Earth Mover's Prototypes: a Convex Learning Approach for Discovering Activity Patterns in Dynamic Scenes

Gloria Zen
DISI, University of Trento
via Sommarive 14, Povo, Italy
gloria.zen@disi.unitn.it

Elisa Ricci
Fondazione Bruno Kessler
via Sommarive 18, Povo, Italy
eliricci@fbk.eu

Abstract

We present a novel approach for automatically discovering spatio-temporal patterns in complex dynamic scenes. Similarly to recent non-object centric methods, we use low level visual cues to detect atomic activities and then construct clip histograms. Differently from previous works, we formulate the task of discovering high level activity patterns as a prototype learning problem where the correlation among atomic activities is explicitly taken into account when grouping clip histograms. Interestingly at the core of our approach there is a convex optimization problem which allows us to efficiently extract patterns at multiple levels of detail. The effectiveness of our method is demonstrated on publicly available datasets.

1. Introduction

Complex and crowded scenes depicting public spaces (*e.g*. city roads, subway stations) are especially challenging for video surveillance systems based on the traditional detection/tracking paradigm. This is mainly due to the presence of frequent occlusions in the scene and to the lack of appropriate models taking into account the spatial and temporal correlations between multiple objects. To face these difficulties, in the last few years there has been a growing interest in developing non-object centric methods for dynamic scene understanding [3, 12, 4, 13, 14].

Following this trend, we propose a novel approach for extracting spatio-temporal patterns in complex scenes. As in previous works [3, 4, 14], we use simple features computed from moving foreground pixels to locate atomic events and we cluster them into atomic activities. Then, given a video stream we divide it into short clips and for each clip we compute a histogram by summing up the occurrences of atomic activities. Previous works [3, 4, 14] adopting this 'word-document' representation assume that words do not follow a specific order into documents, thus ignoring that atomic activities are not independent. This implies an information loss since it is not possible to distinguish among groups of atomic activities corresponding to similar motion patterns (*e.g*. cars moving in the same direction in the same lane) and those representing objects of different size/speed in faraway regions of the scene.

In this paper, we propose to overcome this drawback by taking into account the similarity between atomic activities whilst learning the underlying behavioral patterns of the scene. We present a novel algorithm which adopts the Earth Mover's Distance (EMD) [9] in the objective function of the learning problem and outputs a set of histograms prototypes representing the discovered patterns. The main contributions of this paper are the following:

1. We formulate the task of mining typical behaviors in dynamic scenes as a *prototype learning* problem. Our approach is based on a *convex* optimization problem, specifically a Linear Program (LP), thus, it is not prone to local minima, it is rather scalable and it is easy to implement. To run experiments on medium-large scale datasets, following the idea in [7], we also develop a variant of our algorithm drastically reducing its computational cost.

2. We show how, with the proposed approach, salient patterns at *multiple scales* can be discovered. Differently from previous works and thanks to the theory of Parametric LP [1, 8], our algorithm performs a multiscale temporal segmentation of the video scene in *one shot*.

3. Comparing salient aspects extracted at multiple scales can also be useful in individuating anomalous patterns. To this aim we propose a Multiscale Anomaly Score (*MAS*).

We evaluate extensively our approach on four datasets (three of which are publicly available), showing that it offers competitive performance w.r.t. state-of-the-art methods. To our knowledge few resources (annotated video sequences, source codes) are publicly available for complex scenes analysis. To help filling this gap our code and the results from our experiments are made available to the community (http://disi.unitn.it/∼zen).

Related Works. Among previous works on complex

	Our	[6]	[3]	[14]	[4]	[12]
atom. activities correlation	√	×	×	×	×	×
multiscale analysis	√	×	×	√	×	×
convexity	√	×	×	×	×	×
learning temporal rules	×	∠	√	×	√	√
anomaly detection	√	√	√	×	√	√

Table 1. Qualitative comparison of the proposed and previous works: (√=yes, ×=no, ∠=partially).

scene analysis, those based on Probabilistic Topic Models (PTMs) [3, 12, 4, 6] have shown great potential. Recent works on hierarchical PTMs [3, 4] focus mainly on modeling the behaviors'correlation over time and inferring global rules. In this paper, we address different aspects trying to overcome some drawbacks of PTMs. First, we go beyond the usual word-document paradigm by taking into account the similarities among words during learning. Second, as our learning algorithm is based on a LP problem, we avoid the risk of being stacked into bad local minima typical of EM-like procedures. Third, by parametric LP, spatio-temporal activity patterns at multiple scales can be discovered efficiently and, under special conditions, without retraining from scratch. Another work similar to ours is [14], where a multiscale scene analysis is performed using diffusion maps in a preprocessing step before clustering. Differently, our multi-resolution analysis takes place in the clustering phase and it is also used for individuating unusual behaviors. Table 1 resumes the main features of our approach compared to previous works.

2. Preliminaries

LP and Parametric LP. Given $\mathbf{A} \in I\!R^{n\times m}$, $\mathbf{c}, \mathbf{a} \in I\!R^m$, $\mathbf{b} \in I\!R^n$, a LP in standard form [1, 8] is:

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

If the matrix $\mathbf{A}$ is of full rank and the polyhedron $\mathcal{D} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$ is bounded and non-empty, the LP has a bounded optimal solution. Let $\mathcal{B} \in \mathcal{I} = \{1, ..., m\}$ 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 $\mathbf{A}_\mathcal{B}^{-1}\mathbf{b} \geq 0$. A column $\mathbf{A}_i, i \in \mathcal{B}$ is called a basic column, otherwise it is a non-basic column and belongs to the set $\mathcal{N} = \mathcal{I} - \mathcal{B}$. A *basic feasible solution* (bfs) $\hat{\mathbf{x}}$ associated to a feasible basis $\mathcal{B}$ is obtained by $\hat{\mathbf{x}}_\mathcal{B} = \mathbf{A}_\mathcal{B}^{-1}\mathbf{b}$ and $\hat{\mathbf{x}}_\mathcal{N} = \mathbf{0}$. A bfs is *optimal* if it corresponds to a solution of the LP. There is a bijection between bfs and vertices of $\mathcal{D}$. The simplex method systematically explores the extreme points (bfs) of $\mathcal{D}$ starting from an initial extreme, until an optimal one is found. Given $\lambda \in \mathbb{R}$ a parametric LP has the form:

$$\min_{\mathbf{x}\geq 0} \ (\mathbf{c} + \lambda\mathbf{a})'\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{1}$$

Earth Mover's Distance. When comparing histograms, standard bin-to-bin distance functions (*e.g.* L_p distances, *KL* divergence) assume that the domains of the histograms are aligned, an assumption that is often violated due to noise. On the contrary EMD [9] addresses the alignment problem being a cross-bin distance function. The $EMD(\boldsymbol{h}, \boldsymbol{k})$ between two histograms $\boldsymbol{h}$ and $\boldsymbol{k}$ is obtained as the solution of the transportation problem:

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

provided that the histograms are normalized to unit mass, *i.e.* $\sum_q k^q = 1$ and $\sum_t h^t = 1$. 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.

3. Discovering Patterns in Complex Scenes

3.1. Overview

In this Subsection we describe the proposed approach for discovering spatio-temporal patterns in dynamic scenes. Fig. 1 illustrates the main steps of our method.

In the first phase low level features are extracted from the video and are used to define *atomic events*. We first apply a background subtraction algorithm to extract pixels of foreground. We found a simple dynamic Gaussian-Mixture background model [11] sufficient in our scenarios. We also compute the optical flow vector using the Lucas-Kanade algorithm. By thresholding the magnitude of the optical flow vectors we classify foreground pixels into static and moving pixels. We further differentiate among moving pixels quantizing the optical flow into 8 directions. Then we divide the scene into $p \times q$ patches. For each patch we consider the foreground pixels and their optical flow vectors and 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 represents the percentage of foreground pixels in the patch and $\bar{d}_{of}$ and $\bar{m}_{of}$ are respectively the mode of the orientations distribution and the average magnitude of optical flow vectors in the patch. For patches of static pixels we set $\bar{d}_{of} = \bar{m}_{of} = 0$. We define an *atomic event* as a patch descriptor $\boldsymbol{v}$ such that $f_g \geq T_{fg}$, *i.e.* we exclude patches with few pixels of foreground.

In the second phase we group atomic events with K-medoids clustering, rather than with K-means due to its increased robustness to noise and outliers. Each cluster represents an *atomic activity*. Then, we divide the video into short clips and for each clip we construct an *activity histogram* $\boldsymbol{h}_c$ representing the distribution of atomic activities in the c-th clip. Finally, the clips are grouped according to their similarity. To this aim we propose a novel algorithm which given a set of clips histograms identifies a smaller set of histogram prototypes representing the *salient activi-*

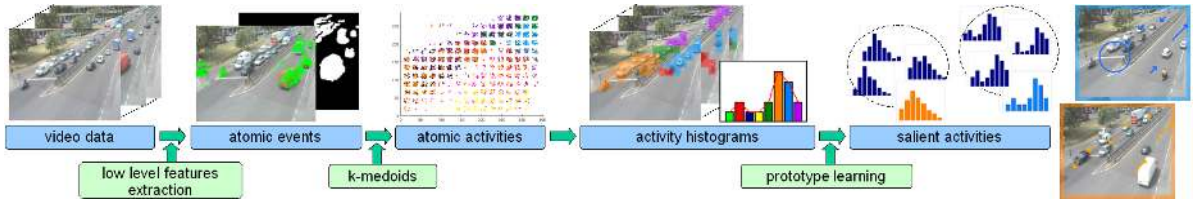

Figure 1. Flowchart of the proposed approach (best viewed in color)

ties occurring in the scene. In the following we describe the prototype learning algorithm.

3.2. Convex Prototype Learning

Suppose a set of histograms $\mathcal{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\}$, $\boldsymbol{h}_i \in I\!R^D$ is given. We aim to learn N representative prototypes $\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\}$, $\boldsymbol{p}_i \in I\!R^D$, each one associated to the original $\boldsymbol{h}_i$, such that their similarity with respect to the original data is maximized. Moreover, we want the set of prototypes to be a sparse representation of the original dataset $\mathcal{H}$, *i.e.* the number of different prototypes to be small. This can be obtained by minimizing their reciprocal differences. The overall task can be formalized as follows:

$$\min_{\boldsymbol{p}_i \in \boldsymbol{\Omega}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{h}_i, \boldsymbol{p}_i) + \lambda \sum_{i \neq j} \eta_{ij} \mathcal{J}(\boldsymbol{p}_i, \boldsymbol{p}_j) \tag{3}$$

The feasible region $\boldsymbol{\Omega} = \{\boldsymbol{p} : \forall t\ p_t \geq 0, \sum_t p_t = 1\}$ is meant to ensure that the prototypes are histograms normalized to unit mass. The objective function consists of two terms. The first term or loss should penalize the difference between the given histograms and the associated prototypes. The second term or regularization term must enforce the smoothness among related prototypes. Their relative importance 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 different prototypes between N and 1 can be obtained.

Learning Prototypes with EMD. In this paper we focus our attention on the cases where $\mathcal{L}(\cdot)$ and $\mathcal{J}(\cdot)$ are convex functions and specifically we present a formulation of (3) where the EMD is adopted as loss function:

$$\min_{\boldsymbol{p}_i \in \boldsymbol{\Omega}} \sum_{i=1}^{N} EMD(\boldsymbol{h}_i, \boldsymbol{p}_i) + \lambda \sum_{i \neq j} \eta_{ij} \max_{q=1\ldots D} |p_i^q - p_j^q| \tag{4}$$

The advantage of using EMD is motivated by the fact that the ground distances d_{qt} can encode information about the similarity of atomic activities. To this aim, since each atomic activity q is represented by the associated exemplar $\boldsymbol{m}_q$ computed by K-medoids, we set $d_{qt} = \|\boldsymbol{m}_q - \boldsymbol{m}_t\|^2$.

As stated above the rightmost term in (4) is meant to minimize the number of different prototypes. The adoption of the L_1 norm induces sparsity, thus producing a small number of prototypes. Generally a comparison among all possible pairs $\boldsymbol{p}_i, \boldsymbol{p}_j$, $i \neq j$, is required imposing all prototypes to be close to each other. However this implies an increased computational cost when solving (4). To alleviate this fact we introduce the binary coefficients $\eta_{ij} \in \{0, 1\}$ in order to select only a subset of pairs of histograms which must be merged. In the absence of prior knowledge, we can simply identify for each histogram $\boldsymbol{h}_i$ a set of P nearest neighbors and set $\eta_{ij} = 1$ if $\boldsymbol{h}_j$ is a neighbor of $\boldsymbol{h}_i$. Alternatively, temporal dependencies can be encoded: if histograms represent temporally adjacent clips we set $\eta_{ij} = 1$ if $i = j - 1, \forall j = 2 \ldots N$, $\eta_{ij} = 0$ otherwise. As shown in the experimental section, we tested both approaches: in the following we refer to them respectively as nearest neighbors clustering and temporal segmentation. Note that the coefficients η_{ij} are fixed and do not change during learning. To compute histogram prototypes we substitute the EMD definition (2) in the loss in (4) we get the following LP:

$$\begin{aligned} \min_{p_i^q, f_{qt}^i, \zeta_{ij} \geq 0} \quad & \sum_{i=1}^{N} \sum_{q,t=1}^{D} d_{qt} f_{qt}^i + \lambda \sum_{i \neq j} \eta_{ij} \zeta_{ij} \\ \text{s.t.} \quad & -\zeta_{ij} \leq p_i^q - p_j^q \leq \zeta_{ij},\ \forall q, \forall i, j \\ & \sum_{q=1}^{D} f_{qt}^i = h_i^t,\ \forall t \quad \sum_{t=1}^{D} f_{qt}^i = p_i^q,\ \forall q, \forall i \end{aligned} \tag{5}$$

where we introduced the slack variables ζ_{ij}. Note that the constraints $\sum_t p_i^t = 1$ are removed since they are automatically satisfied. 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 it produces the effects that all coordinates of a prototype go to zero together and it reduces significantly the computational cost of solving (5), limiting the number of slack variables.

Learning Prototypes with EMD-L_1. For large N and D solving (5) 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 $\boldsymbol{p}_i$. To speed up calculations we propose a modification of (5) which adopts EMD with L_1 distance over bins as ground distance, *i.e.* $d_{qt} = |q - t|$. The idea is that similar atomic activities should correspond to neighboring bins in activities histograms. To this aim we sort the cluster medoids computed with K-medoid according to the associated motion information from exemplars corresponding to static events to those

associated with optical flows with large magnitude. The orientation is also taken into account in this phase.

In case of EMD-L_1 every positive flow between faraway histogram bins can be replaced by a sequence of flows between neighbor bins [7]. Thus (2) can be simplified as:

$$\min_{g_{q,q+1},g_{q,q-1}\geq 0} \quad \sum_{q=1}^{D-1} g_{q,q+1} + \sum_{q=2}^{D} g_{q,q-1} \tag{6}$$
$$\text{s.t. } g_{q,q+1} - g_{q+1,q} + g_{q,q-1} - g_{q-1,q} = h^q - p^q \; \forall q$$

With (6) the number of variables and constraints decreases significantly. In particular the number of flow variables involved reduces from $O(D^2)$ 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. With these premises, we simplify (5) by substituting the EMD-L1 definition (6) in the loss:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N}\sum_{q=1}^{D-1} g^i_{q,q+1} + \sum_{i=1}^{N}\sum_{q=2}^{D} g^i_{q,q-1} + \lambda \sum_{i\neq j} \eta_{ij}\zeta_{ij} \qquad (7) \\ \text{s.t.} \quad & -\zeta_{ij} \leq p^q_i - p^q_j \leq \zeta_{ij}, \; \forall q, \; \forall i,j, \; i \neq j \\ & g^i_{q,q+1} - g^i_{q+1,q} + g^i_{q,q-1} - g^i_{q-1,q} = h^q_i - p^q_i, \; \forall q, i \\ & p^q_i, g^i_{q,q+1}, g^i_{q,q-1}, \zeta_{ij} \geq 0 \end{aligned}$$

The resulting optimization problem is a LP with $n_{var} = 2N(D-1) + ND + \frac{1}{2}N(N-1)$ variables if we impose each prototype to be close to each other, *i.e.* $\eta_{ij} = 1 \; \forall i \neq j$. In this case for large datasets ($N \gg D$) the computational cost of (7) is dominated by the number of slack variables that is quadratic w.r.t. the number of datapoints. However, as we discussed above, by setting some of the $\eta_{ij} = 0$, (7) can be solved efficiently even in case of large datasets.

Learning Prototypes with Bin-to-bin Distances. To demonstrate the advantages of considering cross-bin distances when learning prototypes, we briefly discuss the form that (3) assumes when bin-to-bin distances are considered as loss functions. When $\mathcal{L}(\cdot)$ is the square loss, $\mathcal{J}(\cdot)$ is a sum of L_1 norms and $\eta_{ij} = 1$ if $i = j-1$ and $\eta_{ij} = 0$ otherwise, we get something very close to the "total variation denoising" [10] or to the fused lasso[1] [2]. When the L_1 norm is chosen as loss and a combination of L_1-L_∞ norms is used for $\mathcal{J}(\cdot)$, (3) assumes the form of the following LP:

$$\min_{\boldsymbol{p}_i \in \boldsymbol{\Omega}} \quad \sum_{i=1}^{N}\sum_{q=1}^{D} |h^q_i - p^q_i| + \lambda \sum_{i\neq j} \eta_{ij} \max_{q=1\ldots D} |p^q_i - p^q_j| \tag{8}$$

In the experimental section we show that bin-to-bin distances are less effective than EMD when learning prototypes for dynamic scene understanding.

[1]Note also that in the fused lasso formulation [2] the feasible region is slightly different from $\boldsymbol{\Omega}$ since the $\boldsymbol{p}_i$ are not specifically constrained to be normalized histograms.

Algorithm 1 One shot temporal segmentation

Input: $\mathbf{H} = (\boldsymbol{h}'_1 \ldots \boldsymbol{h}'_N)'$, $i = 0$, $\mathcal{I}_p = \{ND^2 + N + k : k = 0, \ldots, ND\}$.
Set (5) in standard form (1) according to Proposition 1.
Find an optimal bfs $\mathcal{B}_0$ for $\lambda_0 = \infty$.
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 $j \in \mathcal{N}_i$.
  $m = \arg\max_j \{-\bar{c}_j/\bar{a}_j : \bar{a}_j > 0\}$ (*entry index*)
  $\lambda_{i+1} = -\bar{c}_m/\bar{a}_m$
  $\mathbf{u} = \mathbf{A}^{-1}_{\mathcal{B}_i}\mathbf{A}_m$.
  if the support $I(\mathbf{u})$ is empty **then return**
  $\ell = \text{arglexico-min}_t \{\mathbf{A}^i_t/u_t : t \in I(\mathbf{u})\}$ (*exit index*)
  Update $\mathcal{B}_{i+1} = \mathcal{B}_i \cup \{m\} \backslash \{\ell\}$
  Set $\mathbf{P}_i = \mathbf{x}^i_{\mathcal{I}_p}$
  $i \leftarrow i + 1$
end
Output: The set $\{\mathbf{P}_1, \ldots \mathbf{P}_N\}$

3.3. Multiscale Analysis and MAS

A crucial property of (5) and (7) is that the sparsity achieved is controlled by a single parameter, *i.e.* the regularization constant λ. For λ varying between ∞ and 0 a different number of prototypes $M(\bar{\lambda})$ between 1 and N is obtained. Instead of trying to find the value of λ which provides the best prototype representation we propose to exploit the solutions of (5) and (7) for different values of λ. This corresponds to discover different salient activities at multiple scales. For example, in case of temporal segmentation in a traffic scene, for large values of λ we can obtain two prototypes, corresponding to a rough description of the scene, distinguishing among clips with moving vehicles and with vehicles stopped at the traffic lights. As λ decreases we gradually enhance the level of detail differentiating among vehicles flows of different intensity.

Comparing clustering results at multiple scales we can detect unusual behaviors corresponding to atypical histograms. To this aim we define for each $\boldsymbol{h}_k$ an associated anomaly score. The idea 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 $\boldsymbol{h}_k$ corresponds to a different $\boldsymbol{p}_k$), the anomaly score of $\boldsymbol{h}_k$ is computed as the sum of the ratios of the clusters size containing $\boldsymbol{h}_k$ at two subsequent scales. Formally we first introduce the notion of sets of fused histograms as they are generated by our algorithms.

Definition 1. (Sets of Fused Histograms) *Let λ be a fixed value $\bar{\lambda}$ and $\mathcal{H}^{\bar{\lambda}}_k$ be a set of histograms with $k = 1, \ldots, M(\bar{\lambda})$. Then a valid set of fused histograms $\mathcal{H}^{\bar{\lambda}}_k$ satisfies the following properties:*

- ◇ *the collection of the sets $\mathcal{H}^{\bar{\lambda}}_k$ is a partition of $\mathcal{H}$*

$\diamond$ $\forall\, \boldsymbol{h}_\ell, \boldsymbol{h}_m \in \mathcal{H}_k^{\bar{\lambda}}$ *we have* $p_\ell^q = p_m^q\ \forall\, q = 1 \dots D$

$\diamond$ $\forall\, \boldsymbol{h}_\ell \in \mathcal{H}_k^{\bar{\lambda}}$ *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 fused histograms are generated for different λ. By looking at the sets of fused histograms we can define the MAS.

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

$$MAS = 1 - \frac{1}{NL} \sum_{i=1}^{L} \frac{|\mathcal{H}_{k'}^{\lambda_{i-1}}|}{|\mathcal{H}_k^{\lambda_i}|} \tag{9}$$

Thus analyzing multiple scales 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 MAS is higher. Note that MAS definition is possible since with our approach if two histograms belong to the same fused set for $\lambda = \bar{\lambda}$ they will generally remain fused for any $\lambda \geq \bar{\lambda}$. As a final remark we should say that while large values of L usually result in more accurate estimates of MAS, this also increases the computational cost since (5) must be solved L times. In the following we show how in some cases all possible sets of fused histograms can be obtained with computational cost comparable with that of solving (5) once.

3.4. Multiscale Analysis in One Shot

In this section we focus our attention on EMD prototype learning, and we show that since (5) is a parametric LP an algorithm based on a variant of the revised simplex method [8] can be used to compute all possible sets of prototypes for increasing values of λ. We consider the specific case of temporal segmentation *i.e.* we set $\eta_{ij} = 1$ for $i = j-1, j = 2 \dots N$ and $\eta_{ij} = 0$ otherwise.

Let $\mathbf{H}' = (\boldsymbol{h}'_1 \dots \boldsymbol{h}'_N)$, $\mathbf{P}' = (\boldsymbol{p}'_1 \dots \boldsymbol{p}'_N)$, $\mathbf{H}, \mathbf{P} \in I\!R^{ND}$. Let $\mathbf{1} \in I\!R^D$ denote the vector $\mathbf{1} = (1 \ \dots 1)$, $\mathbf{I}$ be the identity matrix and $\mathbf{0}$ be the zero matrix of appropriate dimension. We first define the following block diagonal matrices $\mathbf{D} \in I\!R^{(N-1)D \times (N-1)}$, $\mathbf{D} = \mathrm{diag}(\mathbf{1})$, $\mathbf{F}, \mathbf{G} \in I\!R^{ND \times ND^2}$, $\mathbf{F} = \mathrm{diag}(\mathbf{Q}), \mathbf{G} = \mathrm{diag}(\mathbf{T}), \mathbf{Q} \in I\!R^{D \times D^2}$, $\mathbf{Q} = \mathrm{diag}(\mathbf{1}')$, the block Toeplitz matrix $\boldsymbol{\Sigma} \in I\!R^{(N-1)D \times ND}$, and the matrix $\mathbf{T} \in I\!R^{D \times D^2}$ such that:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{I} & \mathbf{-I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{-I} & \dots & \mathbf{0} \\ \vdots & \ddots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I} & \mathbf{-I} \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \vdots \\ \mathbf{E}_D \end{pmatrix},$$

where $\mathbf{I}$, $\mathbf{0}$ and $-\mathbf{I} \in I\!R^{D \times D}$ and $\mathbf{E}_i = (\mathbf{e}'_i\ \mathbf{e}'_i \dots \mathbf{e}'_i)$, $\mathbf{e}'_i \in I\!R^D$, $\mathbf{e}'_i = (0 \dots 0\ 1\ 0 \dots 0)$ with a 1 in the i-th position.

Proposition 1. *Let* $\boldsymbol{f} \in I\!R^{ND^2}$ *be the vector of flow variables,* $\boldsymbol{\delta}_+, \boldsymbol{\delta}_- \in I\!R^{ND}, \boldsymbol{\zeta} \in I\!R^{N-1}$ *be slack variables. Let* $\boldsymbol{\omega} \in I\!R^{ND^2}$ *be the vector containing the ground distance values i.e.* $\boldsymbol{\omega} = (\mathbf{d} \dots \mathbf{d})$, $\mathbf{d} \in I\!R^{D^2}$, $\mathbf{d} = (d_{11}, \dots d_{1D}\ d_{21} \dots d_{DD})$. *The following elements:* $\mathbf{x}' = \left(\boldsymbol{f}'\ \boldsymbol{\zeta}'\ \mathbf{P}'\ \boldsymbol{\delta}_+{}'\ \boldsymbol{\delta}_-{}'\right)$, $\mathbf{a}' = \left(\boldsymbol{\omega}'\ \mathbf{0}'\ \mathbf{0}'\ \mathbf{0}'\ \mathbf{0}'\right)$, $\mathbf{c}' = \left(\mathbf{0}'\ \mathbf{1}'\ \mathbf{0}'\ \mathbf{0}'\ \mathbf{0}'\right)$,

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{-D} & \boldsymbol{\Sigma} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{-D} & \mathbf{-\Sigma} & \mathbf{0} & \mathbf{I} \\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{G} & \mathbf{0} & \mathbf{-I} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{H} \\ \mathbf{0} \end{pmatrix}$$

define (5) in the standard form (1) of a parametric LP.

In [15] 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 (1) and the tableau simplex method can be used for solving (1) for all possible values of λ simultaneously. In this paper we propose to solve (5) using a variation of the algorithm proposed in [15] by considering rather than the tableau simplex method, the revised simplex with the lexico-min rule since it offers computational advantages for sparse LPs and avoid situations of degeneracy. The resulting algorithm is presented in Algorithm 1. The main difficulty when applying Algorithm 1 is how to individuate an optimal bfs $\mathcal{B}_0$. An optimal bfs $\mathcal{B}_0$ 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 proposition shows an example of a bfs $\bar{\mathcal{B}}_0$ for (5).

Proposition 2. *The set of indices* $\bar{\mathcal{B}}_0 = \mathcal{I}_1 \cup \mathcal{I}_2$ *with* $\mathcal{I}_1 = \{kD+1 : k = 0, \dots, ND-1\}$, $\mathcal{I}_2 = \{ND^2+N+k : k = 0, \dots, 3ND - 2D - 1\}$ *individuates a bfs for (5).*

Similar results can be obtained for (7) and (8) in case of temporal segmentation. When the coefficients η_{ij} assume arbitrary values, (5), (7) and (8) are also parametric LP problems and Algorithm 1 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 further works.

4. Experimental results

We tested our method on four datasets. Due to lack of space we encourage the reader to look at the supplementary material submitted with this paper to see the videos associated to our results. The proposed approach is fully implemented in C++ using the publicly available libraries OpenCV and GLPK 4.2.1 (GNU Linear Programming Kit).

The first dataset consists of a **traffic scene** sequence depicting a crossroads. In this scenario different events occur at regular periods as the vehicles flow is controlled by traffic lights. The second dataset is publicly available and is taken from **APIDIS**[2]. Here players involved in a basketball match

[2] http://www.apidis.org/Dataset/

	Traffic	Basket	Junction	Roundabout
n^o frames	6000	6000	90000	93500
fps	12	23	25	25
n^o clips	300	100	240	311
frame size	276×336	320×368	288×360	288×360
patch size	23×21	16×16	12×12	12×12
D	8	16	16	16

Table 2. Details of the setting used in our experiments.

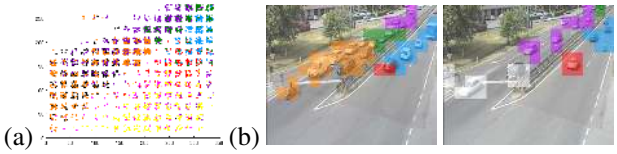

Figure 2. Traffic dataset: (a) K-medoids results (b) Example of atomic activities.

are depicted. In detail we pick the sequence of camera 5 from 20080409T184900 to 20080409T185400. The images are resized and cropped in a way to include only the basketball court. The third and forth datasets [3] are also publicly available[3,4]. They depict two complex traffic scenes in London (**Junction** and **Roundabout**) and they both correspond to a video of about 1 hour duration. More details about the datasets and the experimental setup are reported in Table 2. We chose the first dataset since it is suitable for testing our temporal segmentation approach, as it corresponds to few cycles of the traffic lights status and it contains some interesting anomalous events. The nearest neighbor clustering is adopted for experiments on the other datasets.

Traffic dataset. We report some results from the first phase of our analysis, which aims to classify the atomic events according to their position and motion. Labeled atomic events as they are obtained with K-medoids are depicted in Fig.2.a, where each color corresponds to a specific atomic activity (note that for visualization purposes we plot only a small subset of collected events and we add a small random shift to their position (x, y) in the image plane). It is easy to observe that neighboring events belong to the same group and where the same region contains two clusters, the clusters correspond to activities with different motions. Fig.2.b shows this situation: the orange activity corresponds to vehicles stopped due to red traffic light, while the white one shows moving vehicles in the same area when the traffic lights are on green.

We further show the effectiveness of our approach in segmenting scene activities in time. Figure 4 shows the results on 100 clips that we obtained by solving (7) with $\lambda = 5$. From each of the 10 clusters obtained we extract one frame, representative for the salient activities (due to lack of space we just show 9 of them). The orange, yellow, and red clusters correspond to the activity of parallel vehicle flows (green traffic light), while the light blue, white and

[3]http://www.eecs.qmul.ac.uk/ jianli/Junction.html
[4]http://www.eecs.qmul.ac.uk/ jianli/Roundabout.html

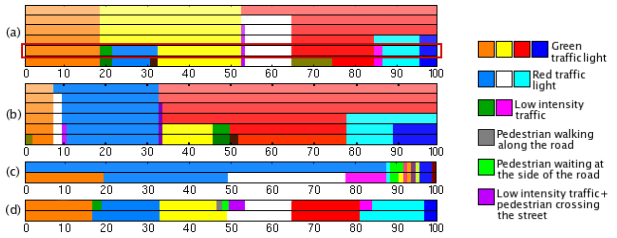

Figure 3. Traffic dataset: comparison of temporal segmentation results obtained with (a) EMD-L_1(5), (b) L_1(8) and (c) Fused lasso. (d) Human annotation at different levels of detail. The red rectangle indicates the segmentation corresponding to Fig.4.

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. Fig.3.a shows the multiscale segmentation obtained on the same sequence solving (7) for different values of λ. 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. The violet cluster, instead, "survives" for several levels. This is an expected result because the clip is associated to an anomalous activity (a jaywalker). We also compare the multiscale temporal segmentation obtained with (7) with the ones we get by solving (8) (Fig.3.b) and with the fused lasso method (Fig.3.c). A visual inspection confirms that the temporal segmentation obtained with EMD distance is more consistent with the results provided by a human annotator (Fig.3.d). This demonstrates that bin-to-bin distances are less powerful as they do not take into account correlations among atomic activities.

A quantitative comparison of the proposed methods ((5) and (7)) and bin-to-bin approaches (Fused lasso and (8)) for the entire sequence is shown in Table 3. The segmentation provided by a human annotator is used as a ground truth. The performance is measured in terms of percentage of break points correctly individuated. It is worth noting that (7) can be considered as a good approximation of (5) as confirmed by the first two columns of the table. By computing the MAS on the entire sequence we detected some anomalous activities (persistent clusters of small size). An example of an unusual pattern is the violet cluster shown in Fig.4 corresponding to the jaywalker. Another example is shown in Fig.5 where a motorbike makes a U-turn.

An important observation concerns the computational cost of our multiscale analysis. As (5) is a parametric LP, this allows us to find *all* solutions (*i.e.* *all* possible prototypes) with a slightly increased computational cost w.r.t. computing just one solution (corresponding to a fixed value of λ). Therefore, the speedup is huge. For example all possible prototypes associated to the 100 clips in Fig.4 can be

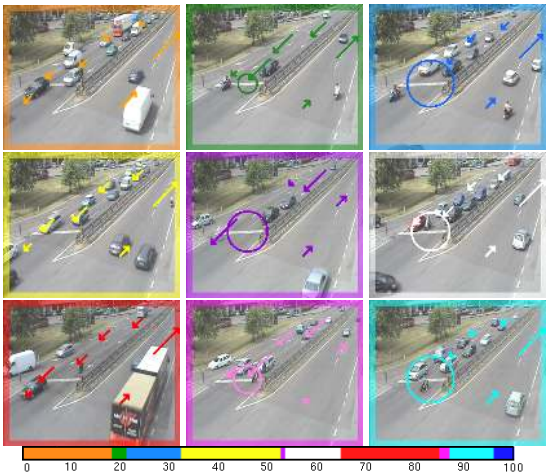

Figure 4. Traffic dataset: (top) salient activities and (bottom) EMD-L_1(7) results.

Figure 5. Anomaly: U-turn of a motorbike.

EMD (5)	EMD-L_1(7)	L_1(8)	Fused Lasso
83.2	82.4	72.5	68.7

Table 3. Temporal segmentation accuracy for the traffic dataset

computed in approximately 5 minutes whilst the solution for just one value of λ takes about 1 minute.

Basket dataset. We chose this dataset to demonstrate that the proposed approach can be used for applications other than traffic scene analysis, such as to obtain a rough synthesis and useful statistics of a sport match. In particular in the APIDIS sequence five main activities can be identified: (A) when the yellow team is on defence 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 yellow team is on the defence, (D) when the players are moving back towards the yellow team's side. Moreover, due to the asymmetric disposition of the camera w.r.t. 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 as they are automatically extracted by our algorithm (7) is shown in Fig.6(top). Furthermore, Fig.6(bottom) compares the results (*i.e.* cluster assignments) for 100 clips obtained with (7) and the ground truth. The ground truth is taken from the APIDIS website. In detail, we select the timestamps of annotated events (*e.g.* 'Ball possession', 'Lost-ball', 'Free-throw','Rebound', etc.) and consider them as breakpoints. We also added some missing breakpoints, *e.g.* the ones representing a switch from events B to C or from D to A.

Table 4 shows the results of a quantitative evaluation of

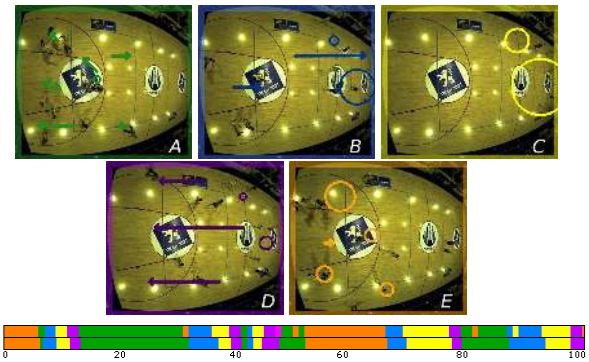

Figure 6. Basket dataset: (top) salient activities and (bottom) (a) EMD-L_1(7) results, (b) ground truth.

n^o clusters	EMD-L_1(7)	L_1(8)	pLSA	pLSA-bin
5	**90.84**	75.17	83.5	77.5
2	**98.42**	**98.42**	94.15	92.25

Table 4. Clustering accuracy (percentage of correctly labeled clips) for the basket dataset

our method (7) compared to (8) and to probabilistic Latent Semantic Analysis (pLSA) with binary and tf-idf features representation. PLSA has been chosen as a baseline since it has been extensively used in previous works [12, 6]. We consider the results for 2 and 5 clusters. In case of the 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-L_1 w.r.t. competing methods. For example, in case of the 5 clusters (7) outperforms the best competing method with 7% more of accuracy. Moreover it is worth noting that pLSA results depends upon initialization conditions, as training relies on a non-convex problem.

London's traffic datasets. We chose these datasets since they have been extensively used in previous works [3, 4, 5, 6]. However few of them provide the ground truth annotations and quantitative results. One of the few exceptions is [5]. Table 5 compares the results in [5] with those we get on the same data (and same clip size) by applying (7) with $P = 3$. On both datasets the proposed algorithm outperforms both pLSA and hierarchical pLSA used in [5]. It is interesting to observe that choosing a suitable order of atomic activities when constructing histograms is crucial: using a random order the performance decreases significantly. These results refer to the situation where only two salient activities are considered. Fig.7 shows an example of the typical activities for the dataset Roundabout.

We also apply (7) for individuating more than two salient activities. In this case we only provide a qualitative evaluation since quantitative results are not available in literature. For the Junction dataset (Fig.8, top) 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 [3, 4], with the difference that in

Figure 7. Roundabout dataset: example of typical activities.

	EMD-L_1(7)	L_1(8)	EMD-L_1(7) random	Standard pLSA [5]	Hierarchical pLSA [5]
J	**92.36**	89.74	86.7	89.74	76.92
R	**86.40**	**86.40**	72.3	84.46	72.30

Table 5. Clustering accuracy (%) for Junction (J) and Roundabout (R) datasets

[3, 4] 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 [3]. In fact, with our algorithm these patterns emerge when refining the analysis with more than three clusters. Finally we show some examples of anomalous activities (Fig.8, middle) found by MAS analysis (Fig.8, bottom). Anomalous activities corresponding to persistent small size clusters show the moments where the traffic lights are on green and vehicles have to stop as a pedestrian is crossing the street (clip 27) and a fireman truck is passing (clip 83). The last case (clip 98) corresponds to a rare event where two large vehicles are passing at the same time. These results, similar to those in [5, 6], confirm the validity of MAS analysis in finding anomalous events. In our experiments the MAS is computed considering $L = 9$ subsequent levels of segmentation. However, with large values of L, more accurate MAS profiles can be obtained, at the expenses of an increased computational cost. How the anomaly detection performance is affected by L will be investigated in future works.

5. Conclusions

We proposed a novel multiscale approach for discovering activity patterns in complex scenes. By taking into account correlations amongst atomic activities, typical patterns can be extracted with improved accuracy w.r.t. previous methods. Moreover, if we also learn the temporal dependencies among behaviors, as other state-of-the-art approaches do, we believe that the potential of our method will be even better exploited. We leave this to future works.

The main novelty of this paper is the EMD prototype learning algorithm: we used for dynamic scene understanding, but we believe that it could be deployed in other tasks, such as facial expression analysis or action recognition.

References

[1] D. Bertsimas and J. N. Tsitsiklis. Introduction to linear optimization. 1997. Athena Scientific. 3225, 3226

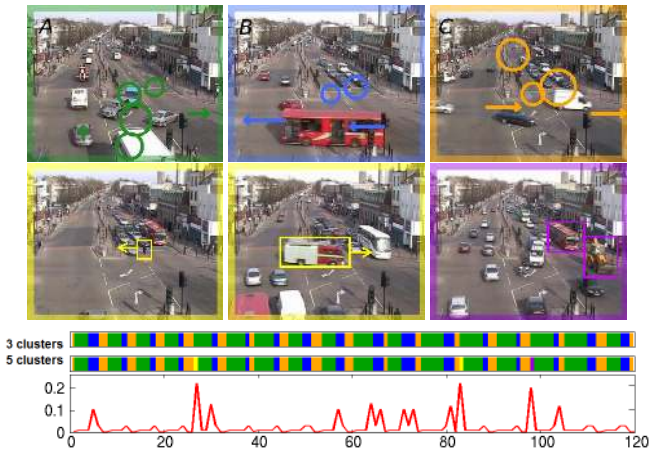

Figure 8. Junction dataset. Three salient activities (top), detected anomalies (middle) and the associated MAS plot and EMD-L_1 clustering results (bottom).

[2] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302–332, 2007. 3228

[3] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. *ICCV*, 2009. 3225, 3226, 3230, 3231, 3232

[4] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. *CVPR*, 2010. 3225, 3226, 3231, 3232

[5] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. *BMVC*, 2008. 3231, 3232

[6] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. *ECCV*, 2008. 3226, 3231, 3232

[7] H. Ling and K. Okada. An efficient Earth Mover's Distance algorithm for robust histogram comparison. *IEEE Trans. on PAMI*, 29(5):840–853, 2006. 3225, 3228

[8] K. Murty. Linear programming, 1983. Wiley, NY. 3225, 3226, 3229

[9] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 3225, 3226

[10] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys.*, 60:259–268, 1992. 3228

[11] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *CVPR*, 2(1):246–252, 1999. 3226

[12] J. Varadarajan, R. Emonet, and J.-M. Odobez. Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. *BMVC*, 2010. 3225, 3226, 3231

[13] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *IEEE Trans. on PAMI*, 30(5):893–908, 2008. 3225

[14] Y. Yang, J. Liu, and M. Shah. Video scene understanding using multi-scale analysis. *ICCV*, 2009. 3225, 3226

[15] Y. Yao and Y. Lee. Another look at linear programming for feature selection via methods of regularization. 2007. Techn. Report 800, Dept. of Statistics, Ohio State University. 3229