Temporal Poselets for Collective Activity Detection and Recognition

Moin Nabi Alessio Del Bue Vittorio Murino

Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia (IIT)

Introduction on Group Activity Analysis

Detection and recognition of actitivities in the wild, some example:

Clutter, crowd

Dynamic scenes

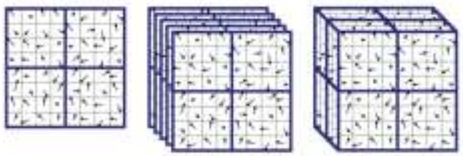
Camera view change

Descriptors for Activity Recognition

Feature-based methods

- 3D-SIFT
- extended SURF
- HOG3D
- STIP
- Cuboid detector and more…

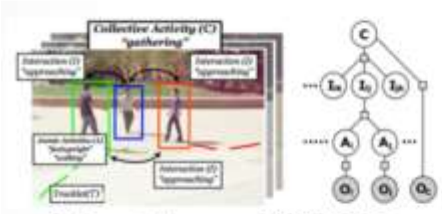
H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In BMVC 2009.



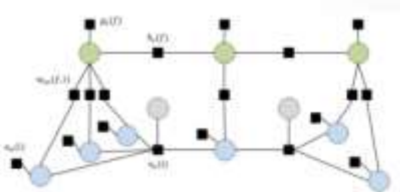
People-based methods

- Spatio-temporal local (STL)
- Action Context (AC)
- The Randomized Spatio-Temporal Volume (RSTV)
- Choi and Savarese, ECCV 2012
- Khamis et al., ECCV 2012

J. Aggarwal and M. Ryoo. Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3):16, 2011.



Choi and Savarese, ECCV 2012



Khamis et al., ECCV 2012

A new descriptor for activities

Properties of **feature-based** methods for Activity Analysis:

- They are general purpose descriptors and they work very well even in the presence of clutter, i.e. crowded scenes.
- They have a tendency to model general motion in the scene (i.e. foreground and background) and they do not discriminate if the temporal information is related to human activities.

Properties of **people-based** methods for Activity Analysis:

- They contain information with a high semantic meaning (context of the area and people detection)
- In clutter or crowded environments their performance is highly diminished.

Is there a **mid-representation** between low-level and high-level features?

Feature based → **?** → **People based**

Temporal Poselet Descriptor (TPOS)

Is there a **mid-representation** between low-level and high-level features?

Feature based → **TPOS** → **People based**

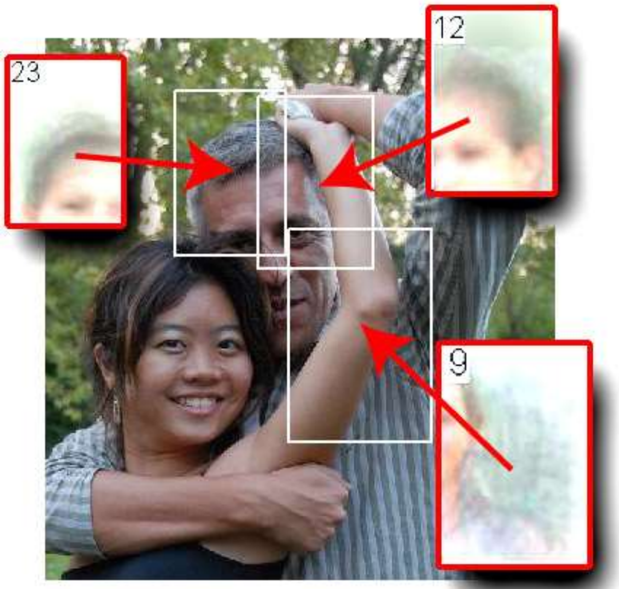
Properties of the **Temporal Poselet Descriptor** for Activity Analysis:

- They are general purpose descriptors and they work very well even in the presence of clutter, i.e. crowded scenes.
- They contain information with a high semantic meaning

TPOS is designed to model semantically meaningful body parts and their motion using **poselets activations in time.**

What is a Poselet?

Poselets are a bank of detectors that respond to a part of the pose of a person from a given viewpoint

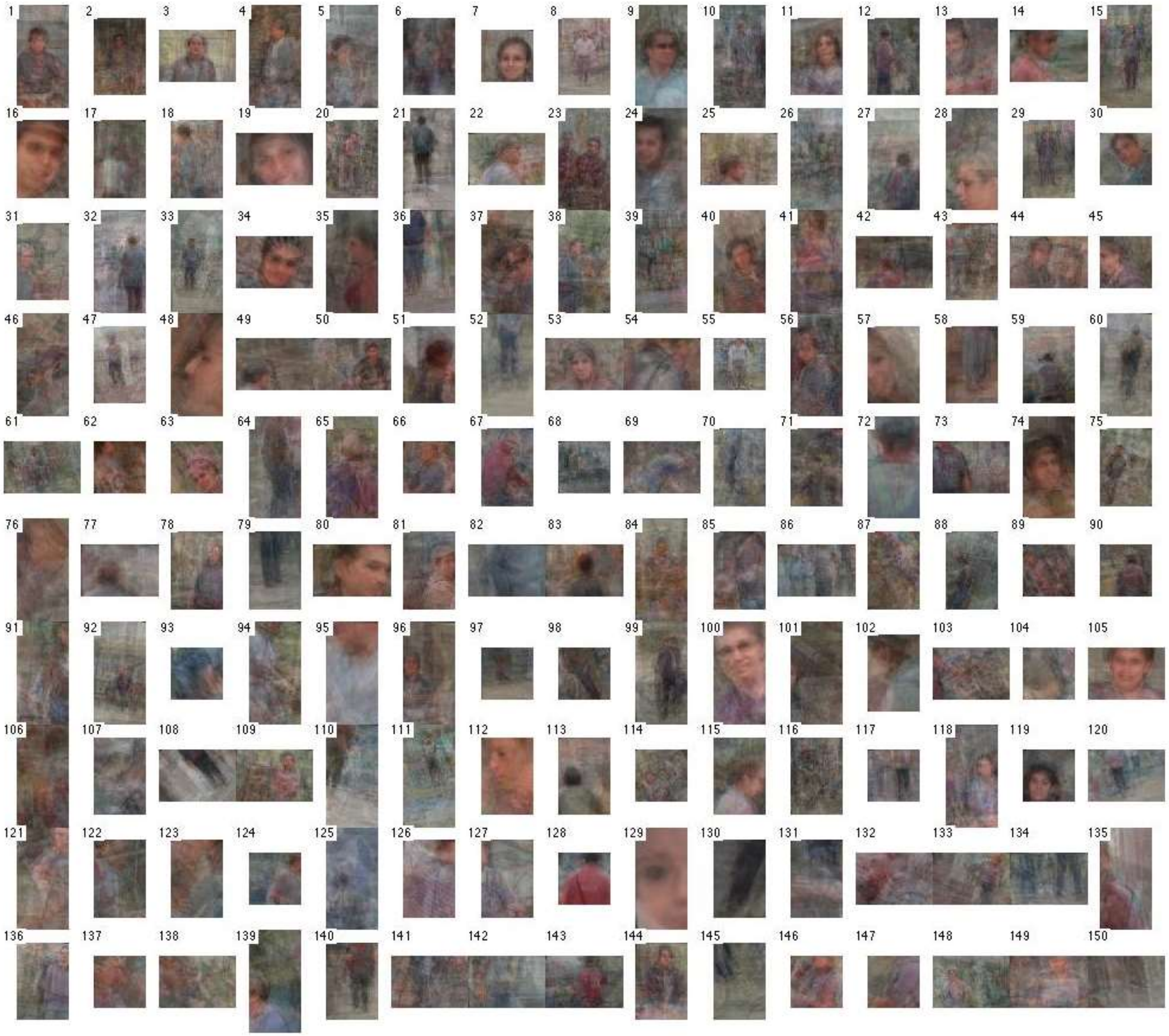


Poselets strongest activations are likely to localized in specific body parts

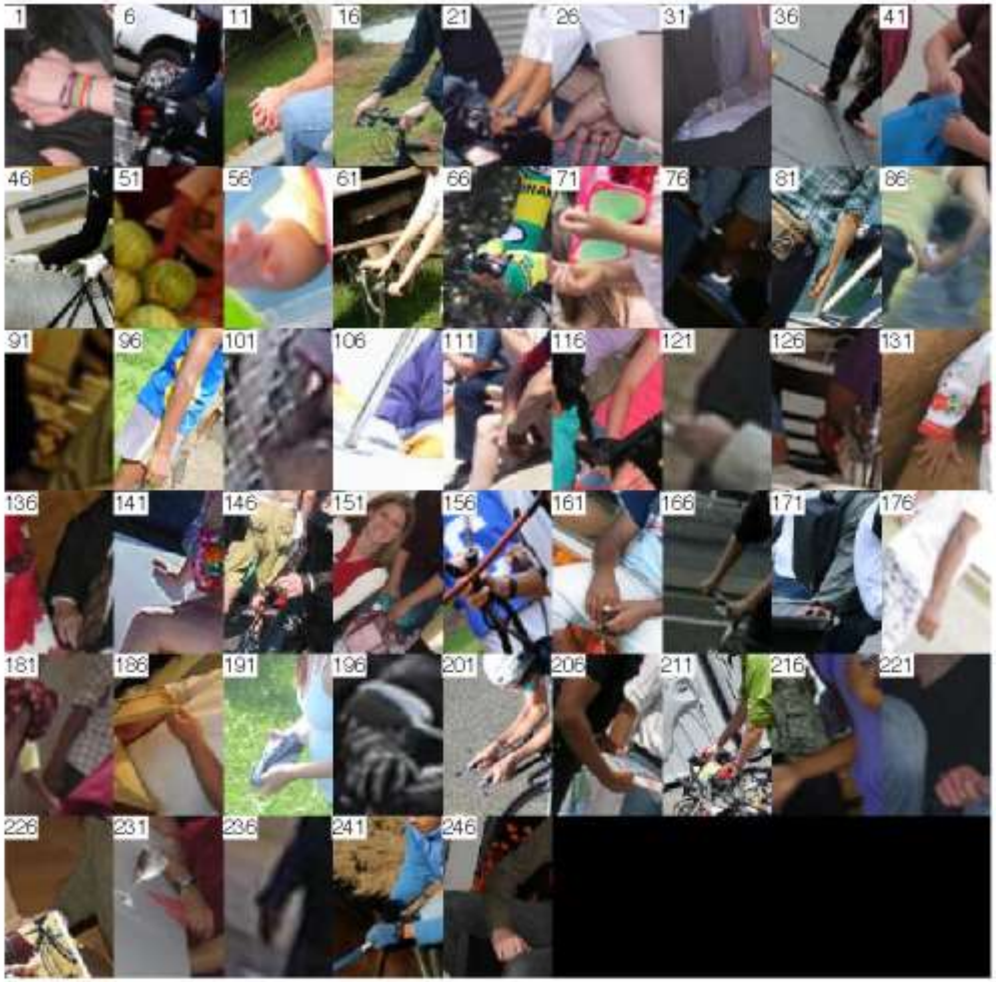
Poselets are parts that are tightly clustered in both appearance and configuration space

[Bourdev & Malik, ICCV09]

150 Poselets



Poselets Details



POSELET #107

Poselets Details

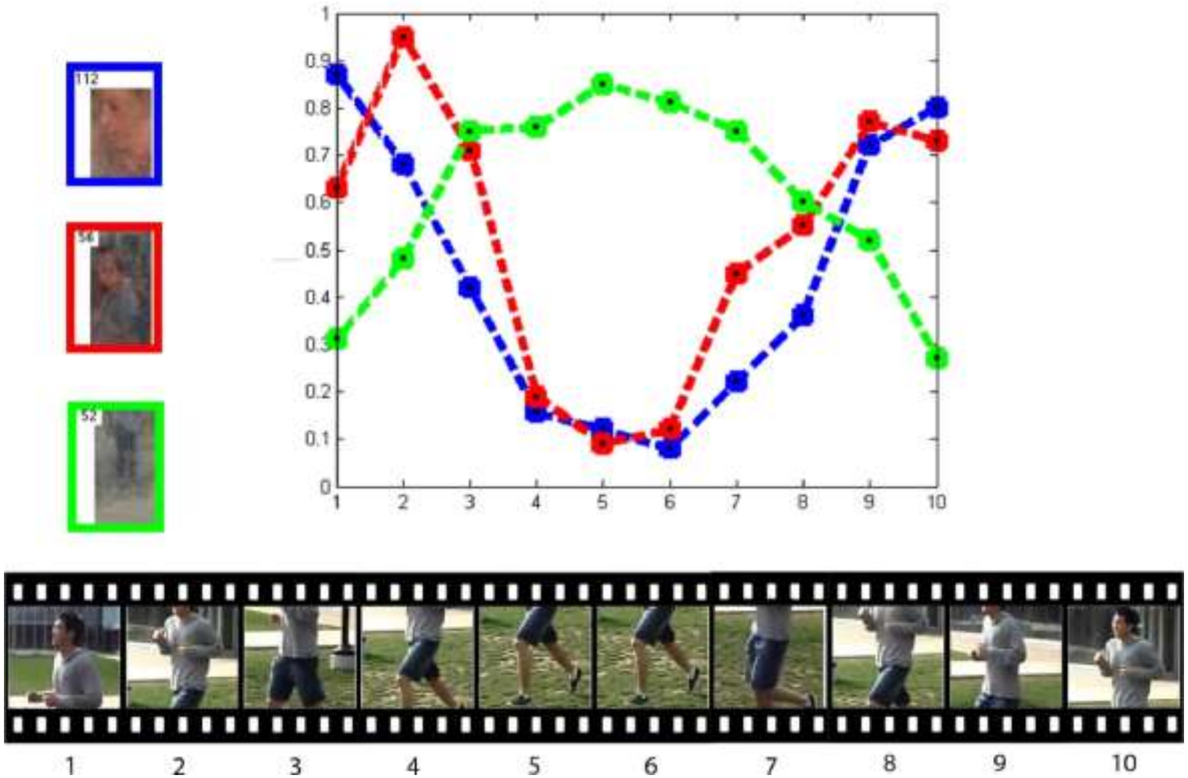
POSELET #95

Poselets Details

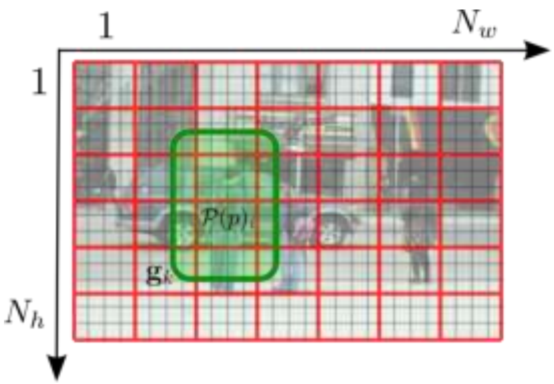
POSELET #130

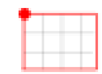
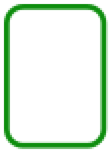
Poselets Activation in time

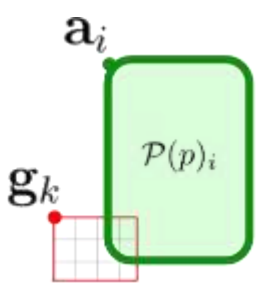
Our approach implies that in time, given a specific action, poselets activations extracted at each frame are correlated.



Measuring Poselets Activations



- An image is partitioned using a regular grid (in red) 
- A poselet *p* with activation *i* (green bounding box) 



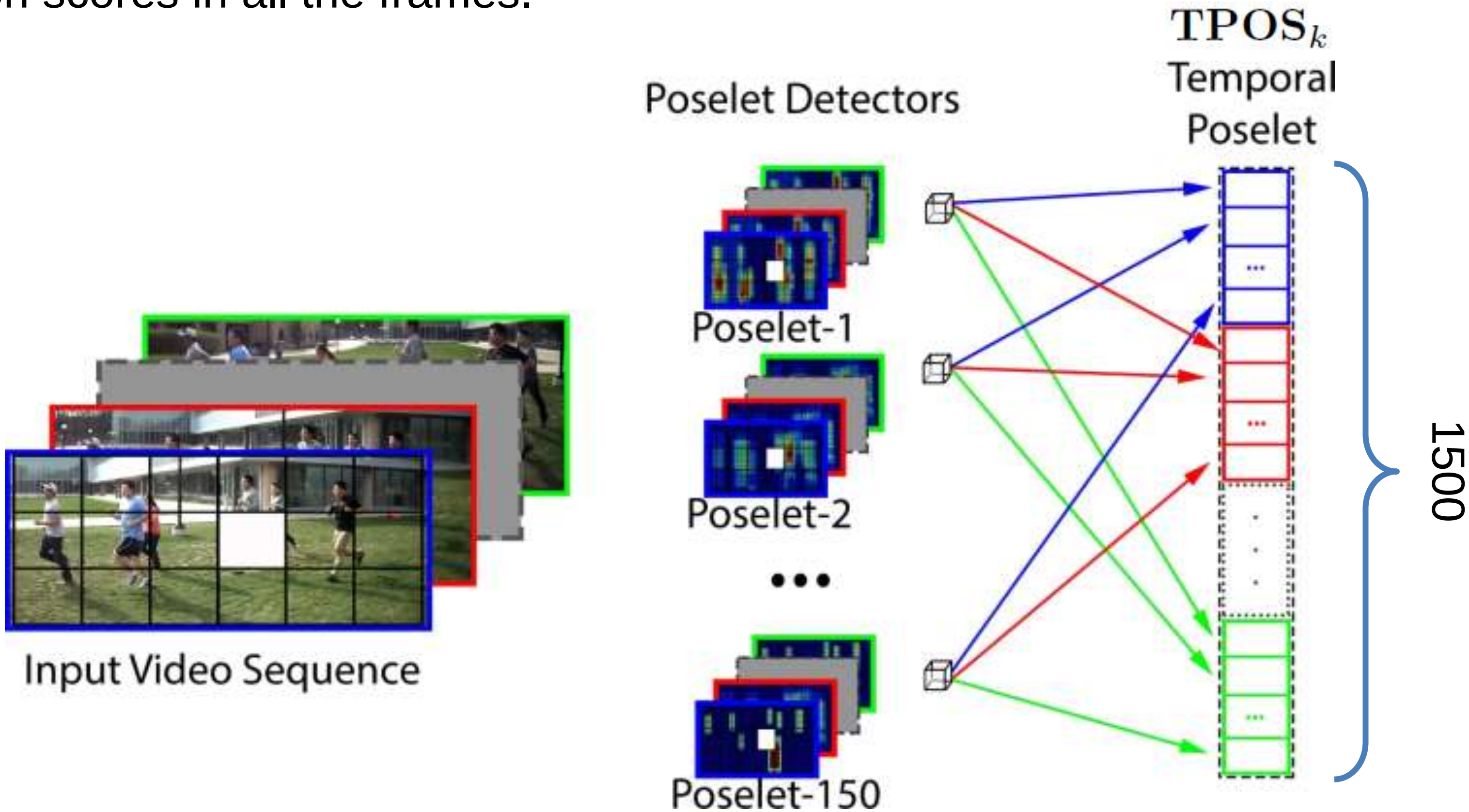
A spatial poselet activation feature is defined as the intersection of the green box and the red cell. Notice that the same poselet activation may intersects and/or include several cell grids in the image.

$$v(p)_{ik} = \frac{area(\mathbf{a}_i \bigcap \mathbf{g}_k)}{area(\mathbf{a}_i \bigcup \mathbf{g}_k)}$$

Similar to Maji et al. CVPR 2010 on single image action classification.

Temporal Poselet (TPOS)

We consider a **video block** of 10 frames and measure each poselet activations in time. The final descriptor is given by the **concatenation** of all the poselets activation scores in all the frames.

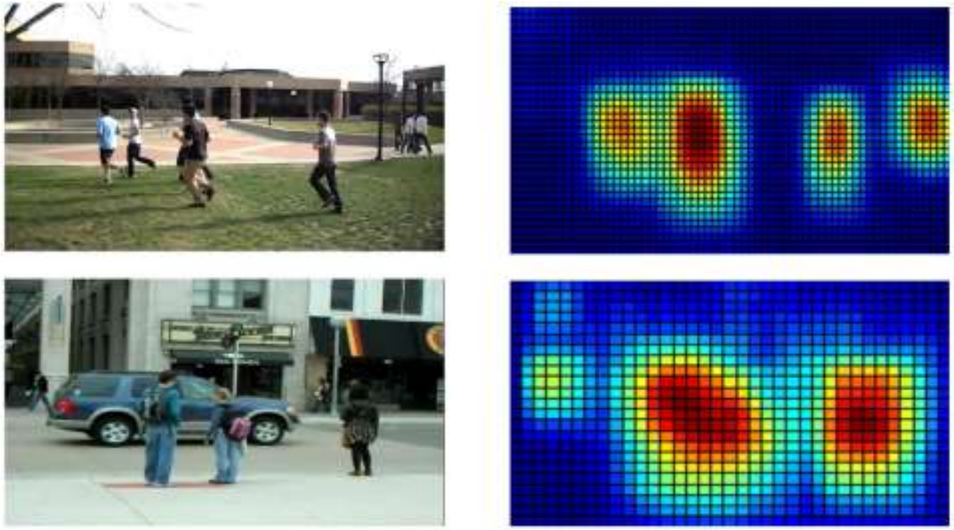


First task: Group Detection

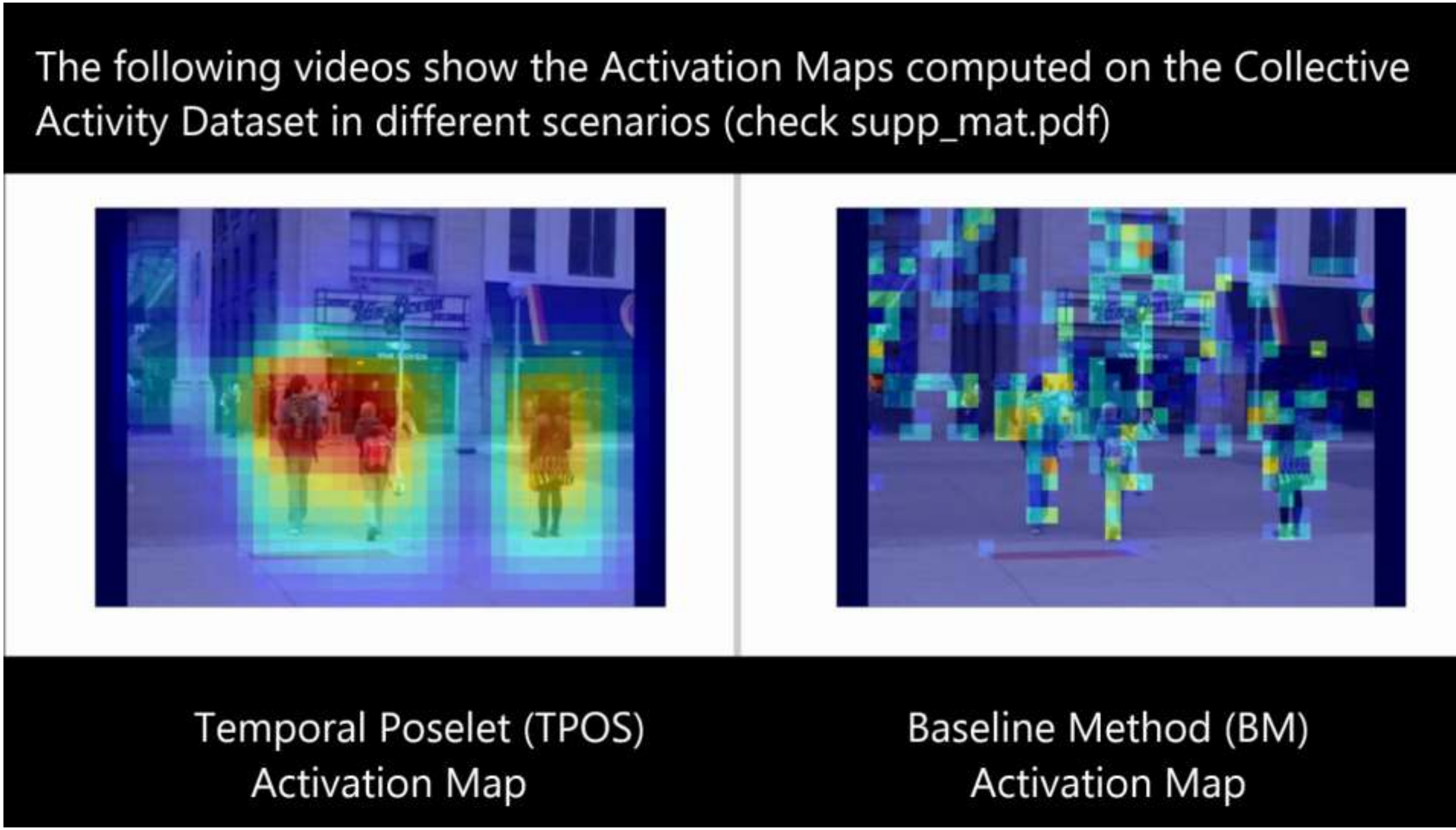
We compute a *saliency measure* that may be used to discard video blocks with few activations:

$$s_k = \|\mathbf{TPOS}_k\|_1$$

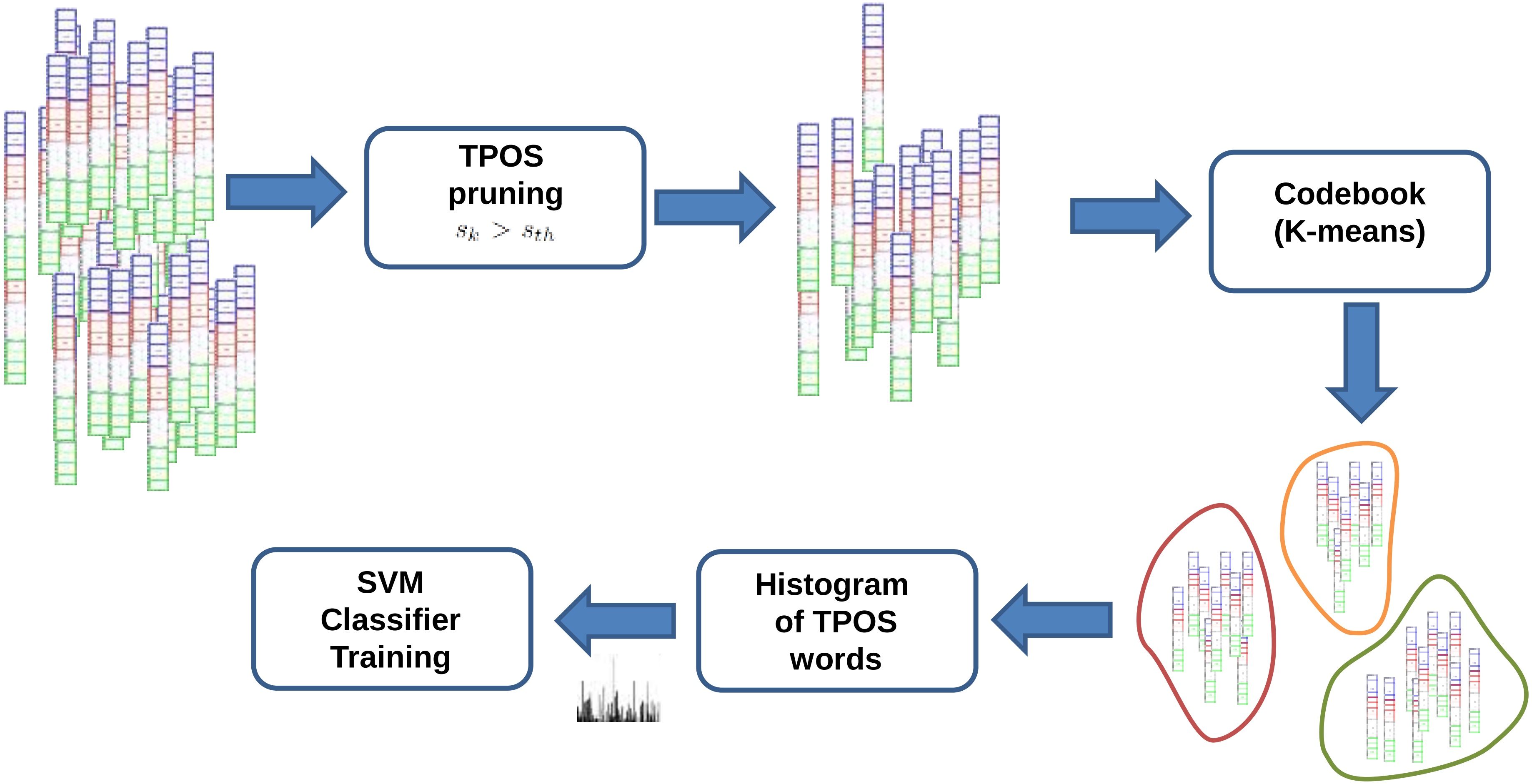
This measure is an indication of the overall activations of a video block



Experimental Results



Second task: Action Recognition



Experimental Results – CAD2/CAD3

CAD2: crossing, waiting, queueing, talking, dancing, jogging.

CAD3: gathering, talking, dismissal, walking together, chasing, queueing

Crossing

Waiting

Queueing

Walking

Talking

Experimental Results – Confusion Matrix

	Crossing	Waiting	Queueing	Talking	Dancing	Jogging
Crossing	47.1%	3.9%	4.4%	6.6%	22.8%	15.2%
Waiting	11.3%	33%	12.2%	12.2%	2.6%	28.7%
Queueing	3.9%	3.3%	63%	7.8%	6.3%	15.7%
Talking	5.4%	1.8%	11.6%	68.8%	2.6%	9.8%
Dancing	5.6%	0%	2.5%	5.6%	83.8%	2.5%
Jogging	6.3%	5.1%	4.4%	0%	3.1%	81.1%

CAD2 Baseline method

	Crossing	Waiting	Queueing	Talking	Dancing	Jogging
Crossing	66.9%	2.9%	0%	6.7%	11%	12.5%
Waiting	5.3%	57.4%	18.3%	13%	4.3%	1.7%
Queueing	3.2%	10.2%	69.3%	11.8%	3.2%	2.3%
Talking	2.7%	8.1%	8.9%	76.8%	2.6%	0.9%
Dancing	3.1%	4.4%	2.5%	3.1%	86.3%	0.6%
Jogging	9.4%	1.9%	0%	2.5%	5.1%	81.1%

CAD2 TPOS method

	Gathering	Talking	Dissmissal	Walking	Chasing	Queueing
Gathering	60%	0%	0%	17.8%	20%	2.2%
Talking	1.5%	70.5%	12.4%	10.1%	0%	5.5%
Dissmissal	0%	37.2%	32.6%	0%	0%	30.2%
Walking	8.2%	16.8%	0%	45.9%	9.2%	19.9%
Chasing	3.7%	0%	0%	35.2%	61.1%	0%
Queueing	3.7%	16%	1.3%	28.4%	3.7%	46.9%

CAD3 Baseline method

	Gathering	Talking	Dissmissal	Walking	Chasing	Queueing
Gathering	47.1%	11.8%	0%	32.4%	8.7%	0%
Talking	0.7%	92.6%	1.2%	5.5%	0%	0%
Dissmissal	0%	33.3%	66.7%	0%	0%	0%
Walking	4.9%	3.9%	0%	83%	1.1%	7.1%
Chasing	2.4%	0%	0%	9.6%	83.3%	4.7%
Queueing	0%	0%	0%	25.2%	13.3%	61.5%

CAD3 TPOS method

Experimental Results – Accuracy

	Base	**TPOS**	RSTV	[9]	[12]	[4]
CAD2	62.8 %	72.9 %	71.7 %	85.7 %	-	-
CAD3	52.8 %	72.3 %	-	-	74.3	79.2%

Average Classification Accuracy

[9] S. Khamis, V. I. Morariu, and L. S. Davis. Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition. In *ECCV 2012*.

[12] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. *NIPS 2010*.

[4] W. Choi and S. Savarese. A Unified Framework for Multi-target Tracking and Collective Activity Recognition. *ECCV 2012*, pages 215–230.

Conclusions and Future Work

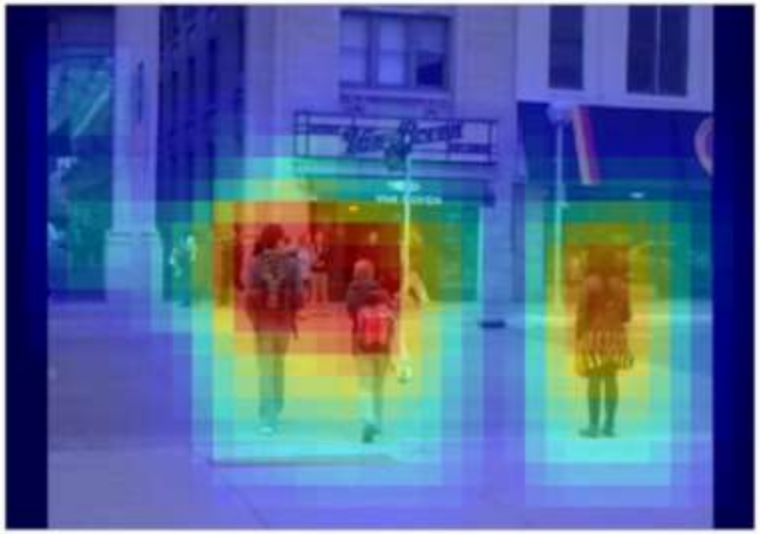
TPOS is a novel descriptor for human activities analysis:

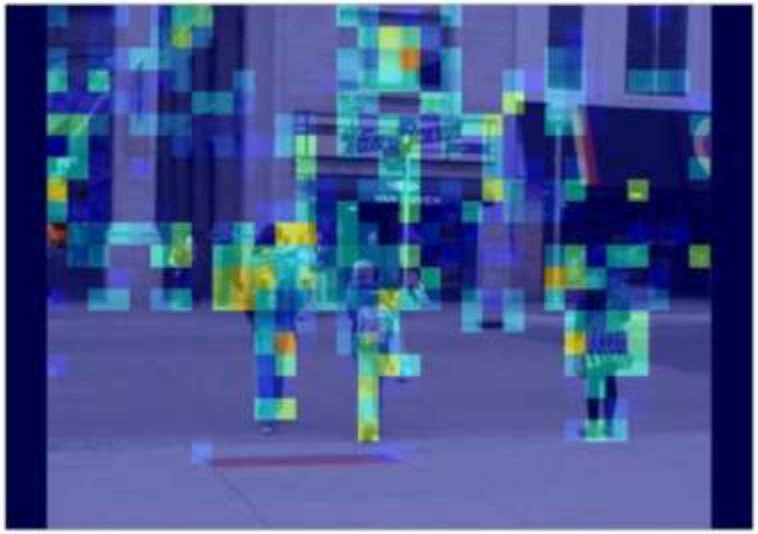
- They are general purpose descriptors and they work very well even in the presence of clutter, i.e. crowded scenes.
- They contain information with a high semantic information about the temporal pose of people in the scene.
- Even without higher-level information (people bounded boxes, tracking information) they are able to obtain reasonable results compared with state of the art approaches.
- Compared to general purpose descriptors, the performance are strongly improved on CAD2/CAD3

Future Work:

- Solve jointly for action segmentation and recognition using TPOS
- Model more deeply the correlation among poselets activation in time

The following videos show the Activation Maps computed on the Collective Activity Dataset in different scenarios (check supp_mat.pdf)





Temporal Poselet (TPOS) Activation Map

Baseline Method (BM) Activation Map