# RECOGNIZING HUMAN ACTIONS BY FUSING SPATIO-TEMPORAL APPEARANCE AND MOTION DESCRIPTORS

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

Media Integration and Communication Center, University of Florence, Italy
http://www.micc.unifi.it/vim

## ABSTRACT

In this paper we propose a new method for human action categorization by using an effective combination of a new 3D gradient descriptor with an optic flow descriptor, to represent spatio-temporal interest points. 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. We extensively test our approach on the standard KTH and Weizmann actions datasets, showing its validity and good performance. Experimental results outperform state-of-the-art methods, without requiring fine parameter tuning.

***Index Terms***— Action recognition, spatio-temporal descriptors, bag-of-words

## 1. INTRODUCTION AND PREVIOUS WORK

Automatic human action recognition in videos has attracted significant interest in recent years since it is useful for many applications such as video annotation and retrieval, human-computer interaction and video-surveillance. For example, considering the video-surveillance domain, an action classification system that alerts an operator of actions 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.

Existing action recognition methods can be classified as using *holistic* [1, 2] or *part-based* information. Most of the holistic-based approaches are computationally expensive due to the requirement of pre-processing the input data and they perform better in a controlled environment. Part-based representations, that exploit interest point detectors combined with robust description methods, have been used very successfully for object and scene classification tasks in images. In particular, an approach that has become very popular is the Bag-of-Words (BoW) model [3, 4, 5]; 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). More recently, this approach has been successfully applied also to the human action classification problem in videos [6, 7, 8], because it overcomes some limitations of holistic models such as the necessity of performing background subtraction and tracking. To this purpose, several spatio-temporal detectors and descriptors have been proposed in the literature [6, 9, 10, 11, 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. This approach results in a denser sampling of the spatio-temporal volume but does not provide a scale-selection criterion. Finally, Willems *et al.* [11] extended the SURF detector and descriptor to the spatio-temporal case obtaining computationally efficient scale-invariant features; Kläser *et al.* [12] proposed a new space-time descriptor based on 3D gradients.

In this paper, we present a new method for human action classification based on the BoW model. In particular we define a novel spatio-temporal descriptor that combines 3D gradient and optic flow descriptors; the gradient part encodes mostly the visual appearance, while the optical flow descriptor encodes the motion information. The experimental results, obtained on KTH and Weizmann datasets, show that our method outperforms state-of-the-art approaches. The rest of the paper is organized as follows: Sect. 2 presents the interest point detector and descriptors; Sect. 3 introduces the techniques for action representation and categorization. Experimental results, with an extensive comparison with state-of-the-art, are discussed in Sect. 4 and conclusions are drawn in Sect. 5.

## 2. DETECTOR AND DESCRIPTORS

Following the approach commonly used for local interest points in images, the detection and description of spatio-temporal interest points are separated in two different steps.

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

Two interest point detectors have recently received most of the attention from the scientific community [9, 10]. The main drawback of the one proposed by Laptev [9] is the excessive sparseness of the spatio-temporal patches extracted from the video. This issue has been addressed by the same author [8] removing the scale-selection step of his algorithm. The single scale detector proposed by Dollár *et al.* [10] treats time and space in a different way resulting also in a denser sampling of the space-time. We believe this approach is more suitable for space-time interest point detection and we propose an extension to this operator running the filter at multiple combinations of spatial and temporal scales.

### 2.1. Detector

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

$$\begin{aligned} R = &(I(x,y,t) * g_\sigma(x,y) * h_{ev}(t))^2 \\ &+ (I(x,y,t) * g_\sigma(x,y) * h_{od}(t))^2 \end{aligned} \tag{1}$$

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 $\sigma$, $h_{ev}$ 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 $\tau = \{2,4\}$.

### 2.2. Descriptors

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 positional dependent statistics of each volume we divide it in 18 subregions (respectively three along the spatial directions and two along the temporal). We use two measures to create the final descriptors: three-dimensional image 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. The two descriptors are presented in the following. For both descriptors we use a polar coordinate representation.

The 3D gradient magnitude and orientations are:

$$M_{3D} = \sqrt{G_x^2 + G_y^2 + G_t^2}, \tag{2}$$

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

$$\theta = \tan^{-1}(G_y / G_x). \tag{4}$$

where $G_x$, $G_y$ and $G_z$ 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 $L_\sigma(x,y,t+1) - L_{\sigma_d}(x,y,t-1)$. Where $L$ is obtained by filtering the signal $I$ with a Gaussian kernel of bandwidth $\sigma_d$, we are therefore using approximated Gaussian derivatives. We compute two separated orientation histograms quantizing $\phi$ and $\theta$, weighting them by the magnitude $M_{3D}$. The $\phi$ (with range, $-\frac{\pi}{2}, \frac{\pi}{2}$) and $\theta$ ($-\pi$,$\pi$) are quantized in four and eight bins, respectively. To increase the robustness of the feature description the spatio-temporal gradient is computed using two adjacent differentiation scales $\sigma_d$. The overall dimension of the descriptor is thus $3 \times 3 \times 2 \times (8+4) \times 2 = 432$. This construction of the three-dimensional histogram is inspired by the approach proposed by Scovanner *et al.* [6], where they construct a weighted three-dimensional histogram normalized by the solid angle value (instead of quantizing separately the two orientations) to avoid distortions due to the polar coordinate representation. Moreover we do not re-orient the 3D neighbourhood, since rotational invariance, which is invaluable in object detection and recognition, is not desired in an action categorization context. We have found that our method is computationally less expensive, equally effective in describing motion information given by appearance variation, and showing a better performance (see comparison results in Tab. 2).

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:

$$M_{2D} = \sqrt{V_x^2 + V_y^2}, \tag{5}$$

$$\theta = \tan^{-1}(V_y / V_x). \tag{6}$$

We compute position dependent histograms as in the gradient based descriptor but, being the optic flow two dimensional, only a single orientation histogram is stored for each of the 18 sub-regions within the voxel. Every sample is weighted with the magnitude $M_{2D}$, as is done for the gradient-based descriptor. We also add an extra "no-motion" bin that, in our initial experiments, has shown to greatly improve the performance. The descriptor size is then $3 \times 3 \times 2 \times (8+1) = 162$.

## 3. ACTION REPRESENTATION AND CLASSIFICATION

The basic idea of the bag-of-words (BoW) model is to represent a text or, as in our case, a visual content as an unordered

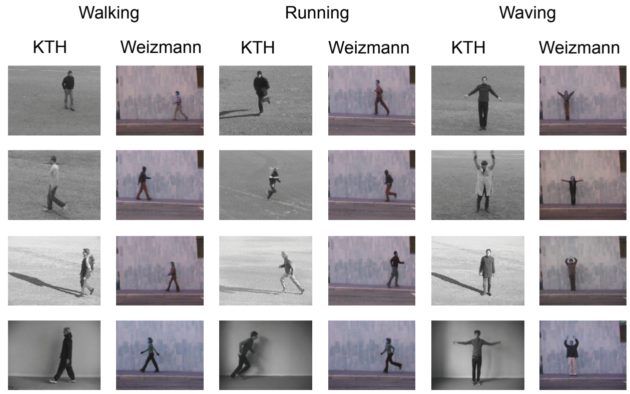


**Fig. 1**. Sample frames from the KTH and Weizmann datasets (Walking, Running and Waving actions) showing the higher variability present in the KTH set.

collection of (visual) words. To this end, it is necessary to define a visual vocabulary from the local features extracted in the video sequences, performing a quantization of the original feature space.

We have analysed four different types of action representation using visual words, considering the two single descriptors presented in Sect. 2.2 and two possible combinations of these descriptors. The proposed combinations are: *i)* a weighted concatenation of the two descriptors and *ii)* 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 flow. The first technique tries to obtain a joint representation of appearance and motion for each spatio-temporal volume. The latter acts at a higher level by first forming two vocabularies, clustering descriptors of each kind separately, and then by concatenating the final visual words histograms; thus the SVM classifiers will be able to pick the best combinations of features.

The visual vocabulary 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. By mapping the features extracted from a video to the vocabulary, we can represent it by the frequency histogram of visual words. Then, this histogram is fed to a classifier to predict the action category. In particular, classification is performed using non-linear SVMs with the $\chi^2$ kernel [5]. To perform multi-class classification we use the *one-vs-one* approach.

## 4. EXPERIMENTAL RESULTS

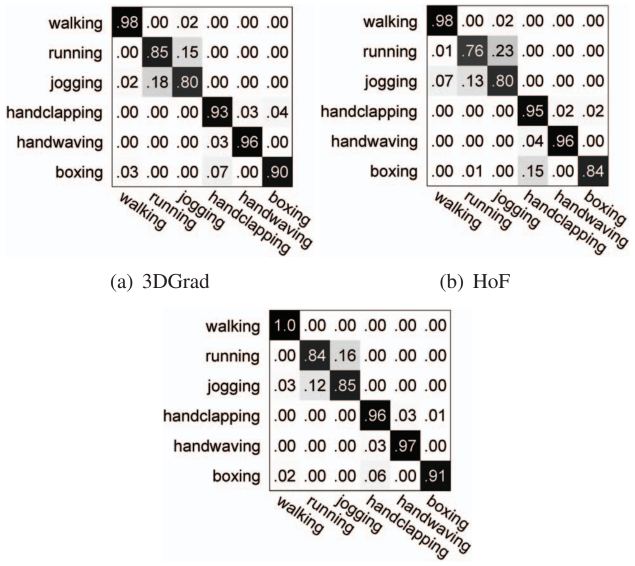
In this section, we report experiments conducted to validate our proposed method. First we have evaluated our descriptors and then we have compared our approach to the state-of-the-art results. The two datasets, the KTH and Weizmann, commonly used for human action recognition are used as benchmarks. The KTH dataset contains 2391 video sequences with 25 actors showing 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 such as 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. Note that due to the large number of actors, clothing changes, shadows and scenario, the KTH dataset can be considered more challenging with respect to the Weizmann. Fig. 1 shows these differences presenting sample frames selected from videos that contain the same action in the two datasets. 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, 7, 13]. 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 9 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 quantization approach used to define the visual vocabulary is k-means clustering, with 4000 visual words for KTH and 1500 for Weizmann, respectively.

### 4.1. Evaluation of our descriptor

Table 1 shows the performance obtained using 3D gradient description and optical flow description (HoF) alone and their two combinations discussed in Sect. 3. In particular, in the first two rows we report results obtained using only one of the two descriptors: 3D gradient in the first row and histogram of optic flow in the second. In the third row are reported the results for the descriptor that is obtained through a weighted concatenation of the two descriptors, while in row four the descriptor is composed by the concatenation of the histograms of the bag-of-words that have been computed from the 3D gradient descriptor and from the histogram of optic flow. The best result is obtained by the concatenation of the histograms. This is due to the fact that the performance of 3D gradient and HoF are quite complementary and because with this concatenation the SVM classifiers improve their implicit feature selection of the descriptor that represents better the action. As an example, the action recognition performance (Fig. 2) for the boxing class on the KTH dataset is lower when using the HoF description instead of the 3D gradient, while for handclapping is the opposite case. It can be observed (Fig. 2 c) that the concatenation of the histograms of the BoWs com-

| **Descriptor** | **KTH** | **Weizmann** |
|---|---|---|
| 3DGrad | 90.38 ± 0.8 | 92.30±1.6 |
| HoF | 88.04 ± 0.7 | 89.74±1.8 |
| 3DGrad_HoF combination | 91.09 ± 0.4 | 92.38±1.9 |
| 3DGrad+HoF combination | **92.10 ± 0.4** | **92.41 ±1.9** |

**Table 1**. Comparison of our descriptors, alone and combined, on the KTH and Weizmann datasets.



| | walking | running | jogging | handclapping | handwaving | boxing |
|---|---|---|---|---|---|---|
| walking | .98 | .00 | .02 | .00 | .00 | .00 |
| running | .00 | .85 | .15 | .00 | .00 | .00 |
| jogging | .02 | .18 | .80 | .00 | .00 | .00 |
| handclapping | .00 | .00 | .00 | .93 | .03 | .04 |
| handwaving | .00 | .00 | .00 | .03 | .96 | .00 |
| boxing | .03 | .00 | .00 | .07 | .00 | .90 |

(a) 3DGrad

| | walking | running | jogging | handclapping | handwaving | boxing |
|---|---|---|---|---|---|---|
| walking | .98 | .00 | .02 | .00 | .00 | .00 |
| running | .01 | .76 | .23 | .00 | .00 | .00 |
| jogging | .07 | .13 | .80 | .00 | .00 | .00 |
| handclapping | .00 | .00 | .00 | .95 | .02 | .02 |
| handwaving | .00 | .00 | .00 | .04 | .96 | .00 |
| boxing | .00 | .01 | .00 | .15 | .00 | .84 |

(b) HoF

| | walking | running | jogging | handclapping | handwaving | boxing |
|---|---|---|---|---|---|---|
| walking | 1.0 | .00 | .00 | .00 | .00 | .00 |
| running | .00 | .84 | .16 | .00 | .00 | .00 |
| jogging | .03 | .12 | .85 | .00 | .00 | .00 |
| handclapping | .00 | .00 | .00 | .96 | .03 | .01 |
| handwaving | .00 | .00 | .00 | .03 | .97 | .00 |
| boxing | .02 | .00 | .00 | .06 | .00 | .91 |

(c) 3DGrad+HoF combination

**Fig. 2**. Confusion Matrices on the test set KTH actions.

puted from both descriptors improves the performance for all the classes except one. The only case where there is no improvement, is when the performance of the two descriptors is too different. The smaller improvement obtained on the Weizmann dataset is probably caused by the smaller training set that is available and the increased size of the representation.

### 4.2. Comparison to state-of-the art

In Table 2 we report a comparison of the average class accuracy of our approach with state-of-the-art results, reported by other researchers. Results obtained on both datasets, KTH and Weizmann, using our method outperform previous works based on a BoW model [10, 12, 7, 11, 13, 12, 7, 6], and also the results reported by Liu *et al.* [14] obtained combining and weighting multiple features. Note that the results that are closer to ours, the works present by Laptev *et al.* [8] and Kläser *et al.* [11], require heavy parameter tuning. In fact in [8] the results are obtained by a fine tuning of different descriptors and grids which add structural information while in [11] is used a single 3D gradient descriptor but with a heavy optimization of its parameters (eight in total), that are dependent on the dataset used. Finally, we cannot compare to results by Gorelick *et al.* [2] or Fathi and Mori [15], because they use an holistic representation and more data given by segmentation masks.

## 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. Our future work will deal with development of new quantization method and evaluation on real world videos.

| Method | KTH | Weizmann |
|---|---|---|
| Our method | **92.10** | **92.41** |
| Laptev *et al.* [8] | 91.8 | - |
| Dollár *et al.* [10] | 81.2 | - |
| Wong and Cipolla [13] | 86.62 | - |
| Scovanner *et al.* [6] | - | 82.6 |
| Niebles *et al.* [7] | 83.33 | 90 |
| Liu *et al.* [14] | - | 90.4 |
| Kläser *et al.* [12] | 91.4 | 84.3 |
| Willems *et al.* [11] | 84.26 | - |

**Table 2**. Comparison of our method with different methods, using KTH and Weizmann datasets.

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

## 6. REFERENCES

[1] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[2] L. Gorelick, M. Blank, E. Schechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, 2007.

[3] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. of CVPR*, 2003.

[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, 2003.

[5] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[6] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. of ACM Multimedia*, 2007.

[7] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[8] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. of CVPR*, 2008.

[9] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[10] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of VSPETS*, 2005.

[11] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of ECCV*, 2008.

[12] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," *Proc. of BMVC*, 2008.

[13] Shu-Fai Wong and Roberto Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. of ICCV*, 2007.

[14] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. of CVPR*, 2008.

[15] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. of CVPR*, 2008.