Human Action Recognition and Localization using Spatio-temporal Descriptors and Tracking

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo
Lorenzo Seidenari, and Giuseppe Serra

Media Integration and Communication Center, University of Florence, Italy
{ballan,bertini,delbimbo,seidenari,serra}@dsi.unifi.it

Abstract. In this paper we propose a system for human action tracking and recognition using a robust particle filter-based visual tracker and a novel descriptor, to represent spatio-temporal interest points, based on an effective combination of a new 3D gradient descriptor with an optic flow descriptor. These points are used to represent video sequences using a bag of spatio-temporal visual words, following the successful results achieved in object and scene classification. The tracker assigns the points to each individual in a scene, allowing the classification of the action performed by each person. The system has been extensively tested on the standard KTH and Weizmann actions datasets, as well as on real world surveillance videos.

Key words: video annotation, action classification, bag-of-words, tracking.

1 Introduction and related works

Human activity recognition and action recognition in videos has attracted significant interest in recent years since it is useful for many applications such as video-surveillance, video annotation and retrieval and human-computer interaction [1]. Perhaps the video-surveillance domain is the one in which is focused the majority of the previous works in the field. This is probably due to the recently increased concern for safety and security issues. Just as an example, an action classification system that alerts a human operator of actions happening in a monitored area, that are possibly dangerous, can reduce human effort and mistakes. However, building a generic human activity recognition system is a challenging problem because of the variations in illumination, environment, size and postures appearance of the people.

Early research on action recognition relied on holistic representations such as spatio-temporal templates [2, 3] or space-time shapes [4]. But most of these approaches are computationally expensive due to the requirement of pre-processing the input data and, therefore, they perform better in a controlled environment. Most recently, part-based methods have been receiving increasing attention due to their simplicity and good performance on object, scene and action recognition problems. In particular, a method that has become very popular is the Bag-of-Words (BoW) approach [5, 6]. It has been originally proposed for document

classification in information retrieval and natural language processing, where each document is represented by its word frequency. In the visual domain, it has been firstly applied to images representing them using the frequency of "visual words" obtained by clustering local image descriptors (e.g. SIFT), typically using the k-means algorithm. More recently, this approach has been successfully applied also for human action classification in videos [7, 8], because it overcomes the limitations of holistic models. To this purpose several spatio-temporal detectors and descriptors have been proposed [7, 9–12] in order to effectively represent motion information. Laptev [9] initially proposed an extension to the Harris-Förstner corner detector for the spatio-temporal case; interesting parts are extracted from voxels surrounding local maxima of spatio-temporal corners, i.e. locations of videos which exhibit strong variations of intensity both in spatial and temporal directions. Dollár *et al.* [10] have followed, in principle, the same approach, but treating time differently from space and looking for locally periodic motion using a quadrature pair of Gabor filters.

These part-based approaches, similarly to those applied to object recognition, usually do not attempt to localize and track the actions. A few exceptions are the works of Niebles *et al.* [13] and Mikolajczyk and Uemura [14]. In the first one the authors exploit space-time patches posteriors, computed with an unsupervised probabilistic topic model, to localize the regions containing an action; however this approach fails in presence of two individuals performing the same action, which is recognized as a single one. In the second one, a vocabulary forest of local motion-appearance features is used for classification; in addition for localization they use a star shape model, able to learn the global structure from a set of examples annotated with bounding boxes. This technique provides a very accurate action segmentation, but the manual procedure of bounding box annotation is very time consuming.

In this paper, we present a system for human action classification, based on the BoW approach, that can be applied to videos containing multiple persons. In particular we propose a novel combined spatio-temporal descriptor: a 3D gradient part encodes mostly the visual appearance, while an optical flow part encodes the motion information. The detected spatio-temporal points are associated to each person present in the scene by a particle filter visual tracker, obtaining a precise spatio-temporal localization of each action. The rest of the paper is organized as follows: Sect. 2 shows the architecture of the proposed system and introduces the techniques for action representation, categorization and tracking. Sect. 3 presents the interest point detector and descriptors; experimental results, with an extensive comparison with state-of-the-art, are discussed in Sect. 4 and conclusions are drawn in Sect. 5.

2 Action Classification Architecture

The architectural design of the proposed solution, based on a bag-of-words model, is shown in Fig. 1. The basic idea of this model is to represent visual content as an unordered collection of visual "words". To this end, it is necessary

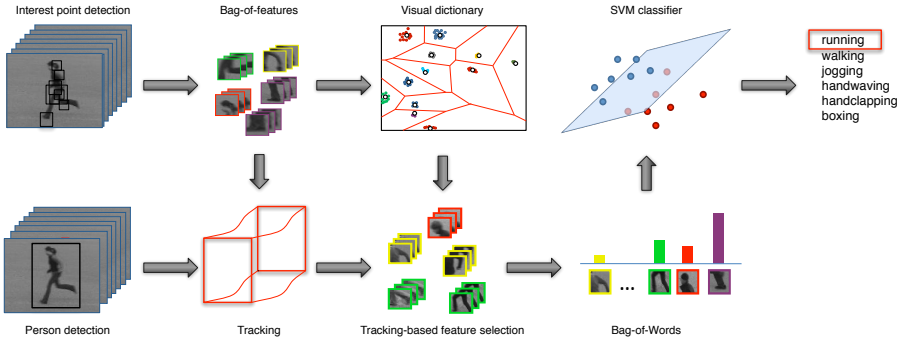

Fig. 1. The proposed solution architecture.

to define a visual dictionary from the local features extracted in the video sequences, performing a quantization of the original feature space. The descriptors used to represent the spatio-temporal interest points are presented in detail in Sect. 3.

The visual dictionary is generated by clustering of a set of interest points and each cluster is treated as a visual word. In particular, we use the k-means algorithm because of its simplicity and convergence speed.

Person tracking is used to assign the detected spatio-temporal interest points to each person present in a video, to localize both in space and time each recognized action. The tracker adopted in our system implements a particle filter based tracking algorithm, presented in [15], that tracks position, size and speed of the target, describing the target appearance with its color histogram. The tracker is initiated using the human detector of Dalal and Triggs [16], implemented in OpenCV. The state update equation, defined over the 8-dimensional state vector x_k, realizes a 1^{st}-order dynamic model:

$$x_k = Ax_{k-1} + v_{k-1} \tag{1}$$

where A is an 8×8 matrix and v_{k-1} is an additive, zero mean, isotropic Gaussian uncertainty term that represents the uncertainty in the state update. To improve the particle filter capability to effectively track the target, even if its appearance is not strongly characterized, the tracking method implements a particular technique to manage the uncertainty in the state update equation, by on-line adaptation of the error v_{k-1}. This allows the tracker to switch between two different behaviors: one that relies on the predicted motion of the target and one that behaves like a random-walk model.

By mapping the features associated to each tracked person in a video to the vocabulary, we can represent it by the frequency histogram of visual words. In order to reduce outliers, histograms of tracks that contain too few interest points, are discarded. Then, the remaining histograms are fed to a classifier to predict the action category. In particular, classification is performed using non-linear SVMs with the χ^2 kernel [6]. To perform multi-class classification we use

the *one-vs-one* approach. Fig. 2 shows an example of the tracker results and features association.

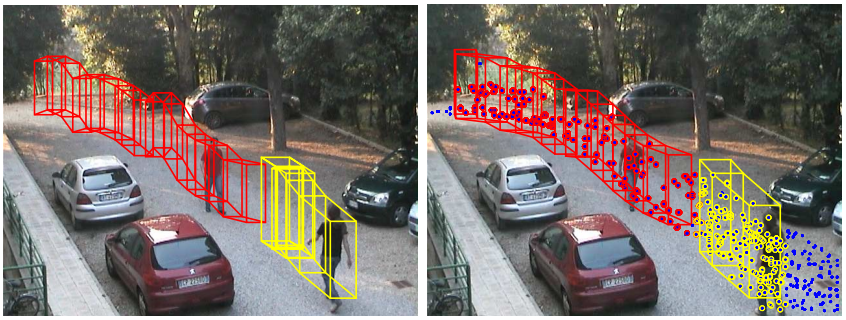

Fig. 2. Example of multiple person tracking, spatio-temporal interest point detection and their association to the tracks.

3 Fusing 3D gradients and Optical flow Descriptors

Typically the action recognition approaches that use a part-based representation perform detection and description of spatio-temporal interest points in two separate steps. Recently the detector proposed by Dollár *et al.* [10] has received a large attention from the scientific community and it has been adopted in several recent works (e.g. [13]). It runs on a single scale, and treats time and space in a different way, so that it generally produces a high number of detections. In our system we propose an extension to this detector, running it at multiple combinations of spatial and temporal scales.

The detector applies two separate linear filters to spatial and temporal dimensions, respectively. The response function is computed as follows:

$$R = (I(x,y,t) * g_\sigma(x,y) * h_{ev}(t))^2 + (I(x,y,t) * g_\sigma(x,y) * h_{od}(t))^2 \qquad (2)$$

where $I(x,y,t)$ is a sequence of gray-level images over time, $g_\sigma(x,y)$ is the spatial Gaussian filter with kernel σ, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied along the time dimension. They are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$, where $\omega = 4/\tau$ and they are explicitly designed to give high responses to periodical intensity changes. The interest points are detected at locations where the response is locally maximum. Representing motion patterns through spatio-temporal patches detected at multiple scales allows to describe events happening over different spatial and temporal extents. This kind of modelling introduces robustness w.r.t. actions happening at various distances from the observer and speed of execution. In particular the spatial scales used are $\sigma = \{2,4\}$ and the temporal scales are $\tau = \{2,4\}$.

A spatio-temporal volumetric patch is extracted in correspondence of each detected interest point. Its volume is proportional, both in space and time extensions, to the detected scale. To compute the representation of each volume we define a combined descriptor based on the fusion of three-dimensional gradients and optical flow. The motivation of this choice is that we expect these quantities to encode different information. For example the gradient descriptor, even taking into account the time dimension, is mostly an appearance descriptor while the optical flow is purely a motion representation. Each volume is divided in 18 subregions (three along each spatial direction and two along the temporal); each subregion is described using the two descriptors that are presented in the following. For both descriptors we use a polar coordinate representation.

The 3D gradient magnitude and orientations are:

$$M_{3D} = \sqrt{G_x^2 + G_y^2 + G_t^2},\ \phi = \tan^{-1}(G_t/\sqrt{G_x^2 + G_y^2}),\ \theta = \tan^{-1}(G_y/G_x) \quad (3)$$

where G_x, G_y and G_t are respectively computed using finite difference approximations: $L_{\sigma_d}(x+1,y,t) - L_{\sigma_d}(x-1,y,t)$, $L_{\sigma_d}(x,y+1,t) - L_{\sigma_d}(x,y-1,t)$ and $L_{\sigma}(x,y,t+1) - L_{\sigma_d}(x,y,t-1)$. L is obtained by filtering the signal I with a Gaussian kernel of bandwidth σ_d. We compute two separated orientation histograms quantizing ϕ and θ, weighting them by the magnitude M_{3D}. The ϕ (with range, $-\frac{\pi}{2}, \frac{\pi}{2}$) and θ ($-\pi$,π) are quantized in four and eight bins, respectively. The overall dimension of the descriptor is thus $3 \times 3 \times 2 \times (8+4) = 216$.

The optic flow is estimated using the Lucas&Kanade algorithm. Considering the optic flow computed for each couple of consecutive frames, the relative apparent velocity of each pixel is (V_x, V_y). These values are expressed in polar coordinates as in the following:

$$M_{2D} = \sqrt{V_x^2 + V_y^2},\ \theta = \tan^{-1}(V_y/V_x). \quad (4)$$

We compute the optical flow descriptor quantizing θ in eight bins. Every sample is then weighted with the magnitude M_{2D}. We also add an extra "no-motion" bin that, in our experiments, has shown to greatly improve the performance. The descriptor size is then $3 \times 3 \times 2 \times (8+1) = 162$.

The combination strategy adopted is a concatenation of the histograms of the bag-of-words that have been computed from the 3D gradient descriptor and from the histogram of optic. This means that the visual words are computed differently for each descriptor and the SVM classifiers are able to pick the best combinations of features, practically resulting in an implicit feature selection.

4 Experimental Results

First we evaluate the effectiveness of the proposed interest point detector and descriptor, comparing it to the state-of-the-art results. To this end we have employed two datasets, the KTH and Weizmann, commonly used as benchmarks for human action recognition. The KTH dataset contains 2391 video sequences with

25 actors performing six actions (walking, running, jogging, hand-clapping, hand-waving, boxing). The Weizmann dataset contains 93 video sequences showing nine different people, each performing ten actions (run, walk, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop-sideways, wave-two-hands, wave-one-hand and bend). All the videos of both datasets show only one individual. Our experimental setup is the same of the most recent works in action recognition domain and thus is suitable for a direct comparison [12, 8, 13, 17]. The SVM classifiers used for the KTH dataset were trained on videos of 16 actors and the performance was evaluated using the videos of the remaining nine actors. Measures have been taken according to a 5-fold cross-validation. In the Weizmann dataset the classifiers were trained on actions from eight actors and tested on the remaining one. Measures have been taken using the leave-one-out cross-validation. The dimension of the visual vocabulary is of 4000 visual words for KTH and 1500 for Weizmann, respectively.

Comparison to state-of-the art. The performance in terms of average class accuracy of our combined descriptor is shown in Table 3. The good results are due to the fact that the performances of 3D gradient and optical flow are quite complementary (see [18] for more details). Table 3 reports also a comparison of our approach with state-of-the-art results, as reported by other researchers. Results obtained on both datasets, KTH and Weizmann, using our method outperform previous works based on a BoW model [7, 10–13, 17], and also the results reported by Liu *et al.* [19] obtained combining and weighting multiple features. Note that the results that are closer to ours (i.e. [8, 11]) require a heavy parameter tuning that is not required in our approach.

Method	**KTH**	**Weizmann**
Our method	**92.1**	**92.41**
Laptev *et al.* [8]	91.8	-
Dollár *et al.* [10]	81.2	-
Wong and Cipolla [17]	86.62	-
Scovanner *et al.* [7]	-	82.6

Method	**KTH**	**Weizmann**
Schüldt *et al.* [20]	71.7	-
Niebles *et al.* [13]	83.33	90
Liu *et al.* [19]	-	90.4
Kläser *et al.* [12]	91.4	84.3
Willems *et al.* [11]	84.26	-

Table 3. Comparison of our method with different methods, using KTH and Weizmann datasets.

Real world surveillance. In addition, we test our approach on real world surveillance videos. In particular we use five complex video sequences containing multiple actions performed concurrently (two examples are shown in Fig. 3). These sequences are taken at different times of the day with different durations (from a minimum of 4 sec. to a maximum of 15 sec). Our method has been applied to recognize and localize two basic actions: walking and running. As training set we recorded 63 videos with eight different actors. Each video contains a person performing the same action multiple times.

Tab. 4 shows the performance of our approach on surveillance videos. For each sequence we report the detected tracks identified from our person detector and tracker. The tracks that contain less than 30 interest points are discarded and

Fig. 3. Example of two sequences from our surveillance dataset. In the first sequence (seq. 3) our actors perform a pickpocketing event. In the second sequence (seq. 5) a snatch is performed.

the filtered tracks are then used to perform action classification. These tracks are manually annotated in walking, running and unknown (in Table $\mathbf{W}_{\text{GT}}$, $\mathbf{R}_{\text{GT}}$,$\mathbf{U}_{\text{GT}}$ respectively). Details of their classification and accuracy are shown. We note that 21/27 tracks are recognized correctly. In the first and last sequences we note that our method makes only a mistake classifying a walking in running action. Instead in the second, third and fourth sequence there is an incorrect filtering of the detected tracks; in these cases our method mistakenly considers a track which does not contain a known action.

Video Seq.	**Detected**	**Filtered**	$\mathbf{W}_{\text{GT}}$	$\mathbf{R}_{\text{GT}}$	$\mathbf{U}_{\text{GT}}$	$\mathbf{W}_{\text{TP}}$	$\mathbf{W}_{\text{FP}}$	$\mathbf{R}_{\text{TP}}$	$\mathbf{R}_{\text{FP}}$	**Accuracy**
1	8	5	3	2	0	2	0	2	1	4/5
2	7	6	3	2	1	3	1	2	0	5/6
3	11	5	2	2	1	2	1	2	0	4/5
4	8	6	2	3	1	2	1	2	1	4/6
5	8	5	3	2	0	2	0	2	1	4/5
										21/27

Table 4. System performance on complex video sequences: for each action ground-truth ($\mathbf{W}_{\text{GT}}, \mathbf{R}_{\text{GT}}, \mathbf{U}_{\text{GT}}$), true positives ($\mathbf{W}_{\text{TP}}, \mathbf{R}_{\text{TP}}$) and false positives ($\mathbf{W}_{\text{FP}}, \mathbf{R}_{\text{FP}}$) are reported.

5 Conclusions

In this paper we have presented a novel method for human action categorization based on a combination of a new 3D gradient with an optical flow descriptor, for spatio-temporal interest points. The approach was validated on two popular datasets (KTH and Weizmann), showing results that outperform state-of-the-art methods, without requiring parameter tuning. In addition, we have tested our approach on a realistic dataset, designed for video surveillance purposes. Our future work will deal with integration of visual cues, action classification and trajectories to recognize composite and group actions.

Acknowledgments. This work is partially supported by the EU IST VidiVideo Project (Contract FP6-045547) and IM3I Project (Contract FP7-222267).

References

1. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology **18**(11) (2008) 1473–1488
2. Polana, R., Nelson, R.: Detecting activities. In: Proc. of CVPR. (1993)
3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3) (2001) 257–267
4. Gorelick, L., Blank, M., Schechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12) (2007)
5. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proc. of ICCV. (2003)
6. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision **73**(2) (2007) 213–238
7. Scovanner, P., Ali, S., Shah, M.: A 3-Dimensional SIFT descriptor and its application to action recognition. In: Proc. of ACM Multimedia. (2007)
8. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. of CVPR. (2008)
9. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2-3) (2005) 107–123
10. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proc. of VSPETS. (2005)
11. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proc. of ECCV. (2008)
12. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-Gradients. Proc. of BMVC (2008)
13. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3) (2008) 299–318
14. Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: Proc. of CVPR. (2008)
15. Bagdanov, A.D., Dini, F., Del Bimbo, A., Nunziati, W.: Improving the robustness of particle filter-based visual trackers using online parameter adaptation. In: Proc. of AVSS. (2007)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of CVPR. (June 2005) 886–893
17. Wong, S.F., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: Proc. of ICCV. (2007)
18. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Effective codebooks for human action categorization. In: Proc. of ICCV International Workshop on Video-oriented Object and Event Classification (VOEC). (2009)
19. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: Proc. of CVPR. (2008)
20. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proc. of ICPR. (2004)