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

Gloria Zen[1], Elisa Ricci[2,3]

[1] DISI, University of Trento, Trento, Italy.

[2] TEV, Fondazione Bruno Kessler (FBK), Trento, Italy.

[3] DIEI, University of Perugia, Italy.

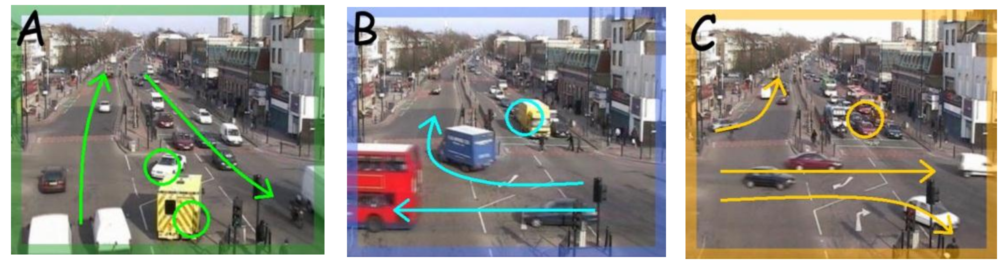
International Conference Computer Vision and Pattern Recognition (CVPR), 2011

Mining behaviors in complex scenes

Goals:

i) to mine patterns of **recurrent activities**

(*e.g.* vertical/horizontal traffic flows)



ii) to detect **anomalies**

(*e.g.* jaywalker, accident, unusual patterns)

video: junction.avi

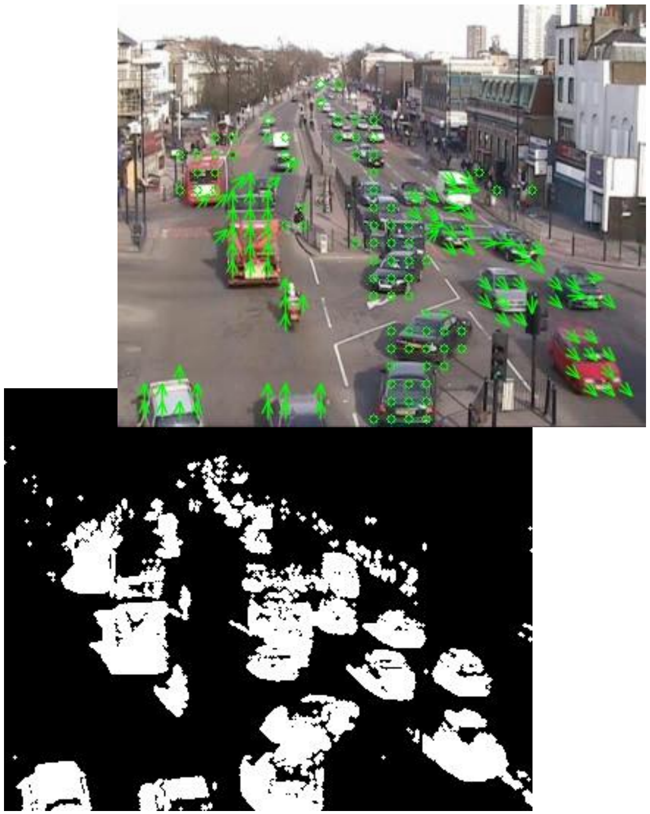
Object-centric methods

- occlusions (broken trajectories)
- several targets (curse of dimensionality)

Not reliable!

Non-object-centric methods

Non-object centric methods in a nutshell



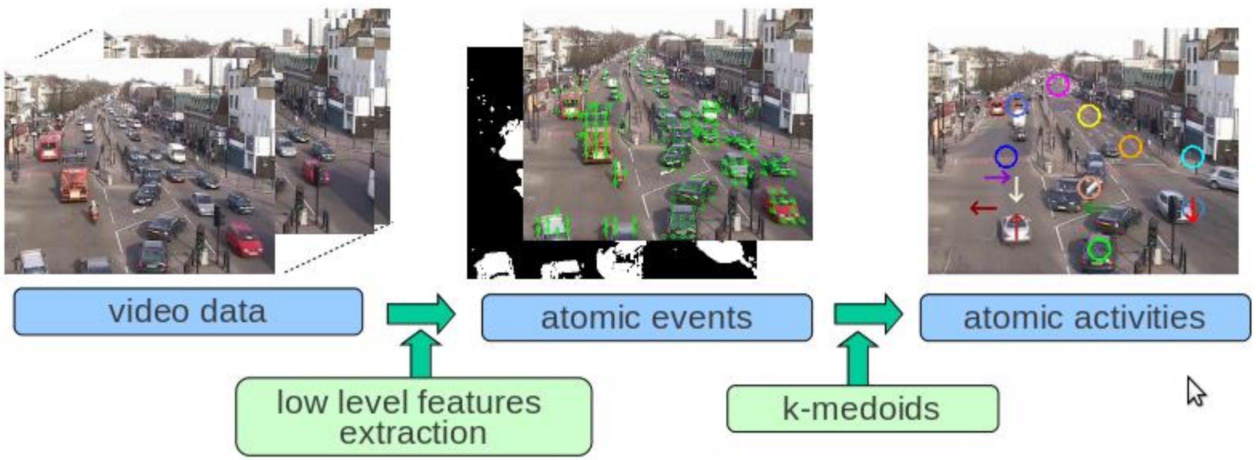
- **Low level cues** (optical flow, foreground) are extracted and quantized (position and motion) to generate **visual words.**

- Short video clips are represented as **visual documents.**

- **Salient activities** (**topics**) of the scenes are extracted based on Probabilistic Topic Models (PTMs) [Kuettel10, Varadarajan10, Hospedales11].

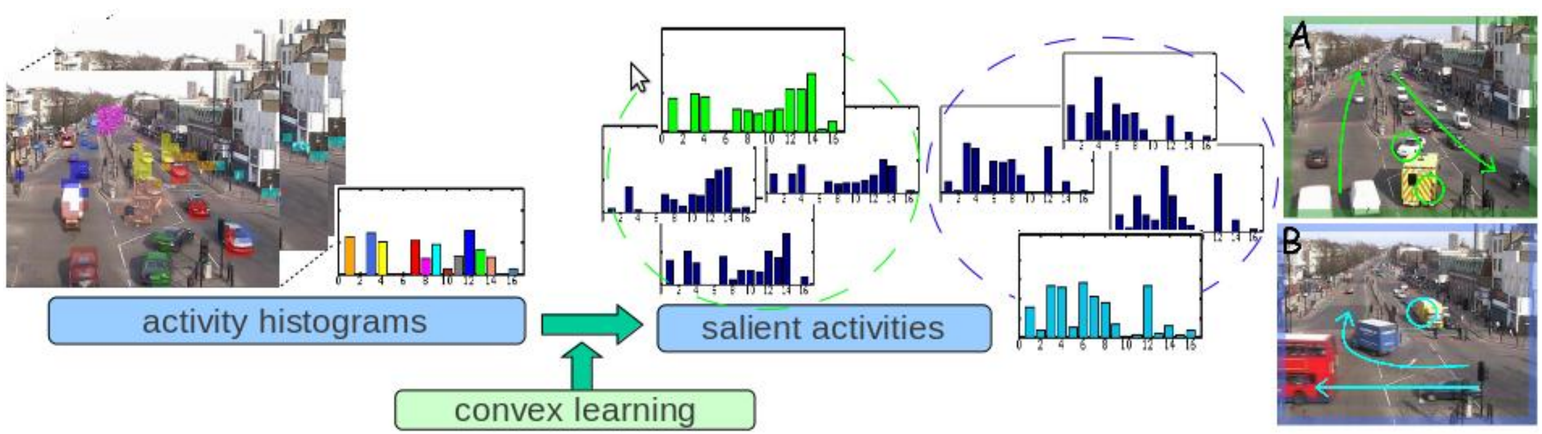
Dependencies between atomic activities (words) are not considered!

Our approach



Step A
detect atomic activities
(codebook generation)

Step B
(*main contribution*)
extract patterns
of typical activities



Main contributions

- The task of extracting typical activities is formulated as a **simple linear programming (LP) problem**.

- The **similarity between atomic activities** is considered by adopting a variation of the Earth Mover Distance (EMD) as distance measure between histograms.

- Anomalous patterns are detected by comparing salient activities extracted at **multiple scales.**

1/3: Convex prototype learning

Given a set of histograms:

$$\mathcal{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\}, \ \boldsymbol{h}_i \in I\!R^D$$

We aim to learn *N* representative prototypes:

$$\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\}, \ \boldsymbol{p}_i \in I\!R^D$$

This task is formalized as a convex optimization problem:

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

Loss: similarity between prototype $\boldsymbol{p}_i$ and associated histogram $\boldsymbol{h}_i$

Regularization: smoothness among neighboring prototypes ($\boldsymbol{p}_i$ is fused into $\boldsymbol{p}_j$)

η_{ij} = {0, 1} indicates prototype's neighborhood

Temporal segmentation: η_{ij} based on a temporal distance criterion

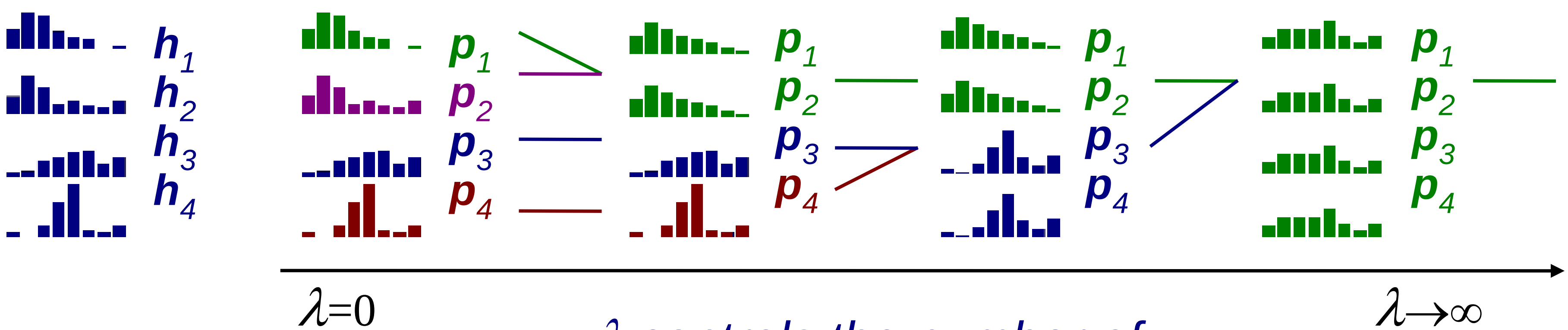
Clustering: η_{ij} based on the distance between histograms' values

1/3: Convex prototype learning

Given the objective function:

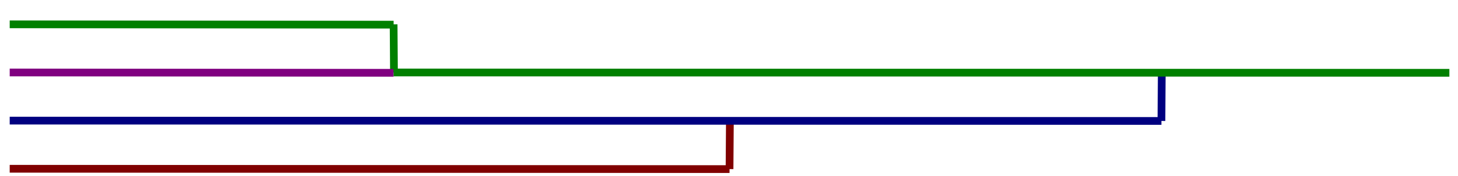
$$\min_{\boldsymbol{p}_i \in \boldsymbol{\Omega}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{h}_i, \boldsymbol{p}_i) + \lambda \sum_{i \neq j} \eta_{ij} \max_{q=1 \ldots D} |p_i^q - p_j^q|$$

What happens in practice:



λ controls the number of different prototypes!

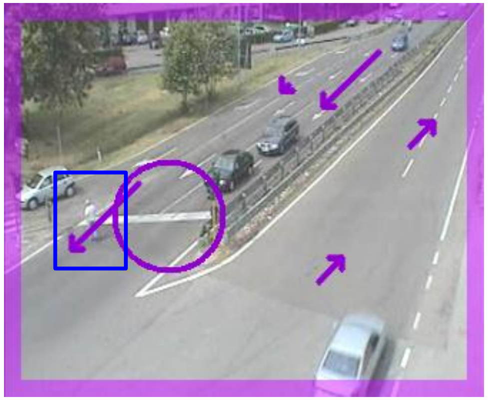
Resulting dendrogram:



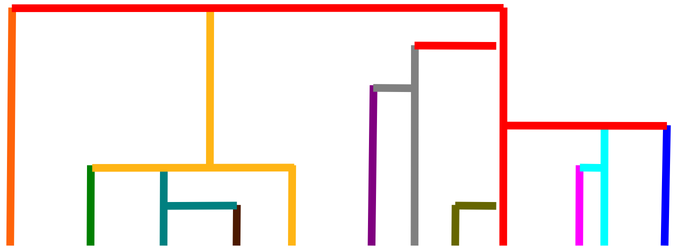
2/3: Multi-scale analysis

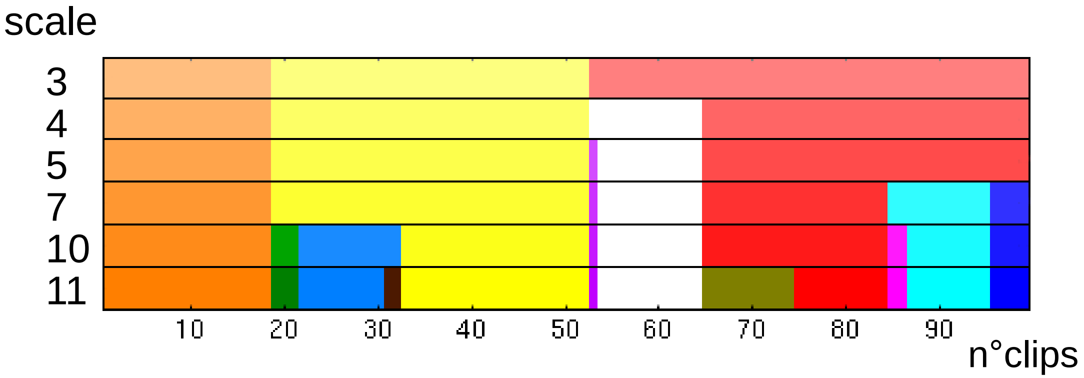
- Comparing clustering results at multiple scales we can detect unusual behaviors corresponding to atypical histograms.

- A high **Multiscale Anomaly Score** (**MAS**) is assigned to small clusters which persist (do not fuse) along the multi-scale analysis.

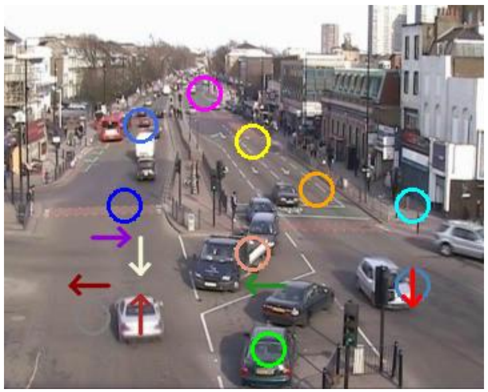


- Detection of a jaywalker

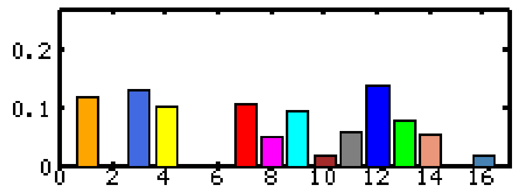




3/3: Correlation among activities



Atomic activities



Histogram of activities for corresponding clip

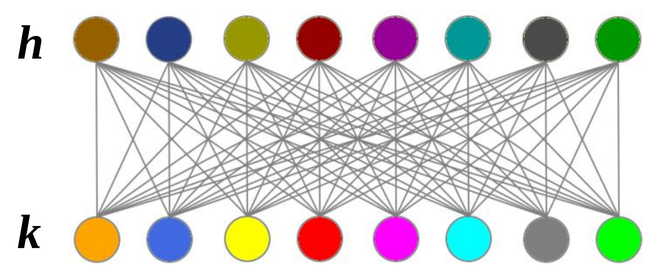
We want to consider **similarity among activities** when comparing clip histograms

A *cross-bin* (Earth Mover Distance) instead of a simple *bin-to-bin* distance is adopted:

$$EMD(\boldsymbol{h},\boldsymbol{k}) = \min_{f_{qt}\geq 0} \sum_{q,t=1}^{D} d_{qt} f_{qt} \qquad \text{s.t.} \sum_{q=1}^{D} f_{qt} = h^t, \sum_{t=1}^{D} f_{qt} = k^q$$

f_{qt} : ***amount of flow*** we want to transfer from bin *q* to *t*.

d_{qt} : ***ground distance,*** encodes similarity among activities.



3/3: Earth Mover's Prototypes

1) We adopt the ***EMD*** in the Loss function:

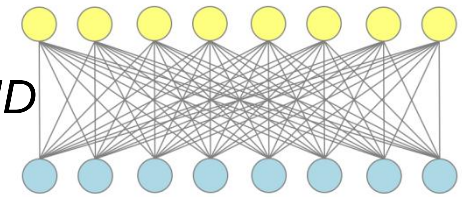
$$\min_{\boldsymbol{p}_i \in \boldsymbol{\Omega}} \sum_{i=1}^{N} EMD(\boldsymbol{h}_i, \boldsymbol{p}_i) + \lambda \sum_{i \neq j} \eta_{ij} \max_{q=1\ldots D} |p_i^q - p_j^q|$$

Complexity: $O(D^2)$. This is computationally expensive...

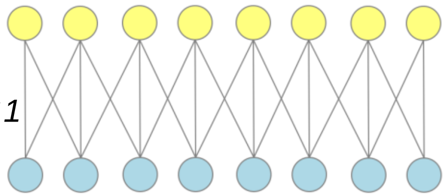
2) An efficient variation of EMD (***EMD-L_1***) is adopted [Ling06], with *L_1* distance over bins as ground distance:

$$d_{qt} = |q - t|$$

EMD



→ *EMD-L_1*



Complexity:
reduced from $O(D^2)$ to $O(D)$!

***Sorting**: similar activities must correspond to neighboring bins in the histogram!*

3) A bin-to-bin distance (***L_1***) is also considered for performance evaluation.

Complexity: $O(D)$

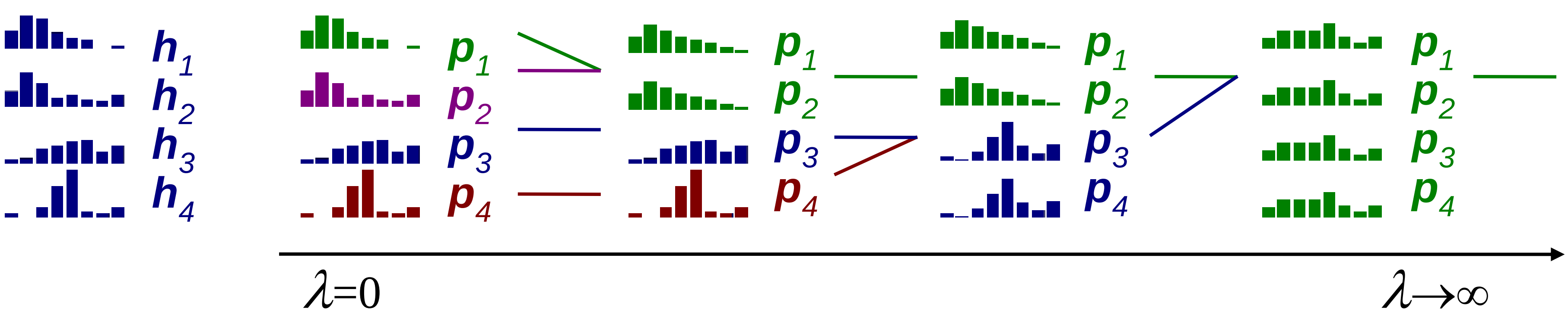
3/3 Earth Mover's Prototypes

The overall optimization problem is a parametric LP:

$$\min_{\boldsymbol{p}_i,\ f_{qt}^i \geq 0} \quad \sum_{i=1}^{N} \sum_{q,t=1}^{D} d_{qt} f_{qt}^i + \lambda \sum_{i \neq j} \eta_{ij} \max_{q=1\ldots D} |p_i^q - p_j^q|$$

$$\text{s.t.} \quad \sum_{q=1}^{D} f_{qt}^i = h_t^i, \quad \sum_{t=1}^{D} f_{qt}^i = p_q^i$$

A variant of the revised simplex method can be used to compute the entire regularization path:



Results: datasets

We tested our method on 4 datasets (3 of them publicly available):

Traffic	
public	no
n°frames	6000
fps	12
frame size	276x336

Junction[1]	
public	yes
n°frames	90000
fps	25
frame size	288x360

Roundabout[1]	
public	yes
n°frames	93500
fps	25
frame size	288x360

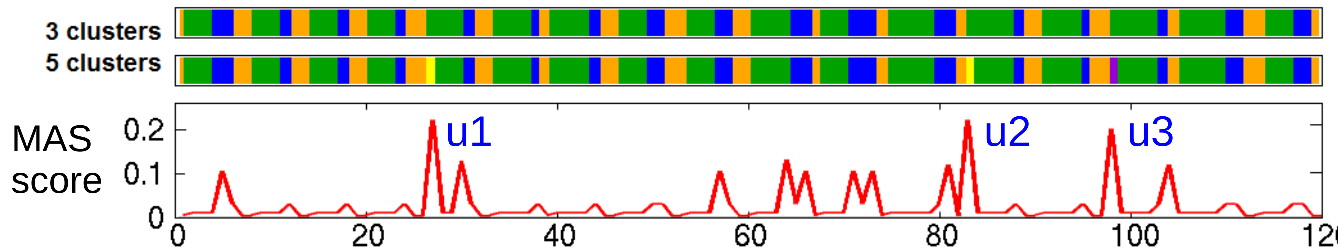
Basket - APIDIS[2]	
public	yes
n°frames	6000
fps	23
frame size	320x368

[1] QMUL dataset, available on www.eecs.qmul.ac.uk/~jianli/

[2] APIDIS dataset, available on www.apidis.org/Dataset/

Results: multiscale analysis

Junction dataset	
n° clips	120
cliplen	375
n° activities	16
Tot video duration	30 min



Vertical

Horizontal ←

Horizontal →

- Three main traffic flows


u1: Jaywalker


u2: Fire engine


u3: Heavy traffic

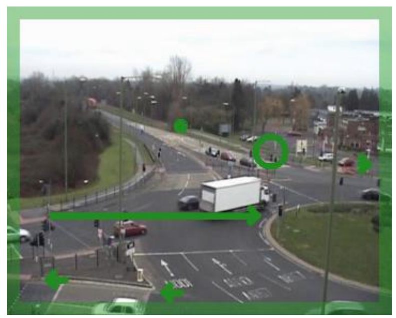
- Unusual events

video: junction.avi

Results: clustering

- Salient activities discovered

Vertical flow



Horizontal flow

video:
roundabout.avi

Roundabout dataset	
n°clips	148
cliplen	300
n°activities	16
Tot video duration	30 min

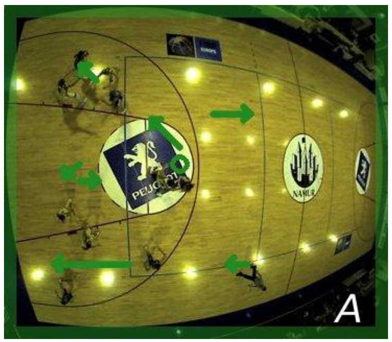
- Accuracy (ground truth[1])

	EMD-L_1	L_1	EMD-L_1 random	Standard pLSA [Li08bmvc]	Hierarchical pLSA [Li08bmvc]
Junction	**92.36**	89.74	86.7	89.74	76.92
Roundabout	**86.40**	**86.40**	72.3	84.46	72.30

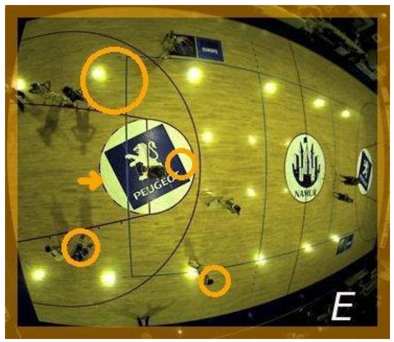
[1] [Li08bmvc] J. Li, S. Gong, and T. Xiang. *Global behaviour inference using probabilistic latent semantic analysis*. BMVC, 2008

Results: basket dataset

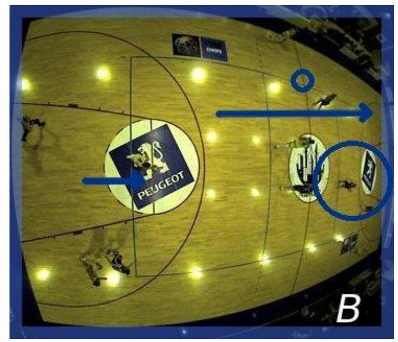
- Five salient activities discovered



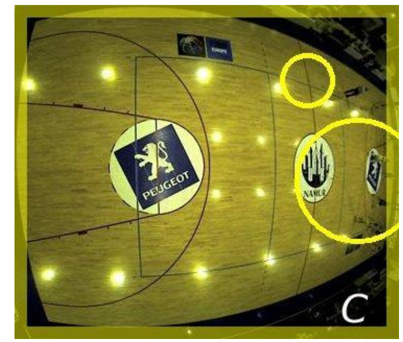
Blue team on attack



Blue team in a free-throw



Towards blue team's court

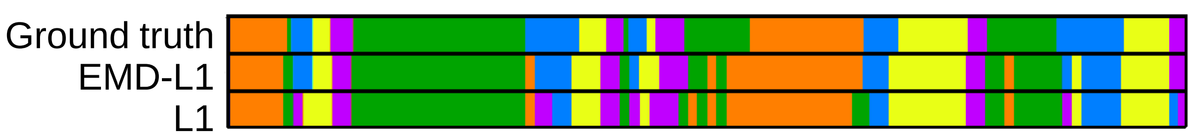


Yellow team on attack

Towards yellow team's court

- Accuracy

video: basket.avi



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

Basket dataset	
n°clips	100
cliplen	60
n°activities	16
Tot video duration	5 min

Conclusions

- Our approach has shown to be effective for **multiscale analysis** of complex video scenes. It relies on a **convex optimization problem**.

- Up to our knowledge this is the first work which considers the **similarity between the atomic activities**.

- A variant of the EMD allows to embed this information at a feasible cost, while the sorting of atomic activities in the histogram becomes crucial for good performance.

- Future work will focus on improving scalability (ad hoc solver needed) and learning of temporal rules.

Data and code will be available on: *www.disi.unitn.it/~zen*

References

[Hospedales11] T. Hospedales, J. Li, S. Gong, T. Xiang. *Identifying Rare and Subtle Behaviours: A Weakly Supervised Joint Topic Model*. IEEE Trans. on PAMI, 2011

[Kuettel10] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. *What's going on? Discovering spatio-temporal dependencies in dynamic scenes*. CVPR, 2010.

[Varad10] J. Varadarajan, R. Emonet, and J.-M. Odobez. *Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes*. BMVC, 2010

[Hospedales09] T. Hospedales, S. Gong, and T. Xiang. *A markov clustering topic model for mining behaviour in video*. ICCV, 2009.

[Yang09] Y. Yang, J. Liu, and M. Shah. *Video scene understanding using multi-scale analysis*. ICCV, 2009.

[Li08] J. Li, S. Gong, and T. Xiang. *Scene segmentation for behaviour correlation*. ECCV, 2008 .

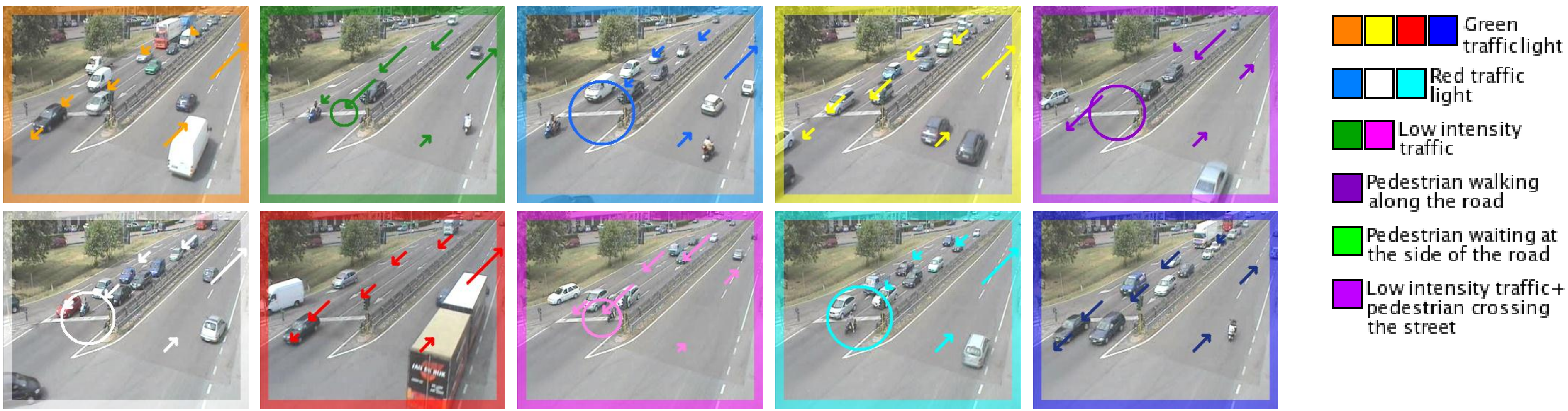
[Li08b] J. Li, S. Gong, and T. Xiang. *Global behaviour inference using probabilistic latent semantic analysis*. BMVC, 2008.

[Ling06] H. Ling and K. Okada. *An efficient Earth Mover's Distance algorithm for robust histogram comparison*. IEEE Trans. on PAMI, 2006

Results: temporal segmentation

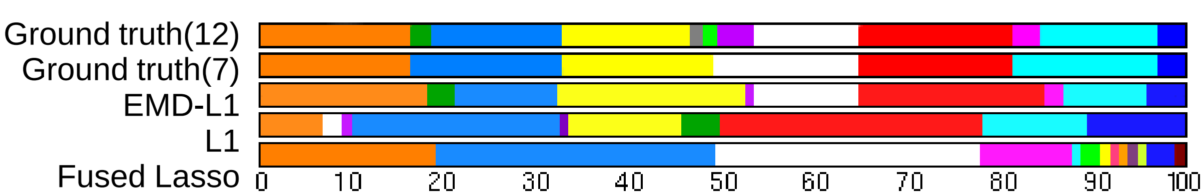
➢ Traffic dataset

- Salient activities



video: traffic.avi

- Accuracy



EMD-L_1	L_1	Fused Lasso
82.4	72.5	68.7