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)
Via Morego 30, Genova, Italy

Abstract

Detection and recognition of collective human activities are important modules of any system devoted to high-level social behavior analysis. In this paper, we present a novel semantic-based spatio-temporal descriptor which can cope with several interacting people at different scales and multiple activities in a video. Our descriptor is suitable for modelling the human motion interaction in crowded environments – the scenario most difficult to analyse because of occlusions. In particular, we extend the Poselet detector approach by defining a descriptor based on Poselet activation patterns over time, named TPOS. *We will show that this descriptor can effectively tackle complex real scenarios allowing to detect humans in the scene, to localize (in space-time) human activities, and perform collective group activity recognition in a joint manner, reaching state-of-the-art results.*

1. Introduction

Understanding human activity is a problem whose solution has a clear impact in modern Computer Vision applications. In general, people activities convey rich information about the social interactions among individuals, the context of a scene, and can also support higher-level reasoning about the ongoing situation. In such regard, much attention has been recently posed in the detection and recognition of specific human activities from images and videos, mainly focusing on crowded scenes. Such task has been formalized in the literature as a classification problem where a label corresponding to a collective activity has to be assigned to a specific video frame or a video clip, possibly also identifying the spatial location in the video sequence where such activities occur.

This paper addresses this open issue and aims at proposing a spatio-temporal descriptor called TPOS, based on Poselets [2]. This descriptor is effective in detecting in space and time *multiple* activities in a *single* video sequence, so providing a semantically meaningful segmentation of the footage, without resorting to elaborated high-level features or complex classification architectures. In the literature, apart from the core classification aspects, a substantial debate has been posed over the type of features which are more discriminative for the activity detection/recognition problem. In this context, two classes of approaches can be identified, which are related to the level of semantics embedded in the descriptor.

On one hand, *feature-based* methods adopt the classical strategy of detecting first a set of low-level spatio-temporal features, followed by the definition of the related descriptors. These descriptors should be representative of the underlying activity and they are typically defined as a spatio-temporal extension of well-known 2D descriptors, such as 3D-SIFT [21], extended SURF [24], or HOG3D [10]. Among the best performing features, we can quote the Laptev's space-time interest points (STIP) [14], the cuboid detector [7] and descriptor based on Gabor filters [3]. A number of other descriptors also deserves to be mentioned like dense trajectories [22], spatial-time gradient [15], optical flow information [7], and Local Trinary Patterns [25]. In [16], an unsupervised spatio-temporal feature learning scheme is proposed which only uses the intensity value of the pixels in an extended ISA (Independent Subspace Analysis) framework. An interesting comparative evaluation was presented in [23], which reports a performance analysis of different combinations of feature detectors and descriptors in a common experimental protocol for a number of different datasets.

On the other hand, a different set of methods, named *people-based* approaches, directly use higher-level features which are highly task-oriented, i.e. they are tuned to deal with people. They rely on a set of video pre-processing algorithms aimed at extracting the scene context, people positions (bounding boxes, trajectories) and head orientations. For instance, the context around each individual is exploited by considering a spatio-temporal local (STL) descriptor extracted from each head pose and the local surrounding area [6]. The Action Context (AC) descriptor is introduced in [11] in a similar way as STL, but it models action context rather than poses. In [12, 13], two new types of contextual information, individual-group interaction and individual-individual interaction are explored in a latent variable framework. The Randomized Spatio- Temporal Volume (RSTV) model, a generalization of STL method, is introduced in [5] by considering the spatio-temporal distribution of crowd context. Khamis et al. [8] presented a framework by solving the problem of multiple target identity maintenance in AC descriptors. Recently, Choi and Savarese [4] put a step forward presenting a more extensive model which estimates collective activity, interactions and atomic activities simultaneously. Finally, [9] introduced a model which combines tracking information and scene cues to improve action classification. The comprehensive survey in [1] reports an excellent overview of recent and past works in this domain.

Both classes of approaches have their expected pros and cons which are more evident for "in the wild" scenarios. In particular, people-based methods highly rely on the accuracy of sophisticated detectors that might fail in the case of dense (i.e., crowded) scenarios because of the presence of occlusions, or due the impossibility to reliably deal with the spatio-temporal correlation among the large number of targets. However, in case of accurate detections and tracking, they have shown to obtain the best performance. This is mainly due to the significant semantic information provided by the descriptors which is tailored to the task of human sensing.
Instead, feature-based approaches utilize low-level spatio-temporal representations regardless the scene complexity, and thus they are less prone to gross misinterpretations as given, for instance, by false positive person detections eventually obtained by people-based methods. However, low-level features lack of an actual semantic interpretation since the identity and the collective human parts peculiar of the interaction are typically disregarded. In other words, such features are extracted at the whole frame level, and not necessarily correspond to high-level descriptions of a spatially and temporally localized activity, or are related to local peculiar parts of the image.

1.1. Discussion and contributions

In order to overcome these limitations, this paper proposes a method to narrow down the gap between feature- and people-based approaches, namely, trying to inject a semantic value to spatio-temporal features, so leading to more powerful discrimination of human activities while maintaining the method complexity to a manageable level. We introduce here a novel semantic descriptor for human activity as a temporal extension of the Poselet approach [2], in which human, semantically meaningful body parts and their motion are modelled by poselets activations in time. In a nutshell, we devised a *temporal poselet* ($\mathbf{TPOS}$) descriptor by analyzing the activation correlation of a bank of (poselet) detectors over time. This provides a video representation composed by joint human detections and activity characterization using the basis of the detection activation patterns of the poselets.

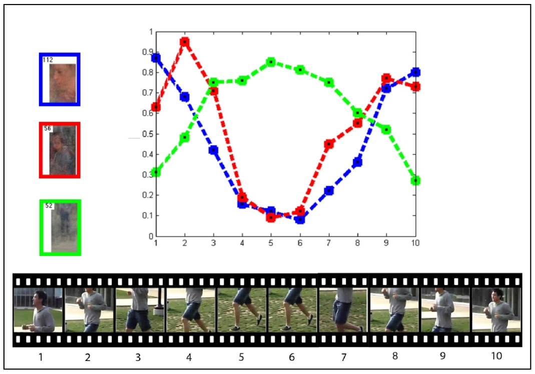


Figure 1. Three types of poselet activations in time in a 10-frame video (bottom): head, torso and legs in the sequence are displayed as blue, red and green profiles, respectively, in a 10-frame sequence, showing the correlation of these types of poselets during the "running" activity.

The underlying idea of the proposed approach is that, since each poselet is activated by a particular static human pose, the collection of poselet activations in time can capture and characterize human motion. In other words, by assuming that people activities can be described as a sequence of peculiar body poses, they can be represented as a sequence of activations of a set of human body part detectors. As an example, in Fig. 1, we show the activations of three different poselets (face, torso and legs) in a 10-frame video block of a video sequence representing a running person. It is worth to note that the specific sequence of appearing/disappearing of the human legs, torso and face in this short video clip is encoded by corresponding activations of the leg, torso and face poselets, respectively. So, the activation profiles of these three poselets suggest that human activity is correlated with temporal profiles of body parts and that can be learned for a discriminative analysis.

Similar to our proposal, in certain (partial) aspects, a few recent works on activity recognition are worth to be mentioned. The closest work to our is [18] where poselet detector activations have been used for activity recognition in still images. The idea is similar to ours, that is to build a robust descriptor for actions considering activation of poselets. In their work the feature level action discrimination was caught by re-training a large dictionary of 1200 action-specific poselets. This set of detectors were trained in the same dataset and tuned for specific action classes. In our case, first we deal with video sequences instead of still images, then, we aim at capturing action discrimination at a representation level and model the pose dynamics of a group of people using the basis pattern of poselet co-occurrence. So, we instead use the activation score of an outsourcing bank of detectors (150 general purpose poselets), as learned in the original formulation [2], and we tested on video datasets, which results in a much more challenging scenario.

Very recently, activities have been modelled as a sparse sequence of discriminative keyframes, that is a collection of partial key-poses of the actors in the scene [19]. Keyframes selection is cast as a structured learning problem, using an *ad hoc* poselet-like representation trained on the same dataset used for testing (still using separate sets), and reinforcing the original poselet set using HOG features with a bag of words (BoW) component. This approach has interesting peculiarities and reaches good performance on the UT-Interaction dataset. Nevertheless, it uses specific descriptors learned from the same dataset and, despite it is claimed to be able to spatially and temporal localize actions and action parts, this is reached only in terms of the keyframes extracted, and it still has the limit to classify an activity per test video sequence.

The main peculiar aspect of our approach lies in the design of a new, yet simple, feature descriptor which results to be expressive for detecting people and to characterize their collective activity. In particular note that our descriptor is not customized (i.e., "tuned") on any specific dataset (for training and testing), but the basic (poselet) detectors are used as given in [2], and tested in completely different datasets. Moreover, our particular formulation potentially allows to deal with different activities in a single video sequence, so promoting its use for segmenting continuous video streaming providing temporal and spatial localization of the semantically meaningful events of the video sequence. These characteristics make TPOS unique in the panorama of the spatio-temporal descriptors for human detection and collective activity recognition. It is also worth to note that single human actions are out of the scope of this work and our approach is devoted to the analysis of groups of people and crowds in time.

We tested the **TPOS** method on two public datasets [5, 4] and our experimental evaluation will show that it can effectively be used for group detection and collective activity recognition in the wild, outperforming the state-of-the-art baseline methods and showing the limitations of the pure feature-based methods.

The remainder of the paper is structured as follows. The temporal poselet approach is presented in Sec. 2 together with its application to the people group detection. Section 3 shows the strength of the semantic representation of the temporal poselets for addressing the group activity classification problem. Experiments in Section 4 will evaluate the proposed approach on two different datasets [5, 4]. Finally Section 5 will draw the direction for future work and further application domains.

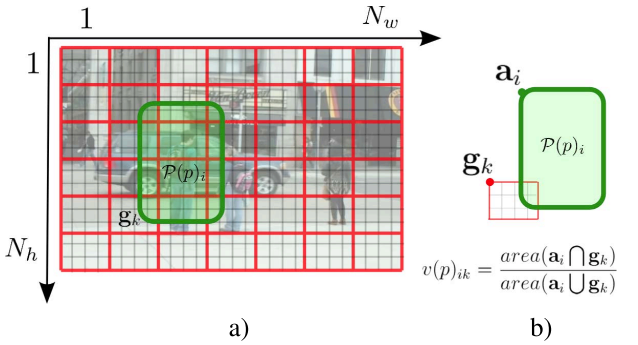

1
N_w
1
N_h
$\mathcal{P}(p)_i$
$\mathbf{g}_k$
a)
$\mathbf{a}_i$
$\mathcal{P}(p)_i$
$\mathbf{g}_k$
$v(p)_{ik} = \frac{area(\mathbf{a}_i \bigcap \mathbf{g}_k)}{area(\mathbf{a}_i \bigcup \mathbf{g}_k)}$
b)

Figure 2. a) The image is partitioned using a regular grid (in red) where each cell grid is defined by $\mathcal{G}_j$ containing the 2D position and the cell grid size. A poselet p with activation i (in green) is defined in the same manner with $\mathcal{P}(p)_i$. Notice here that the same poselet activation may intersects and/or include several cell grids in the image. b) The activation for a cell g_k is given by the intersection between the poselet $\mathcal{P}(p)_i$ bounding box and the grid.

2. Temporal poselets for group activity recognition

Our final goal is to detect the collective activity of a group of people in weakly labelled videos. In this section, we will formalize our poselet-based temporal descriptor to be used for two applications, group detection and collective activity recognition.

2.1. Poselet-based video representation

We first derive the descriptor in 2D and then we follow with the extension in time. A generic image frame is first partitioned in a set of $N_h \times N_w$ grid cells. the overall set of image cell is represented by the set $\mathcal{G}$ such that:

$$\mathcal{G} = \{\mathbf{g}_k\}_{k=1}^{|\mathcal{G}|} \quad \text{and} \quad \mathbf{g}_k = \{\mathbf{z}_k, w, h\},$$

where $\mathbf{z}_k \in \mathbb{N}^2$ represents the coordinates of the top-left corner of the grid cell, w and h the constant grid cell width and height, respectively and $|\mathcal{G}| = N_h \times N_w$. Following this notation, a given cell grid k is fully specified by $\mathbf{g}_k$.

Given an image $\mathcal{I}_f$ we run P poselet detectors as a filter bank. This provides the location of the detected poselets together with their bounding box size.In particular, for each poselet detector p with $p = 1 \ \ \ldots \ P$, we obtain the set of poselet detection $\mathcal{P}$ such that:

$$\mathcal{P}(p) = \{\mathbf{a}_i\}_{i=1}^{|\mathcal{P}(p)|} = \{\mathbf{l}_i, w_i, h_i, c_i\}_{i=1}^{|\mathcal{P}(p)|}$$

where $\mathbf{l}_i \in \mathbb{N}^2$ represents the coordinates of the top-left corner of the poselet bounding box, w_i and h_i gives the bounding box width and height for the detection i respectively and finally c_i is the poselet detection score. As defined before, the activation i of poselet p is fully defined by $\mathcal{P}(p)_i = \mathbf{a}_i$. Notice here that in general a poselet detection $\mathcal{P}(p)_i$ may include several cell grids of the set $\mathcal{G}$ (see Figure 2a).

Now we need to define for each cell grid g_k of the image which type of poselets p has been activated. To this end, we define a *spatial poselet activation feature* $v(p)_{ik}$ as the ratio between the areas of the intersection and union of the bounding boxes, i.e.:

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

where the operators $\bigcap$ and $\bigcup$ define the intersection and union respectively of the image windows i, k whose area is given by the function $area(\,\cdot\,)$. This ratio indicates if $\mathbf{a}_i$, the i^{th} activation of the poselet p includes the image cell $\mathbf{g}_k$. By considering all the possible activations i of poselet p we can define the *spatial poselet feature* over all the possible activations as:

$$v(p)_k = \sum_{i=1}^{|\mathcal{P}(p)|} c_i \, v(p)_{ik}$$

that provides an indication of the persistence of a particular poselet weighted by the score c_i in a given grid of the image. We finally characterize the overall poselet activation by defining the *spatial poselets vector* $\mathbf{v}_k \in \Re^P$ for the image cell j as:

$$\mathbf{v}_k = \left[\begin{array}{cccc} v(1)_k, & v(2)_k, & \ldots, & v(P)_k \end{array} \right]^\top .$$

This vector has an entry for each poselet type which reflects the degree to which a particular poselet is activated in that area.

In order to consider the dynamics of the scene, we extend the previously defined spatial descriptor with a temporal component. This is achieved by further dividing the video sequence in a set of $(N_h \times N_w \times N_t)$ video blocks where the length of each block in time is normally $T = 10$ frames. Such frame length was optimised in [16] and [23]. Given each image frame in the T-frame video clip, we can define for each cell grid $\mathbf{g}_k$ over time t (i.e., the video block) a set of T *spatial poselets vectors* $\mathbf{v}_{k,t}$ for $t = 1 \ldots T$. Thus we define the *temporal poselet* descriptor $\mathbf{TPOS}_k$ for the video block k as the concatenation of all the *spatial poselet vectors* such that:

$$\mathbf{TPOS}_k = \left[\begin{array}{cccc} \mathbf{v}_{k,1}^\top, & \mathbf{v}_{k,2}^\top, & \ldots, & \mathbf{v}_{k,T}^\top \end{array} \right]^\top . \quad (1)$$

The vector $\mathbf{TPOS}_k \in \Re^{TP}$ is including the activations of all the poselets in a particular video-block in space and time and captures human motion regularities in the crowd by analyzing the statistics of their poses given the poselet activations. It also embeds the information of the activations of different poselets in time and not only in space (see Fig. 3 for a graphical description).

This semantic description of the scene helps us toward having a better understanding about the functionality of each region based on the implicit human pose in that region.

2.2. Video representation and group detection

The temporal poselet vectors defined in Eq. (1) are the basic building blocks for creating a people-based representation of a video sequence. In practice, the descriptor here defined is a powerful cue for detecting human groupings in unconstrained scenes. However, as the poselet detectors are subject to false positives, some activations might be noisy or not consistent with the human activity. For this reason we define a saliency measure that discards video blocks with few activations.In practice we define the saliency of a temporal poselet as the sum of the elements of $\mathbf{TPOS}_k$ giving:

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

This measure is an indication of the overall activations of a specific video block and it will be also used for higher-level tasks such as group activity recognition in order to obtain fewer examples for training and testing. We select this measure because, it exploits the strength of the spatial activations as well as the poselet scores.

Such saliency can also be used directly to provide a powerful cue for people detection in unconstrained scenes. Given s_k for all the cell grids, it is possible to graphically visualise the saliency measure as in Fig. 4. The resulting map, overlayed over the images in Fig. 4c, is called Activation Map, and it shows that activations are more predominant where the major density of people are present. It is also interesting to show a comparison between the information implicit in the $\mathbf{TPOS}$ descriptor with respect to a general purpose descriptor such as STIP [14]. For instance, common spatio-temporal descriptors are responsive to motion, regardless if the cause of the movement is given by a car or a pedestrian as visible in the second row of Fig. 4d. Also note that in the video clips in the first and third rows of the figure the camera is shaking, and this create a relevant amount of noise for a standard spatio-temporal descriptor. The Activation Map given by the temporal poselets instead only provides the location of the human motion. For similar reasons, $\mathbf{TPOS}$ does not manage well single person activity as poselet detector responses are sensitive to the density of the scene. Actually, $\mathbf{TPOS}$ being a high-dimensional descriptor is very sparse if the scene is scarcely populated. Such sparsity can be found in other high-level descriptors like those presented in Object-bank [17] and Action-bank [20] works.

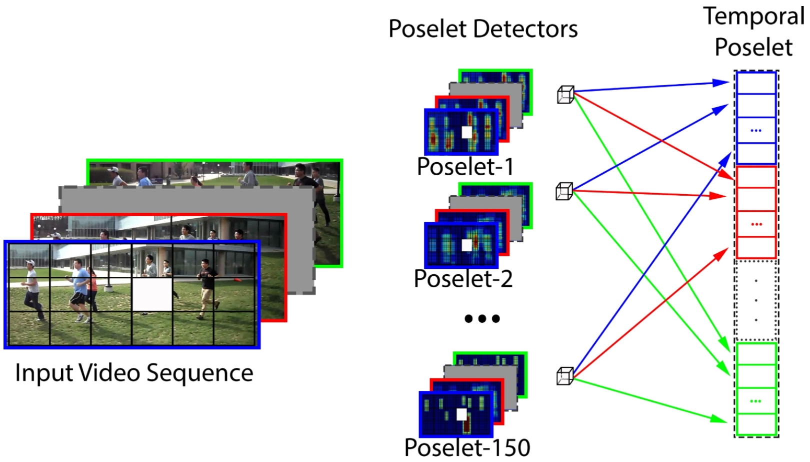


Figure 3. Temporal poselet for a particular video block (shown in white) where colours show poselet activation in different frames.

3. Temporal poselets for group activity recognition

Temporal poselets can be used for group activity recognition reaching a higher classification performance than standard spatio-temporal descriptors. In such case, we have to characterize an activity in a video stream given the temporal poselets and relate such information to the set of class activity labels. The standard procedure using temporal descriptors such as SURF3D and HOG3D follow the line of classical bag of words (BoW) approaches. Here we present an adaptation using the temporal poselets that actually achieves a finer representation power than standard methods.

First, we perform the usual split into training and testing sets using a dataset of videos showing different group activities. Then, our approach extracts all the temporal poselet vectors by dividing all the training video sequences in sub-sets of T frames thus obtaining F video clips in total. Since the overall number of temporal poselets can be arbitrarily high, we remove temporal poselets with a saliency value lower than a prefixed threshold (i.e. we keep a generic $\mathbf{TPOS}_k$ only if $s_k > s_{th}$). After this initial pruning stage we obtain N temporal poselets that are then used to create a codebook with k clusters (in general k is in the order of the hundred and N about 10^5/10^6). A K-means clustering method using the cosine distance as a similarity measure is adopted to compute this dictionary obtaining for each of the N temporal poselets an assignment to each of the k clusters.

Now, for the F labelled video clips we compute a histogram representing the frequency of the k temporal poselet words in each clip. This stage provides a set of F histogram $\mathbf{h}_f \in \Re^k$ with $f = 1 \ldots F$ which represents the frequent correlated Poselets in space and time for the activity classes considered. Representing the crowd motion using bag of the basis temporal poselets, provides more flexibility for representing very complex crowd motions. The histograms for each video clip and their related activity labels are then fed to a SVM classifier. At inference time we create a Bag-of-Word representation for each video clip by assigning each video block to the nearest cluster by using cosine distance. Finally we use trained SVM for classifying the activity of people in the input video clip.

4. Experimental Results

In this section we present the datasets used for evaluation, the baseline methods and our results for group detection and collective activity recognition.

4.1. Dataset description

We use several released versions of the Collective Activity Dataset (CAD), introduced first in [6] for evaluating collective activities. The dataset is suitable for our task because of the presence of multiple people activities in the natural unconstrained setting, while most of the classical activity datasets (i.e., CAVIAR, IXMAS, or UIUC), are not adequate for our purpose, since they either consider only the activity of a single person or few number of people [9]. We test our descriptor on second version of the Collective Activity Dataset (CAD2) [5] and the recently released third version (CAD3) [4]. CAD2 contains 75 video sequences captured with hand held cameras in realistic conditions including background clutter and mutual occlusions of people. We have activities classified with the following 6 categories: crossing, waiting, queueing, talking, dancing, and jogging. Instead CAD3 presents 33 video clips with 6 collective activities: gathering, talking, dismissal, walking together, chasing, queueing. The annotation of the datasets used in our approach are given by people bounding boxes and their trajectories in time. Noticeably, CAD3 has sequences with almost no camera motion with respect

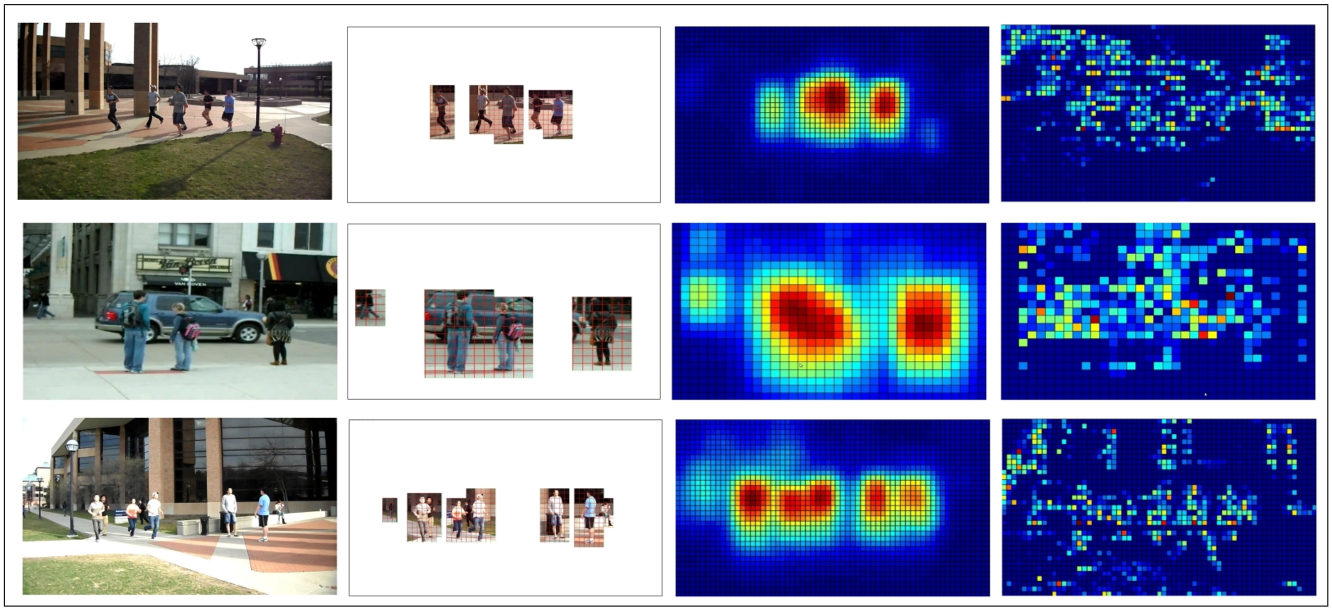


Figure 4. Temporal poselet for group detection: (a) represents a sample from 10-frame different video clips. (b) displays the ground truth positions for the people in the scene (c) shows the color coded activation for the Activation Map using temporal poselets where red represents the stronger activations (d) shows the color coded activation for the Activation Map using STIP [14].

to CAD2 where the camera is subject to relevant motion.

4.2. Implementation of the baseline methods

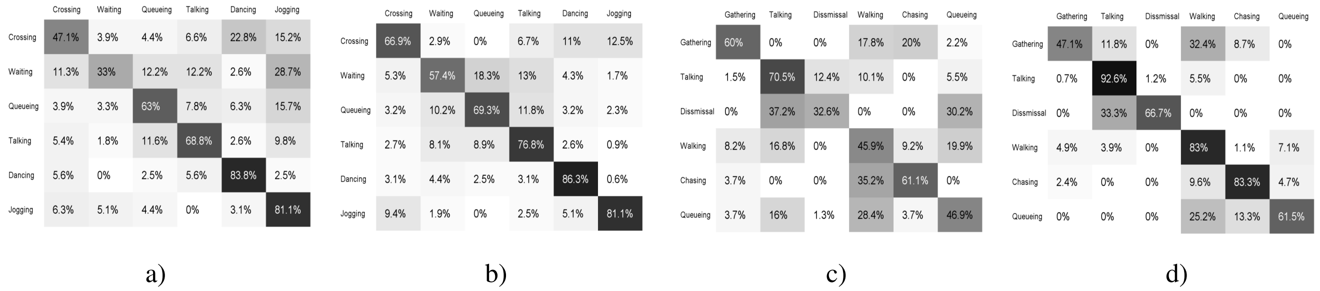
Our final goal is to evaluate the increment of performance given the proposed **TPOS** descriptor. For this reason, we implement two baseline methods with classical low-level temporal descriptors, and we report the increment of performance for collective activity recognition in two datasets (CAD2 and CAD3). When comparing with feature-based approaches we employ a similar pipeline and protocol as described in [23]: we extract local features, we perform vector quantization by using K-means, and we finally classify using χ^2 kernel SVMs. As presented in Sec. 3 our method changes the feature extraction stage: we replaced state-of-art descriptors with the new introduced temporal poselet features.

Activity Detection. In more details, we first run space-time interest points (STIP) [14] on each frame for each video sequence of the training videos [1]. Then, we divide each clip to fixed-size video blocks by applying a 3D grid (of size 20 × 20 × 10 pixels) as described in Sec. 3. Subsequently, we count the frequency of spatio-temporal interest points belonging to each block. This provides a map for each video clip similar to the Activation Map for temporal poselets which is shown in figure 4(d). We consider this approach as our baseline method for group detection.

Activity Recognition. Now, to create a baseline for feature based collective activity recognition, we then select a subset of the video blocks in which the number of spatio-temporal interest points inside them is higher than a pre-fixed threshold. Notice here that we empirically selected the saliency threshold s_{th}=150 in our method and s_{th}=120 for the baseline method since these thresholds gave the best performance for each algorithms.

At each selected video block we extract a HOG3D descriptor followed by the K-means clustering on around $700,000$ video blocks extracted from training data. We set the size of our codebook to 100, and then represent every video clip using bag of these visual words. This parameter K was optimized across the dataset carrying out several experiments with different parameter values finding no big differences in performance. In particular, we initially select K=4000 (same as [23]), and we reduced to K=1000 gradually down to 100, finding no significant improvement in terms of average accuracy. We finally kept K=100 as the best compromise between accuracy and computational cost. Finally we trained multi-class SVM with χ^2 kernel on a BoW representation of our video clips. In the inference phase, for each input video clip we again create a BoW representation by assigning each video block to the nearest cluster by using cosine distance. Finally, we use trained SVM for classifying the activity of people for every input video clip.

[1] For evaluation we use the code made available at: `http://www.di.ens.fr/~laptev/download.html` with its default parameters.



a)

	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%

b)

	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%

c)

	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%

d)

	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%

Figure 5. Confusion Matrix: (a) CAD2- Baseline method. (b) CAD2- **TPOS** method (c) CAD3- Baseline method (d) CAD3- **TPOS**.

4.3. Results and discussion

The main focus of the experiments is to quantitatively evaluate group activity detection and recognition using CAD2 and CAD3. Nevertheless, we also show qualitative results about the robustness of our approach to camera motion and dynamic background (see Fig. 4). Collective action detection is also evaluated with quantitative results in terms of ROC curve (see figure 6) against the saliency threshold parameter. For quantitative evaluation of group detection, we first generate the ground truth information for multiple people in videos using single people bounding boxes. This ground truth information is provided in terms of human/non-human label for each cell of the 3D grid. More precisely, for each video sequence we generate a 3D binary matrix with the same dimension of the Activation Map. This matrix consists of 1s where people are present in that video block and 0 where there is no person inside (see Fig. 4(b)). This ground truth matrix allows us to compare with the baseline method in terms of detecting video blocks as people. Since the density of people in each video sequence is shown as an Activation Map, varying the saliency threshold will result in different video blocks selected as people. We then compute the FP, TP, TN and FN for each threshold value by comparing the assigned label (human or non-human) and the ground truth label for each cell in the grid. This is given by counting the number of same/different labels of the cells in Activation Map in comparison with their corresponding cells in ground truth. The overall results shows 24.50% improvement in terms of the area under the ROC curve.

Our activity recognition results are shown in Fig. 5 in the form of confusion matrices for the 6-class datasets (CAD2 and CAD3). We also report the performance of our method compared to the baseline approach. In summary, we outperform the baseline in average by 10.10% in CAD2 and 19.50% in CAD3 (see Table 1). It is also interesting to notice the higher confusion values among the waiting, queueing and talking classes. This shows that our descriptor has the tendency to confound human poses related to these three classes because they are very similar in the pose of the people taking part on these activities (i.e. mainly standing). On the other hand, when the intensity of motion activity is higher in activities like in crossing and jogging, we have more accurate results. This can be related to the implicit dynamics extracted by the temporal poselet descriptor.

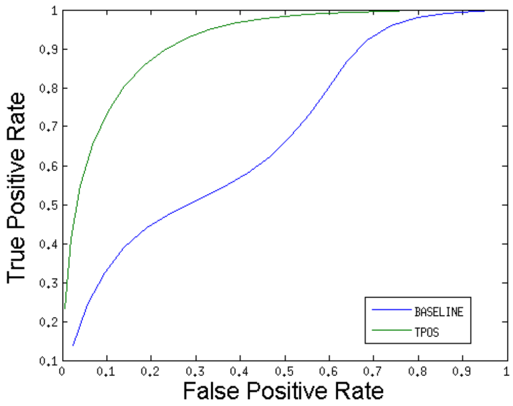


Figure 6. Group detection result: **TPOS** (green) vs. Baseline(blue)

Moreover, such behaviour does not emerge in the confusion matrix of the baseline approach because it captures the low-level statistics of the pixels motion and gradient. Thus, such descriptor is more sensitive to the appearance of the scene and of the background. In fact, temporal poselets are not likely to be activated in such parts of the image and they are directly related to the human content of the scene. It is also worth to note that the baseline method has worse results in CAD3 because of the lack of overall motion in the sequence since this dataset is a fixed camera scenario. Our method instead, since it is responsive to human parts can anyway provide reasonable results even if people are not moving too much.

Table 1 also shows the results of our method along with our baseline and the other recently published people-based approaches like [9, 5] for CAD2 and [12, 4] for CAD3. As mentioned before, the use of higher level features increases the performance of their systems. Actually, [9] employs

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

Table 1. Average Classification Accuracy.

additional information in training time including bounding boxes of all people, the associated actions and their identities. Instead, the RSTV model [5] beside using this information, it also adds additional trajectory information during training, including the location and the pose of every person as well. However, notice that we outperform the feature-level baseline approach significantly, and we slightly outperform the RSTV method when it is not optimized adopting an MRF on top. Obviously, we perform worse than Khamis et al. [9] because of the considerations made previously, which we included here for the sake of completeness of evaluation. In CAD3, our results are comparable with [12](as reported from [4]), and there is only 6.8% difference in average accuracy with respect to [4]. Both these methods are people-based and are based on complex models, the former using contextual information in a latent SVM framework, and the latter by combining belief propagation with a version of branch and bound algorithm equipped with integer programming.

5. Conclusions

We have evaluated our method by using the temporal poselet descriptor for group detection and activity recognition in the wild. The results show significant improvements in detection as well as recognition task in comparison with the baseline methods. Our representation based on temporal poselets can locally discover the presence of collective activities and recognise the type of action in a single video sequence. This aspect is very important for tasks in video-surveillance for crowded environments where there is a serious need to localize possibly anomalous activities. Moreover, this approach could also be used for video summarization, by extracting the video clips mostly representative of the peculiar collective activity in a long sequence. In particular, future work will be directed towards the space-time segmentation of different activities in a single video sequence.

References

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

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *CVPR 2009*.

[3] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *CVPR 2009*.

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

[5] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR 2011*.

[6] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops 2009*.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2nd Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance 2005*.

[8] S. Khamis, V. Morariu, and L. Davis. A flow model for joint action recognition and identity maintenance. In *CVPR 2012*.

[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*.

[10] A. Klaser and M. Marszalek. A spatio-temporal descriptor based on 3D-gradients. *BMVC 2008*.

[11] T. Lan, Y. Wang, G. Mori, and S. Robinovitch. Retrieving actions in group contexts. In *International Workshop on Sign Gesture Activity*, volume 3273, pages 3274–3278, 2010.

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

[13] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *TPAMI*, 34(8):1549–1562, 2012.

[14] I. Laptev. On space-time interest points. *IJCV*, 2005.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR 2008*, pages 1–8.

[16] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011*.

[17] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *NIPS 2010*, 24.

[18] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*.

[19] M. Raptis and L. Sigal. Poselet Key-framing: A Model for Human Activity Recognition. In *CVPR 2013*.

[20] S. Sadanand and J. J. Corso. Action Bank: A High-Level Representation of Activity in Video. In *CVPR 2012*.

[21] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia 2007*.

[22] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR 2011*.

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

[24] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *ECCV 2008*, pages 650–663.

[25] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV 2009*, pages 492–497, 2009.