*)
  $\lambda_{i+1} = -\frac{\bar{c}_m}{\bar{a}_m}$
  $\mathbf{u} = \mathbf{A}^{-1}_{\mathcal{B}_i}\mathbf{A}_m$.
  if the support $I(\mathbf{u})$ is empty **then**
    return *problem is unbounded*
  $\ell = \arg$ lexico- $\min_t\{\frac{\mathbf{A}^i_t}{u_t} : t \in I(\mathbf{u})\}$ (*exit index*)
  Update $\mathcal{B}_{i+1} = \mathcal{B}_i \cup \{m\}\backslash\{\ell\}$
  Create the set $\mathcal{P}_i = \{\boldsymbol{p}^i_1, \ldots, \boldsymbol{p}^i_N\}$ extracting the corresponding coordinates from $\mathbf{x}^i$.
  $i \leftarrow i + 1$.
end
Output: The sets of prototypes $\mathcal{P}_1, \mathcal{P}_2, \ldots \mathcal{P}_{N_{prot}}$

proposition shows how a basic feasible set $\bar{\mathcal{B}}_0$ can be individuated for the proposed problem (8).

Proposition 2. *The set of indices* $\bar{\mathcal{B}}_0 = \mathcal{I}_1 \cup \mathcal{I}_2$ *with* $\mathcal{I}_1 = \{kD + 1 : k = 0, \ldots, ND - 1/D\}$, $\mathcal{I}_2 = \{ND^2 + N + k : k = 0, \ldots, 3ND - 1\}$ *individuates a bfs for (8).*

Proof. See Appendix A.

Algorithm 2 can generally be applied not only to (8) but also to (10), (11), (13) provided that a suitable bfs is found. Due to lack of space in the following we only show the results associated to (13).

Proposition 3. *The following elements:*

$$\mathbf{x}' = \begin{pmatrix}\boldsymbol{p}' & \boldsymbol{\xi}' & \boldsymbol{\zeta}' & \boldsymbol{\delta}_+{}' & \boldsymbol{\delta}_-{}' & \boldsymbol{\theta}_+{}' & \boldsymbol{\theta}_-{}'\end{pmatrix}'$$

$$\mathbf{a}' = \begin{pmatrix}\mathbf{0}' & \mathbf{1}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}'\end{pmatrix}'$$

$$\mathbf{c}' = \begin{pmatrix}\mathbf{0}' & \mathbf{0}' & \mathbf{1}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}'\end{pmatrix}'$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{-I} & \mathbf{-I} & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{-I} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{\Sigma} & \mathbf{0} & \mathbf{D} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{-\Sigma} & \mathbf{0} & \mathbf{D} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{E} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} \mathbf{H} \\ \mathbf{-H} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

with the block diagonal matrix $\mathbf{E} \in I\!R^{N \times ND}$, $\mathbf{E} = \mathrm{diag}(\mathbf{1})$, $\boldsymbol{\xi}$, $\boldsymbol{\theta}_+, \boldsymbol{\theta}_- \in I\!R^{(N-1)D}$ and $\mathbf{1} \in I\!R^D$ *define (13) in the standard form (16) of a parametric LP.*

Proposition 4. *The set of indices* $\bar{\mathcal{B}}_0 = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_4 \cup \mathcal{I}_5$ *with* $\mathcal{I}_1 = \{kD+1 : k = 0, \ldots, N-1\}$, $\mathcal{I}_2 = \{ND+k : k = 1, 2, \ldots, ND\}$, $\mathcal{I}_3 = \{2ND + N - 1 + kD + 1 : k = 0, 1, \ldots, N-1\}$, $\mathcal{I}_4 = \{3ND + N - 1 + k : k = 1, 2 \ldots, ND\}\backslash\{3ND+N-1+kD+1 : k = 0, 1 \ldots, N-1\}$, $\mathcal{I}_5 = \{4ND + N - 1 + k : k = 1, 2, \ldots, 2(N-1)D\}$, *individuates a bfs for (13).*

Proof. See Appendix B.

As a final remark we should note that in general even when the coefficients η_{ij} assume different values that in the case of temporal segmentation, (8) and (13) are also parametric LP problems and Algorithm 2 can be used for computing the entire solution path. However, in this cases (*e.g.* for nearest neighbor clustering) determining a suitable bfs $\mathcal{B}_0$ is more complex and we leave it to future works.

6 Experimental Results

6.1 Datasets and Experimental Setup

Experiments were conducted on five datasets, four of which are publicly available. The first dataset consists of a **Traffic** scene sequence. As the vehicles flow is controlled by traffic lights, different events occur at regular periods. The second video sequence depicts a **basketball match** and is taken from the APIDIS[1] website. The images are cropped to include only the basketball court and resized. The last three datasets **Junction**, **Roundabout**, **Junction2** are also available[2] (for the first two sequences, ground truth for two levels temporal segmentation is available; for the third one, we manually annotated a sequence of 80 clips at 2 and 3 levels, based on the traffic lights' changes). The videos depict some traffic scenes in London and have been extensively used in previous works [6], [7], [31], [38].

In this section, we first show temporal segmentation results obtained with EMD-L_1-*linear* (10); the other experiments are meant to test the proposed approach for nearest neighbor clustering. In the first case, *temporal segmentation* is obtained setting in (10) $\eta_{ij} = 1$ if $i = j-1$ and $\eta_{ij} = 0$ elsewhere; in the case of *clustering*, the nearest neighbor graph for prototype learning is computed based on histograms similarity, using EMD with L_1 ground distance. In all the experiments we found that $N_P = 3$ or $N_P = 4$ correspond to the best performance. A discussion about how to choose the values of α and β is reported in subsection 6.3.3. The value of λ changes in all the different experiments according to the required number of clusters. While for temporal segmentation Algorithm 2 can be used to obtain all possible prototypes at varying λ, for nearest neighbor clustering is necessary to test several λ to get the required number of clusters. More details about the datasets and our experimental setup are summarized in Table 1. The proposed algorithms are listed in Table 2 and are fully implemented in C++ using the publicly available libraries OpenCV for video processing and feature extraction and GLPK 4.2.1 (GNU Linear Programming Kit) as the backend linear programming solver. The code[3] for solving problems (8), (10), (11) and (13) and the video[4] showing our results are available online.

1. http://www.apidis.org/Dataset/
2. http://www.eecs.qmul.ac.uk/~jianli/Dataset_List.html
3. http://disi.unitn.it/~zen
4. http://disi.unitn.it/~zen/demo_emp.html

TABLE 1
Details on datasets and experimental setup

	n^o frames	fps	n^o clips	frame size	patch size	D	clip length [s]
Traffic	6000	12	300	276×336	23×21	8	12
Basket	6000	23	100	320×368	16×16	16	3
Junction	90000	25	300	288×360	12×12	16	12
Roundabout	93500	25	311	288×360	12×12	16	12
Junction2	78000	25	312	288×360	12×12	16,24,30	10

TABLE 2
Proposed approaches tested in our experiments.

	L_1	EMD-L_1-*lin.*	EMD-L_1-*circ.*	EMD-L_1-*2D*
Equation	(13)	(10)	(11)	deducible from (10)

6.2 Temporal Segmentation

We demonstrate the effectiveness of the proposed temporal segmentation approach on the **Traffic** dataset. This sequence despite being short is interesting, as it corresponds to few cycles of the traffic lights status and it contains some interesting anomalous events. When applying temporal segmentation only the similarity among adjacent clips is considered. Therefore several clusters/prototypes correspond to the same patterns (*i.e.* green traffic light). These clusters must be merged manually after learning. This supplementary phase may result annoying when dealing with long sequences (*e.g.* Junction, Roundabout). In these cases nearest neighbor clustering is preferred. For this reason we evaluate the performance of temporal segmentation results in term of correctly individuated breakpoints while for nearest neighbor clustering the accuracy is computed considering the percentage of correctly labeled clips. In the Traffic scene two main traffic flow patterns are distinguished: (i) two parallel flows when the traffic light is on green and (ii) vehicles stopping and forming a queue in the lane on the left when the traffic light is on red. Rare events also occur such as pedestrians crossing the street outside zebra crossing or vehicles making not allowed U-turns. Figure 4 shows the multi-scale segmentation results on 100 clips obtained by solving EMD-L_1-*linear* (10) for different values of λ.. The temporal segmentation results with 10 clusters, obtained with $\lambda = 5$, is highlighted with a red frame. From each of the 10 clusters obtained we extract one frame, representative for the salient activities. As expected, clips with similar activity histograms are associated to the same cluster. Interestingly we successfully detect the changes in vehicles flow triggered by the traffic lights. As shown in Fig. 4, the orange, yellow, red and blue clusters correspond to the activity of parallel vehicle flows (green traffic light), while the light blue, white and cyan clusters are associated to stationary vehicles (red traffic light). The green, violet and pink clusters are still associated to red traffic lights and, in particular, they represent the phase when the traffic queue begins, hence the traffic flow is characterized by low density.

It is interesting to analyze the way clusters merge as λ increases. For example, the clusters associated to the same traffic light status but with different traffic density (*i.e.* pink and cyan, green and light blue) merge at the superior level. A visual inspection confirms that the segmentation results obtained with EMD distance are consistent with the human annotation (Fig.4.b). We manually annotated it. A quantitative comparison of the proposed methods (8) and (10) and bin-to-bin approaches (Fused lasso [19] and (13)) for the entire **Traffic** sequence is shown in Table 3. The performance is measured in terms of percentage of break points correctly individuated. The results clearly demonstrate that bin-to-bin distances are less powerful as they do not take into account similarity among atomic activities. It is worth noting that (10) can be considered as a good approximation of (8). An important observation concerns the computational cost of our multiscale analysis. As (8) is a parametric LP, *all* solutions (*i.e. all* possible prototypes) can be found with a slightly increased computational cost with respect to computing just one solution (corresponding to a fixed value of λ). Therefore, the speedup is huge. For example all possible prototypes associated to 100 clips can be computed in approximately 5 min whilst the solution for a single value of λ takes about 1 min.

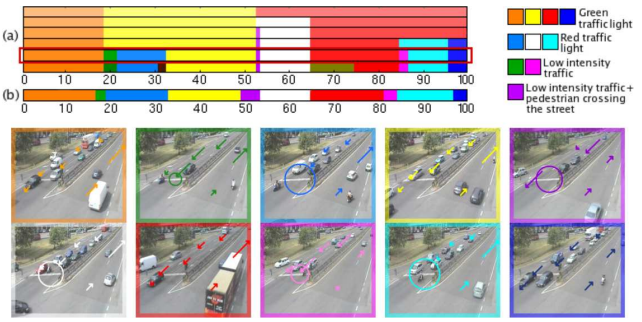


Fig. 4. Traffic dataset. (a) Temporal segmentation results obtained varying λ with EMD-L_1(10). (b) Ground-truth and (top, right) corresponding legend. (Bottom) salient activities automatically extracted from the segmentation result highlighted in red.

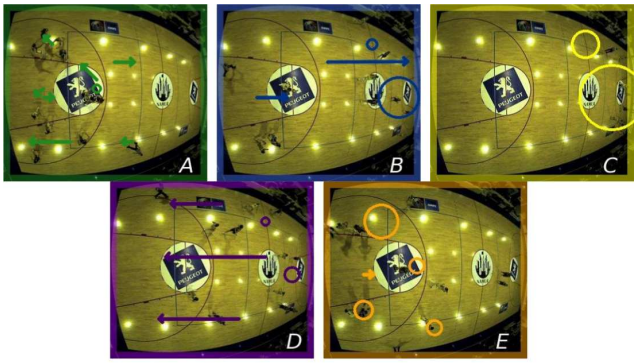
TABLE 3
Traffic dataset: temporal segmentation accuracy

EMD (8)	EMD-L_1-*linear*(10)	L_1(13)	Fused Lasso
83.2	82.4	72.5	68.7

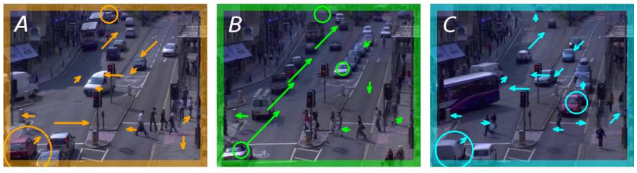
6.3 Clustering

6.3.1 Salient Activities

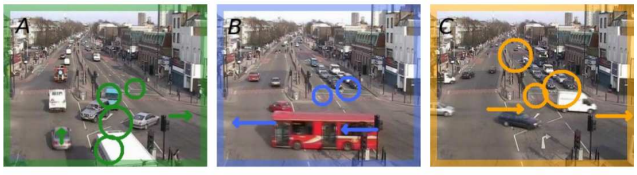
This Section demonstrates that the proposed nearest neighbor clustering approach can be used to detect typical activities in various scenarios. For example for the **Basket** dataset five main activities are automatically identified: (*A*) when the yellow team is on defense and the blue team is trying to shot, (*B*) when the players are moving from the yellow team's court side to the blue team's side, (*C*) when the blue team is on the defense, (*D*) when the players are moving back towards the yellow team's side. Moreover, due to



(a) Basket dataset



(b) Junction2 dataset



(c) Junction dataset

(d) Roundabout dataset

Fig. 5. Salient activities extracted with our method. Circles and arrows represent static and dynamic atomic activities respectively (size is proportional to bin value)

the asymmetric disposition of the camera with respect to the basketball court, different phases of the match can be observed when players are in the yellow team's side, such as the case of free throws (*E*). A representative frame for each of the five activities automatically extracted solving (10) is shown in Fig.5.a.

For the dataset **Junction2** we use our approaches for both two and three classes segmentation. The representative frames corresponding to the 3 clusters case automatically extracted are shown in Fig.5.b. These three flow patterns are regulated by three traffic lights, one in the bottom left, the second in the center and the third one in the right part of the image. Flow (*A*) corresponds to red traffic light in the bottom left lane; flow (*B*) to red traffic light in the central lane; flow (*C*) to red traffic lights in the bottom left and in the right lane. Atomic activities corresponding to pedestrians crossing the main road are also individuated (see the small arrows in the lower part of the images).

For the **Junction** dataset (Fig.5.c) by solving (10) or (11) we discover three main activities which correspond to different phases of the traffic flow: A) vertical flow and B) and C) respectively horizontal traffic flow from right to left and from left to right. These activities are also found in [1], [6], [7], with the difference that the cluster A is split in two different activities, corresponding to vertical flow with and without interleaved turning traffic. This division is less evident as it is confirmed by the transition behavior matrix in Fig.3.e in [6]. In fact, with our algorithm these patterns emerge when refining the analysis with more than three clusters. For the **Roundabout** dataset (Fig.5.d) two salient activities are discovered: they roughly correspond to the vertical (orange cluster) and the horizontal traffic flow (green cluster).

TABLE 4
Comparison of our approach with pLSA

	n° clusters	EMD-L_1 *linear*(10)	EMD-L_1 *circular*(11)	L_1 (13)	pLSA	pLSA bin
Basket	2	**98.42**	**98.42**	**98.42**	94.15	92.25
	5	**90.84**	**90.84**	75.17	83.5	77.5
Junction2	2	**96.20**	93.67	93.67	93.67	86.08
	3	84.81	**86.08**	70.89	79.40	75.60

TABLE 5
Junction2 dataset: clustering accuracy for different number of atomic activities

n°activities	30		24		16	
n°clusters	2	3	2	3	2	3
L_1	93.67	70.89	**96.20**	69.20	**96.20**	70.89
EMD-L_1-*lin.*	**96.20**	84.81	**96.20**	55.70	86.08	56.96
EMD-L_1-*circ.*	93.67	86.08	**96.20**	68.35	**96.20**	70.89
EMD-L_1-*2D*	**96.20**	**89.87**	**96.20**	72.15	**96.20**	73.42
k-means	88.83	68.24	**96.20**	67.05	94.41	59.73

6.3.2 *Comparison with Results in the Literature*

In this Section we perform a quantitative comparison between our methods and PTMs. Table 4 shows the results (percentage of correctly labeled clips) obtained by applying our methods (10) and (11) to the **Basket** sequence compared to (13) and to pLSA with binary and tf-idf features representation. For pLSA clustering labels are obtained by taking the topic with larger probability. pLSA has been chosen as a baseline since it has been extensively used in previous works [8], [31]. We consider the results for 2 and 5 clusters. The ground truth is taken from the APIDIS website[5]. In

5. We consider the timestamps of annotated events (*e.g.* 'Ball possession', 'Lost-ball', 'Free-throw', etc.) and added some missing information, *e.g.* the one representing a switch from events B to C or from D to A (Fig.5.a).

TABLE 6
Comparison with previous works: clustering accuracy

	EMD-L_1 *linear*(10)	EMD-L_1 *circular*(11)	L_1 (13)	Standard pLSA [38]	Hierarchical pLSA [38]	DDP-HMM [7]
Junction	**92.31**	**92.31**	89.74	89.74	76.92	87.18
Roundabout	**86.40**	**86.40**	**86.40**	84.46	72.30	85.14

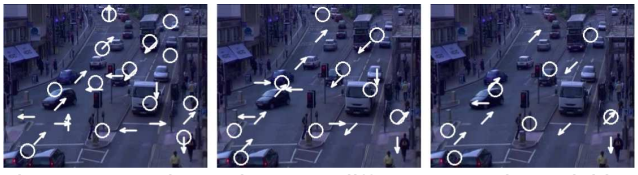


Fig. 6. Junction2 dataset: different atomic activities extracted with D = 30, D = 24 and D = 16.

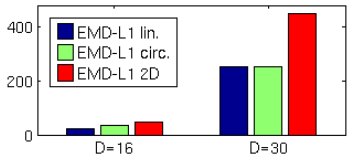


Fig. 7. Junction2 dataset: average computational time (sec) for solving our learning problems at varying D.

the case of 2 clusters the ground truth is created by merging the activities A and E on one side, fusing B, C and D on the other. Table 4 confirms the advantages of EMD-based approaches w.r.t. competing methods. For example, in the case of 5 clusters our methods outperforms pLSA with 7% in accuracy. We explain this with the fact that our approaches differently from (13) and pLSA takes into account atomic activities similarity. Moreover, it is worth noting that pLSA results depend upon initialization conditions, as training relies on a non-convex problem. On the 2 clusters task there is no advantage on using EMD based methods with respect to using bin-to-bin clustering approach (13). We believe that in some easy tasks bin-to-bin distances may suffice.

Similar conclusions can be made for the dataset **Junction2** (see Table 4). Also in this case EMD-based approaches outperform L_1 clustering and pLSA for the most difficult task (3 clusters). Other interesting remarks can be made observing Table 5. Here the results obtained with all proposed approaches are compared at varying number of atomic activities. The table demonstrates that few atomic activities may not suffice for accurate segmentation. This is basically due to the fact that missing atomic activities hinder the recognition of high level behaviour. For example for $D = 16$ the absence of the static atomic activities in upper left corner of the image inhibits the possibility to detect situations of traffic line (see Fig. 6). In these cases a $2D$ histogram representation with appropriate sorting compensates the decrease in accuracy. In this experiment we also report the results associated to *k-means* clustering as a baseline (Table 5). As expected, *ad-hoc* approaches as the ones we developed outperform standard clustering techniques.

In our datasets we found that the EMD-L_1 with linear histograms and EMD with thresholded L_1 distance and circular histograms perform similarly (see Table 4 for the **Basket** and **Junction2** sequences) with a slightly better performance for the latter representation (Table 5). Therefore with our approach we did not found great benefits in using a thresholded ground distance opposite to what was reported in the previous works [28]. This is probably due to the fact that we do not simply compute the EMD between noisy histograms as in [28] but we use EMD as an objective function to calculate the set of prototypes.

An important consideration concerns the computational cost associated to our approaches. Figure 7 reports the average time (sec) for solving the proposed

TABLE 7
Clustering accuracy with and without sorting.

		Junction	Roundabout	Basket	Junction2
EMD-L_1-*lin.*	sorted	92.31	86.4	90.84	84.81
	unsorted	86.7	72.3	82	75.95

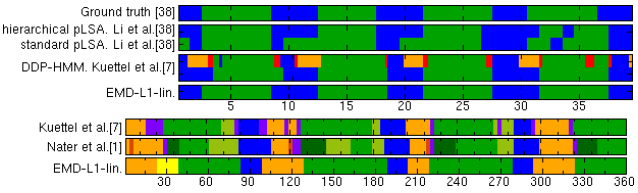


Fig. 8. Junction dataset. Comparison with previous works.

optimization problems (3.5 GHz Intel Xeon machine). As expected the computational costs associated to prototype learning of 1D histograms are comparable, while a 2D representation implies an increased cost due to a larger number of flow variables.

Table 6 compares our approach with previously published results. In particular we consider the results reported in [7], [38]. We apply (10) and (11) on the same data (the datasets **Junction** and **Roundabout**) using the same clip size as [38]. Results reported in [7] are obtained using a slightly different settings, *i.e.* clip length= 3 sec and 6 clusters. We manually merged these clusters to directly compare with the ground truth in [38]. The corresponding temporal segmentation bars for the Junction dataset are shown in Fig.8(top). On both datasets the proposed algorithms outperforms DDP-HMM [7], pLSA and hierarchical pLSA [38] (the experimental setup is slightly different as in [38] a training/test approach is used). EMD-based clustering is also more accurate than prototype learning with L_1 distance (13). These results confirm the fact that higher clustering accuracy can be obtained by considering atomic activities similarity during the learning phase. In the case of the Junction dataset we also compare our approach with the results presented in [1] which correspond to the short sequence of 360 sec, between frame 9201 and 18200, segmented at 7 levels. These results do not refer to the same part of the sequence annotated in [38], so a quantitative comparison is not possible. A qualitative comparison between our approach and [1], [7] is provided in Fig.8(bottom). As shown, the results of all three approaches are similar.

6.3.3 Ordering Atomic Activities

In this Section we present results demonstrating the validity of the proposed approach for sorting atomic

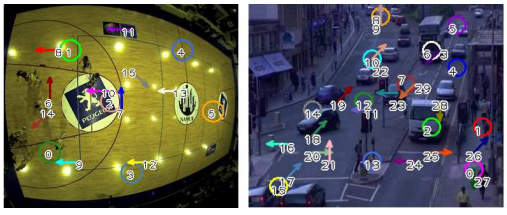


Fig. 9. Automatically sorted atomic activities.

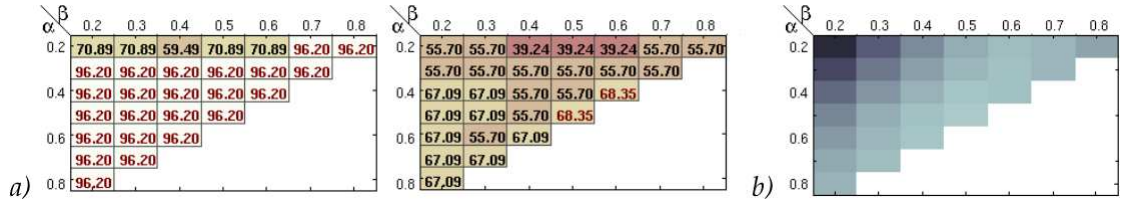


Fig. 10. Junction2 dataset (D=24). **a)** Clustering accuracy for 2 (left) and 3 (right) clusters for different atomic activities orders using EMD-L_1 circular. **b)** Associated distortion matrix (higher is darker)

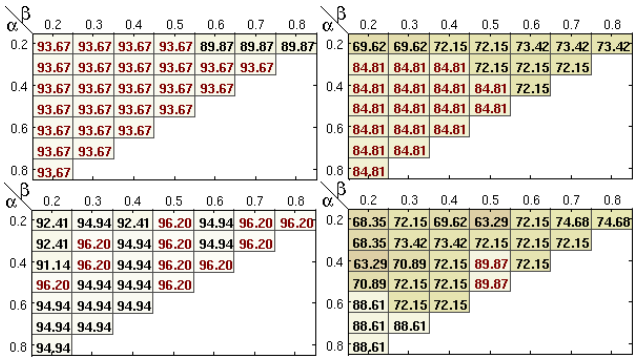


Fig. 11. Junction2 dataset (D=30). Clustering accuracy for 2 (left) and 3 (right) clusters with different atomic activities orders using (top) EMD-L_1-lin. and (bottom) EMD-L_1-2D.

Fig. 12. Traffic dataset: anomaly (motorbike U-turn)

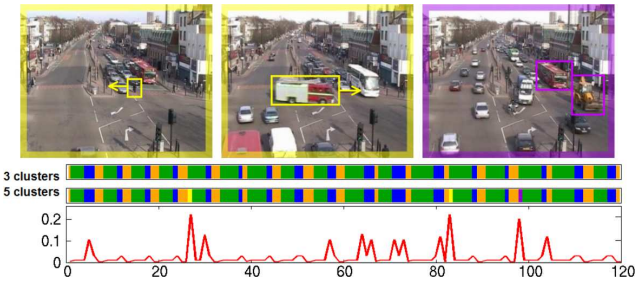


Fig. 13. Junction dataset: detected anomalies (top) and the associated MAS plot (bottom).

activities. Table 7 proves the importance of choosing an appropriate order of atomic activities for EMD prototype learning: for all the datasets a random order of atomic activities entails a decrease in terms of accuracy. Figure 9 shows an example of atomic activities automatically sorted for the **Basket** and the **Junction2** datasets in the case of EMD-L_1 with circular histograms and thresholded ground distance. For Basket dataset, this order corresponds to the highest accuracy (90.84% in the 5 clusters case) and it is obtained for values $\alpha = \beta = 0.5$, *i.e.* considering both the motion and the position information when computing the optimal sorting. It is straightforward to observe that similar atomic activities are grouped (for example the first 5 activities correspond to zero motion). In this way atomic activities typically corresponding to the same cluster (*e.g.* number 0, 1 and 2 for the Free Throw) are close in the histogram representation. Figure 11 reports the performance of the proposed approaches for the **Junction2** dataset at varying values of the parameters α and β, *i.e.* for different sorting. The plots demonstrate that in general while for an easy task (2 clusters) almost all type of sorting produces good results (accuracy around 95%), when more clusters are required it is very important to take into account both the motion and the position information. Temporal correlation is less important. Similar results were also obtained for the other datasets. Therefore as a practical rule of thumb we set $\alpha = \beta = 0.5$. Interestingly, in most of the cases we observe a certain correlation between the values of distortions computed with Eqn. (12) and the clustering accuracy (see Fig. **??**). Therefore looking at the distortion values can also be a valuable hint for sorting atomic activities.

6.4 Detecting Anomalous Patterns

By computing the MAS on an entire video sequence we detected some anomalous activities (persistent clusters of small size). In the case of the **Traffic** dataset an example of an unusual pattern is the violet cluster shown in Fig.4 corresponding to a jaywalker. By looking at the multiscale segmentation in Fig.4.a it is evident that the violet cluster, opposite to the others, "survives" for several levels. This single clip cluster correctly obtains a high MAS score as it is associated to an anomalous activity. Another example of anomalous activity in this sequence is shown in Fig.12. Here a motorbike makes a U-turn. This also corresponds to a single clip cluster which persist at several levels.

Figure 13 (top) shows some examples of anomalous activities found by MAS analysis (Fig. 13, bottom) for the dataset **Junction**. Anomalous activities corresponding to persistent small size clusters show the moments where the vertical traffic flows are interrupted as a pedestrian is crossing the street (clip 27) and a fireman truck is passing (clip 83). The last anomaly (clip 98) corresponds to a rare event where two large vehicles are passing at the same time. These results, similar to those in [1], [31], [38], confirm the

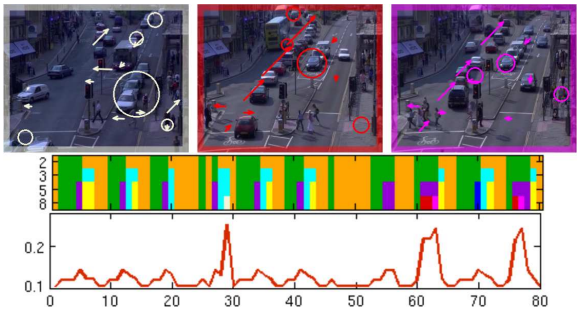


Fig. 14. Junction2 dataset: detected anomalies (top) and the associated MAS plot (bottom).

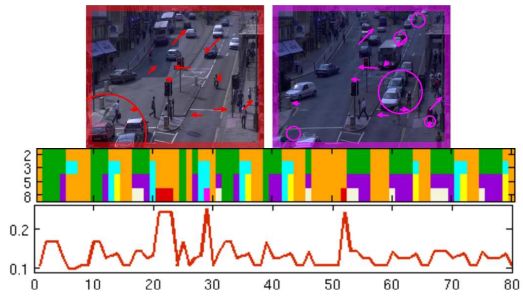


Fig. 15. Junction2 dataset: detected anomalies (top) and the associated MAS plot (bottom) when atomic activities are not correctly sorted.

validity of MAS analysis in finding anomalous events. In our experiments the MAS is computed considering $L = 9$ subsequent levels of segmentation. Figure 14 shows some anomalies detected for the **Junction2** dataset. Anomalies are due to an usual presence of a biker stopped next to the vehicles at the red traffic light (clip 28), a traffic jam (clip 62) and a traffic jam on the left lane when the traffic light is on green for the central lane (clip 77). Finally Fig. 15 demonstrates that a good atomic activities sorting is also crucial for detecting anomalous patterns. In fact in the case of an incorrect order also wrong clips are indicated as anomalous. For example in Fig. 15 clip 28 is correctly individuated but clips 21-23 have a high MAS value even if they do not correspond to critical situations.

7 Discussion and Conclusions

We proposed a multiscale approach for discovering activity patterns in complex scenes. The main novelty of this paper is the EMD prototype learning algorithm. By taking into account similarity amongst atomic activities, typical patterns can be extracted with improved accuracy with respect to previous approaches. The prototype learning algorithm has been presented in the context of dynamic scene analysis, but we believe that it could be successfully deployed in other tasks, such as facial expression analysis or action recognition.

In this work we considered the EMD approximation approach proposed in [26]. Recently, other methods [40], [41] have been proposed to speed-up the EMD distance calculation. These approaches are in general computationally more efficient than the one proposed in [26]. However, we chose Ling and Okada's approximation as it basically provides a simplification of the EMD definition proposing a LP with reduced flow variables. This LP can be easily embedded into our optimization framework and allows us to develop a Multiscale Analysis by Parametric LP theory. Moreover, the EMD wavelet approximation [40], [41] is especially convenient when the histogram size is larger than 200/300 bins. Differently, when very short histograms ($D \leq 50$) are considered as in this paper, the EMD wavelet approach is not advantageous since the overall computational cost is dominated by the initial wavelets coefficients calculation.

Our experiments showed that the proposed approach is a valuable alternative to PTMs in the context of complex scene analysis. Differently from PTMs our approach takes into account words similarity. However, it is worth noting that PTMs can be more versatile in applications when it is necessary to consider a large number of words, to learn the temporal dependencies among behaviors or to model the temporal information within the topics themselves. The proposed prototype learning algorithms can be also extended in several directions. For example in our previous work [11] we have shown how to embed the temporal information present inside the clips into our learning framework. Also we expect that by adopting the EMD approximation in [40], [41] our clustering approach can be applied to other problems where high dimensional histograms are needed. Finally future works include further exploiting the importance of atomic activity sorting: we expect to enhance even more the performance of our approach by introducing some form of weak supervision.

References

[1] F. Nater, H. Grabner, L. Van Gool, "Temporal Relations in Videos for Unsupervised Activity Analysis", *British Machine Vision Conference (BMVC)*, 2011.

[2] M. Marszalek, I. Laptev, C. Schmid, "Actions in context ", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] J. C. Niebles, C.-W. Chen, L. Fei-Fei. "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification ", *IEEE European Conference on Computer Vision (ECCV)*, 2010.

[4] M. D. Breitenstein, H. Grabner, L. Van Gool, "Hunting Nessie – Real-Time Abnormality Detection from Webcams", *IEEE International Workshop on Visual Surveillance*, 2009.

[5] T. Hospedales, J. Li, S. Gong, T. Xiang, "Identifying Rare and Subtle Behaviours: A Weakly Supervised Joint Topic Model", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 12, pp. 2451–2464, 2001.

[6] T. Hospedales, S. Gong, and T. Xiang, "A Markov Clustering Topic Model for Mining Behaviour in Video", *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[7] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering Spatio-Temporal Dependencies in Dynamic Scenes", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[8] J. Varadarajan, R. Emonet, and J.-M. Odobez, "Probabilistic Latent Sequential Motifs: Discovering temporal activity patterns in video scenes", *British Machine Vision Conference (BMVC)*,2010.

[9] X. Wang, K. Tieu, and W.E.L. Grimson, "Learning Semantic Scene Models by Trajectory Analysis", *European Conference on Computer Vision (ECCV)*, 2006.
[10] E.E. Zelniker, S.G. Gong, and T. Xiang, "Global Abnormal Detection Using a Network of CCTV Cameras", *Workshop on Visual Surveillance*, 2008.
[11] G. Zen, E. Ricci, S. Messelodi, and N. Sebe, "Sorting Atomic Activities for Discovering Spatio-temporal Patterns in Dynamic Scenes", *Int. Conf. on Image Analysis and Processing (ICIAP)*, 2011.
[12] G. Zen, and E. Ricci, "Earth Mover's Prototypes: a Convex Learning Approach for Discovering Activity Patterns in Dynamic Scenes", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
[13] T. Xiang, and S. Gong, "Video behavior profiling for anomaly detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, 5, pp. 893–908, 2008.
[14] X. Wang, X. Ma, and W.E.L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 3, pp. 539–555, 2008.
[15] Y. Yang, J. Liu, and M. Shah, "Video Scene Understanding Using Multi-scale Analysis", *IEEE International Conference on Computer Vision (ICCV)*, 2009.
[16] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A survey", *ACM Computing Surveys*, 41, 3, pp. 1–58, 2009.
[17] M. Breunig, H. Kriegel, R. Ng, and J. Sander, "LOF: Identifying density-based local outliers", *ACM SIGMOD International Conference on Management of Data*, 2000.
[18] J. Tang and Z. Chen and A.W. Fu and D.W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns", *Advances in Knowledge Discovery and Data Mining*, 2002.
[19] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, "Pathwise coordinate optimization"*Annals of Appl. Stat.*,1:302-332, 2007.
[20] Y. Yao, and Y. Lee. "Another look at linear programming for feature selection via methods of regularization", *Techn. Report 800, Dept. of Statistics, Ohio State University*, 2007.
[21] R. Sandler, and M. Lindenbaum, "Nonnegative Matrix Factorization with Earth Mover's Distance Metric", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[22] D. Lashkari, and P. Golland, "Convex clustering with exemplar-based models", *Advances in Neural Information Processing Systems (NIPS)*, 2008.
[23] S. Nowozin, and S. Jegelka, "Solution Stability in Linear Programming Relaxations: Graph Partitioning and Unsupervised Learning", *International Conference on Machine Learning*, 2009.
[24] C. Stauffer, and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999.
[25] Y. Rubner, C. Tomasi, and L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval", *International Journal of Computer Vision*, 40, 2, pp.99–121, 2000.
[26] H. Ling, and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29, 5, pp. 840–853, 2006.
[27] O. Pele, M. Werman, "A Linear Time Histogram Metric for Improved SIFT Matching", *European Conference on Computer Vision (ECCV)*, 2008.
[28] O. Pele, M. Werman, "Fast and robust Earth Mover's Distances", *IEEE International Conference on Computer Vision*, 2009.
[29] K. Murty, *Linear Programming*, Wiley, NY, 1983.
[30] D. Bertsimas, and J. N. Tsitsiklis, *Introduction to Linear Optimization., Athena Scientific*, 1997.
[31] J. Li, S. Gong, and T. Xiang, "Scene segmentation for behaviour correlation", *European Conference on Computer Vision*, 2008.
[32] L.I. Rudin, S. Osher and E. Fatemi, "Nonlinear total variation based noise removal algorithms", *Phys.*, 60, pp. 259–268, 1992.
[33] T. Xiang, and S. Gong, "Beyond Tracking: Modelling Activity and Understanding Behaviour", *International Journal of Computer Vision*, 67, 1, pp. 21–51, 2006.
[34] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: representing activities as bags of event n-grams", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
[35] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical bayesian models", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
[36] T. Duong, H. Bui, D. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
[37] S. Rosset and J. Zhu, "Piecewise Linear Regularized Solution Paths", *Annals of Statistics*, 35, 3, 2007.
[38] J. Li, S. Gong, and T. Xiang, "Global Behaviour Inference using Probabilistic Latent Semantic Analysis", *British Machine Vision Conference (BMVC)*, 2008.
[39] Y. Takano, and Y. Yamamoto, "Metric-Preserving Reduction of Earth Mover's Distance", *Asia-Pacific Journal of Operational Research*, 27, 39-54, 2010.
[40] S. Shirdhonkar, and D. W. Jacobs, "Approximate Earth Mover's Distance in Linear Time", *Computer Vision and Pattern Recognition (CVPR)*, 2008.
[41] D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek, "Unsupervised Clustering of Multidimensional Distributions Using Earth Mover Distance", *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.
[42] J. Wagner, and B. Ommer "Efficiently Clustering Earth Mover's Distance "*Asian Conference on Computer Vision (ACCV)*, 2010.
[43] R. Emonet, J. Varadarajan, and J.M. Odobez "Extracting and Locating Temporal Motifs in Video Scenes Using a Hierarchical Non Parametric Bayesian Model "*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Elisa Ricci is Assistant Professor at University of Perugia. She received her PhD from the same university in 2008. During her PhD she was a visiting student at University of Bristol. After that she has been a post-doctoral researcher at IDIAP, Martigny and the Fondazione Bruno Kessler, Trento. Her research interests are mainly in the areas of computer vision and machine learning.

Gloria Zen received her master degree in Telecommunication Engineering from the University of Trento in 2006. After that she worked as a researcher in computer vision industry and in Fondazione Bruno Kessler, Trento. She is currently leading a PhD at the Department of Information Engineering and Computer Science, University of Trento. Her main research interests include computer vision and video scene understanding.

Nicu Sebe is Associate professor at the University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He received his PhD from Leiden University, The Netherlands in 2001. He was involved in the organization of the major conferences addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a general/program chair of FG 2008, ACM CIVR 2007 and 2010, ACM Multimedia 2007, 2011 and 2013.

Stefano Messelodi graduated in computer science from the University of Milan (Italy). Since 1986 he is working in FBK (ITC-irst), Trento, Italy, where he coordinates the Technologies of Vision research unit. His research interests include text localization in scene, semantic image labelling and dynamic scene understanding. He served as a reviewer for several journals and conferences. He is a member of IEEE Society and International Association for Pattern Recognition.