# Representing activities with layers of velocity statistics for multiple human action recognition in surveillance applications

Fabio Martínez, Antoine Manzanera, Eduardo Romero

# Representing activities with layers of velocity statistics for multiple human action recognition in surveillance applications

Fabio Martínez [a], Antoine Manzanera [b] and Eduardo Romero[a]

[a] Computer Imaging and Medical Applications Laboratory – CIM@LAB, Universidad Nacional de Colombia, Bogotá, Colombia

[b] Unité d'Informatique et d'Ingénierie des Systèmes, ENSTA-ParisTech

## ABSTRACT

A novel action recognition strategy in a video-surveillance context is herein presented. The method starts by computing a multiscale dense optical flow, from which spatial apparent movement regions are clustered as Regions of Interest (RoIs). Each ROI is summarized at each time by an orientation histogram. Then, a multilayer structure dynamically stores the orientation histograms associated to any of the found RoI in the scene and a set of cumulated temporal statistics is used to label that RoI using a previously trained support vector machine model. The method is evaluated using classic human action and public surveillance datasets, with two different tasks: (1) classification of short sequences containing individual actions, and (2) Frame-level recognition of human action in long sequences containing simultaneous actions. The accuracy measurements are: 96.7% (sequence rate) for the classification task, and 95.3% (frame rate) for recognition in surveillance scenes.

**Keywords:** Action recognition, optical flow, Motion descriptors, video-surveillance

## 1. INTRODUCTION

Human action recognition is the base of many surveillance applications.[1,2] The challenge is to automatically detect an action occurring in a recorded sequence and to classify selected actions. The great difficulty arises with the extreme variability, in terms of geometry of the scene, people appearance and poorly controlled conditions. Previous action recognition methods are based on motion global descriptors that use optical flow strategies. For instance, Ikizler et al[3] used orientation histograms of a pre-computed optical flow combined with contour orientations. This method can distinguish simple periodic actions but the contour-optical flow integration is too limited to address more complex activities. Chaudhry et al[4] described human activities using histograms of oriented optical Flow (HOOF) with vertical symmetry (i.e. ignoring the difference between motions to-the-left and to-the-right). Such symmetry results in certain invariance but also in limitations to distinguish some actions, for instance antagonist motions of the limbs. Other methods use sparse optical flow, reducing the representation to certain salient features that may not be representative enough to describe a particular action.[5,6] On the other hand, strategies based on local patch relationships highlight and summarize the motion, for example, $3d$ spatio temporal Haar features, proposed in pedestrian applications.[7] Such descriptors are less sensitive to the quality of the action segmentation but remain strongly dependent on the subject appearance and recording conditions.

The aim of the present work is to recognize human actions using motion descriptors from temporal series of orientation histograms, collected within a multi-layer structure, each layer being a potential human action (temporal RoI). This approach starts by computing a dense optical flow that is then spatially clustered into RoIs which are described as orientation histograms. The histograms are gathered within a multi-layer structure that allows to handle and store temporal information. A motion descriptor can then be extracted at any moment by computing temporal statistics from a particular layer. Finally, the descriptors are labelled as actions using SVM classification. Evaluation was performed with Weizmann[8] and ViSOR video-surveillance[9] databases. This paper is organized as follows: Section 2 details the proposed approach, Section 3 demonstrates the effectiveness of the method and the last section presents the conclusions and possible future works.

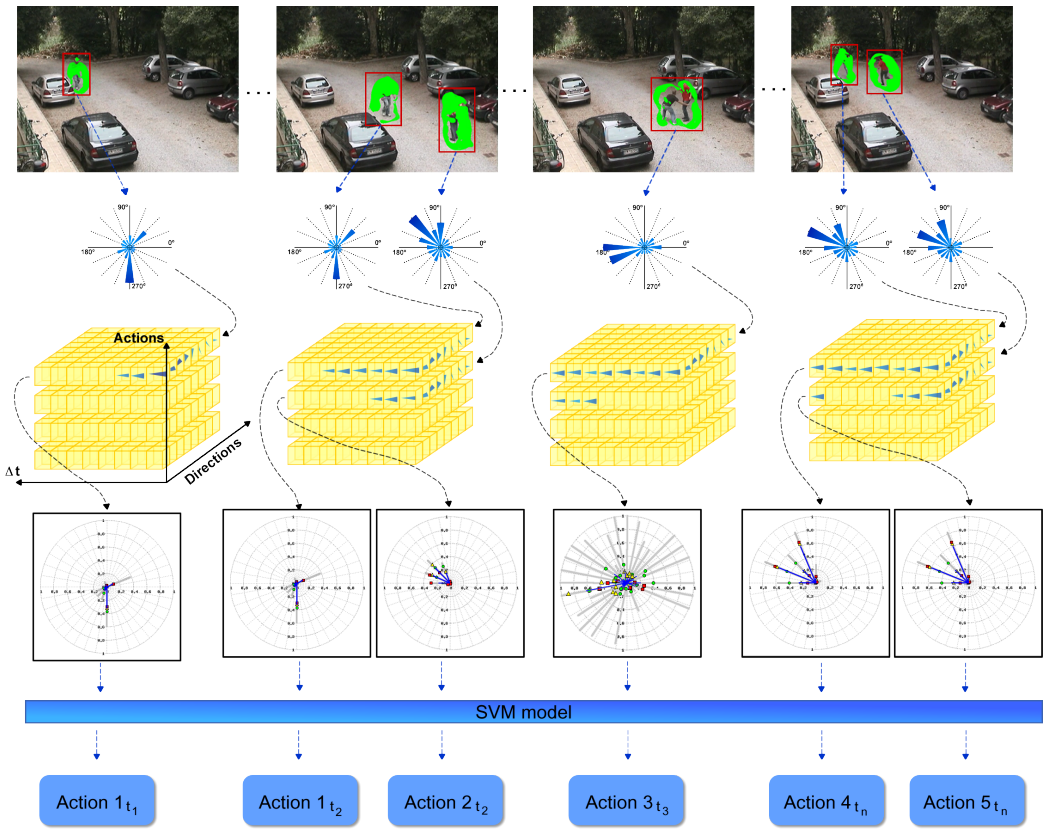


Figure 1. The proposed method starts by computing a dense optical flow, which is clustered as individual motion RoIs. An orientation histogram for each RoI is then calculated and stored in a multi layered data structure. At any time, a vector of characteristics is extracted from every layer and classified using a SVM model to label the action.

## 2. PROPOSED METHOD

In Figure 1 the whole method is summarized.

### 2.1 Optical flow characterization

The computed multiscale dense optical flow estimation[10] consists in projecting every pixel to a feature space composed of spatial derivatives of different orders, at several scales (the local jet). Then, for each frame and every pixel, the apparent velocity vector is estimated by searching the pixel associated to the nearest neighbor in the space of local jet vector at the precedent time. The dense optical flow is firstly used to coarsely segment potential human actions by morphologically closing those pixels whose velocity norm is above a certain threshold and spatio-temporally connecting the resulting regions, according to a distance criterion ( see Figure 1, first row). Afterward, for each RoI, a frame-level descriptor is built, based on the distribution of the instantaneous motion orientations. For a non-null flow vector $\mathbf{V}$, let $\phi(\mathbf{V})$ its orientation, quantized to $N$ values. The motion orientation histogram of each RoI frame is computed as the relative occurrence of flow vectors with a given orientation, weighted by their vector norm[11] ( see Figure 1, second row) :

$$H_t(\omega) = \frac{\sum_{\{x;\phi(\mathbf{V}_t(\mathbf{x}))=\omega\}} \|\mathbf{V}_t(\mathbf{x})\|}{\sum_{\{x;\|\mathbf{V}_t(\mathbf{x})\|>0\}} \|\mathbf{V}_t(\mathbf{x})\|}$$

where $\omega \in \{\omega_0 \dots \omega_{N-1}\}$. $N$ is the number of orientations.

Corresponding author: E. Romero. Email: edromero@unal.edu.co

## 2.2 Multi-layer data structure

The RoI histograms are stored in a FIFO multi-layer data structure, the $x-$ axis being the computed histogram, the $y-$ axis the temporal dimension and the $z-$ axis the actions (layers) present in the video (Figure 1, third row). If a data layer collects more than three consecutive histograms, it is considered as potential motion and then a motion descriptor can be computed. When a layer does not show activity for more than three consecutive frames, it is voided and the corresponding motion is eliminated.

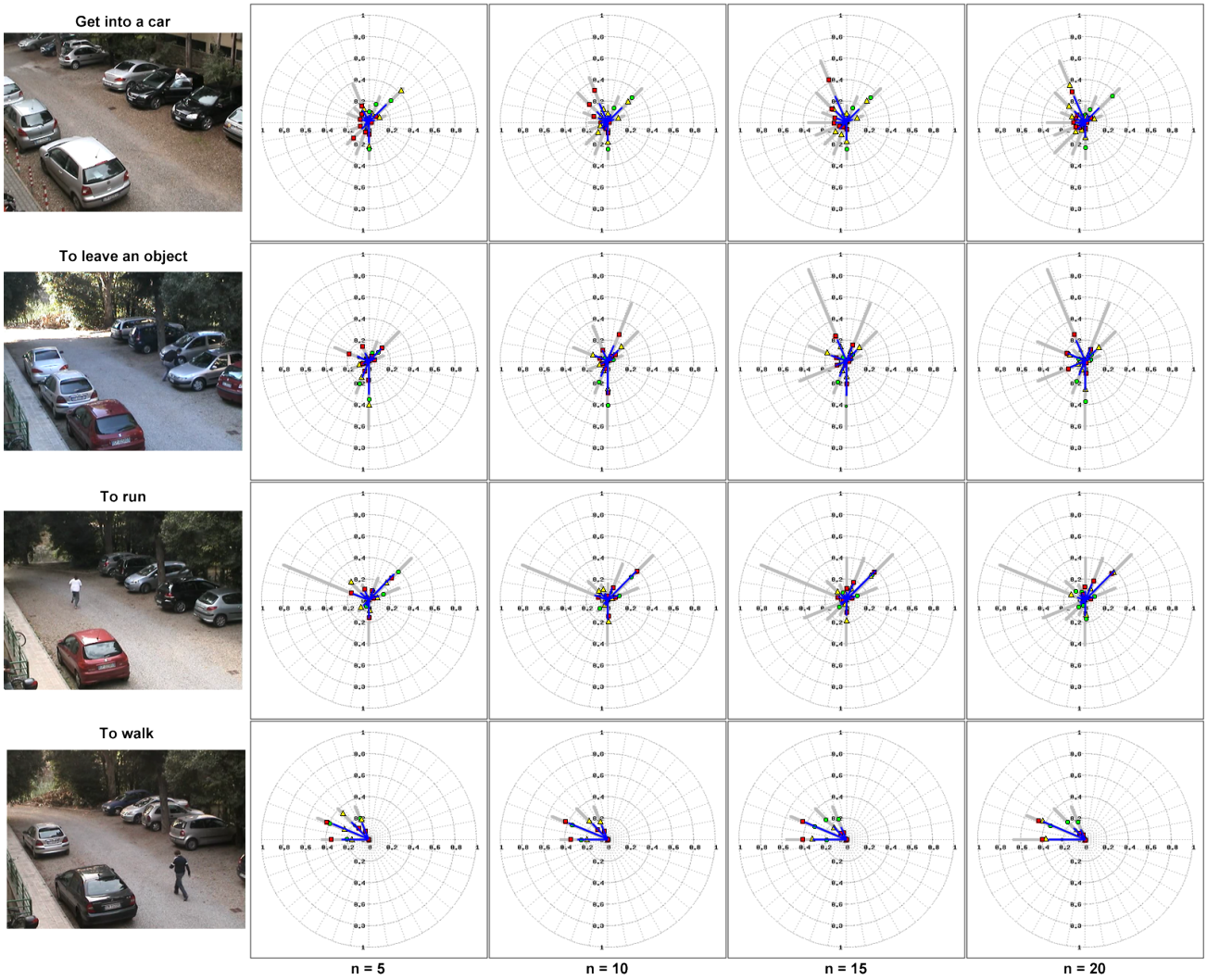


Figure 2. Example of motion descriptors. The blue and gray lines represent the maximum and mean values, respectively. The red square, yellow triangle and green circle represent the mean values for the beginning, middle and end portion of the $n$ histograms respectively

## 2.3 Motion descriptor

For each activity layer of the data structure, a motion descriptor is computed. If the layer contains $n$ histograms $H_t(\omega)$, a set of temporal cumulated statistics are calculated for every orientation $\omega$, as follows:

1. **Maximum:** $M(\omega) = \max_{0 \leq t < n} \{H_t(\omega)\}$

2. **Mean:** $\mu(\omega) = \sum_{0 \leq t < n} \frac{H_t(\omega)}{n}$

**3. Standard deviation:** $\sigma(\omega) = \sqrt{\sum_{0 \leq t < n} \frac{H_t(\omega)^2}{n} - \mu(\omega)^2}$

Afterwards, the histograms stored in the multi-layer structure are also split into 3 intervals of equal durations, and the corresponding means are calculated. Examples of human action descriptors are shown in Figure 2. For the initial computation of the motion descriptor, different activities may show similar patterns, because some actions are composed by two or more simple motions. However after 10 frames, the activities are usually well labeled.

### 2.4 SVM Classification and Recognition

Finally, the recognition of each potential motion stored in a layer is performed using a Support Vector Machine (SVM) classifier, a *one-against-one SVM multiclass classification*[12] ( see an example in Figure 3). As will be shown later, taking into account all the vote values instead of the winner label only is useful to perform time filtering of the recognition. A Radial Basis Function (RBF) SVM model was trained with a set of motion descriptors,[13] extracted from previously labeled human activity sequences. A sensitivity parameter analysis $(\gamma, C)$, was performed under a grid-search using a cross-validation. Additionally, a simple rule was introduced to detect complex activities. If the system detects two simple human actions and they are grouped as a single region, a new activity is defined and tagged as complex.

## 3. EVALUATION AND RESULTS

Experimentation was carried out with different public dat asets. In the first experiment, we tested our approach in an action classification task, using a leave-one-out cross validation scheme. Firstly, we used the Weizmann dataset,[8] which is composed of 9 subjects and 10 actions recorded in 93 sequences. The corresponding confusion matrix for the Weizmann dataset is shown in Table 1. The proposed approach achieves an average accuracy of 95%.

| Category | bend | jack | jump | pjump | run | side | skyp | walk | wave1 | wave2 |
|---|---|---|---|---|---|---|---|---|---|---|
| bend | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jack | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jump | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pjump | 0 | 0 | 0 | 89 | 0 | 0 | 11 | 0 | 0 | 0 |
| run | 0 | 0 | 0 | 0 | 80 | 0 | 20 | 0 | 0 | 0 |
| side | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| skyp | 0 | 0 | 0 | 0 | 0 | 20 | 80 | 0 | 0 | 0 |
| walk | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| wave 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| wave 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

Table 1. Confusion matrix for the Weizmann dataset.

The proposed approach was also tested with the ViSOR dataset (Video Surveillance Online Repository dataset).[9] This dataset is composed of 5 activities, recorded in 150 real world surveillance videos. Each video was divided into two parts, to get more examples. A $k$- fold cross validation scheme was used: for each split, 60 % of the data were used for training and 40 % for testing, obtaining an averaged accuracy of, 96.7%. Results are shown in the confusion matrix (Table 2, top). Performance was also evaluated in terms of classical statistical indices (Table 2, bottom). The obtained results demonstrate both good performance using a very compact action descriptor.

In a second experiment, we evaluated the accuracy of our approach in an action recognition task for 5 long videos from the ViSOR dataset(each one $\sim$ 400 frames). Figure 4 shows the performance of our approach in a long video example w.r.t. ground truth (red line) for two actors. A first raw prediction (blue line), at each frame, achieves an average accuracy of 90.81% with a delay detection time of three frames. Looking at Figure 4,

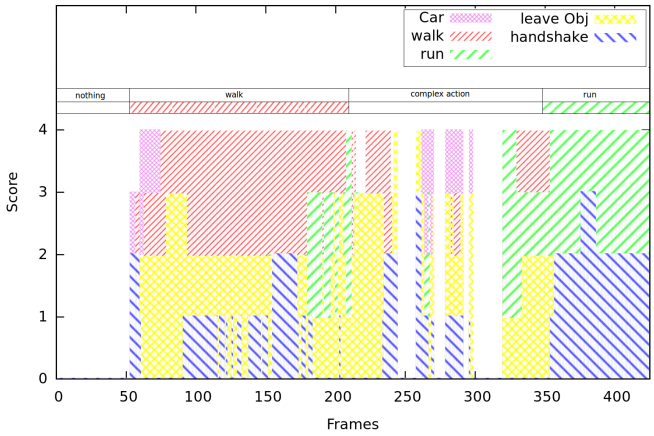


Figure 3. Multiclass SVM voting example for action recognition in a long video. The upper band represents the ground truth for this sequence.

| Category | gc | lo | w | r | h |
|---|---|---|---|---|---|
| get car | 100 | 0 | 0 | 0 | 0 |
| leave Object | 0 | 96.67 | 0 | 0 | 3.33 |
| walk | 0 | 0 | 91.65 | 8.35 | 0 |
| run | 2.38 | 0 | 0 | 97.62 | 0 |
| hand shake | 0 | 0 | 0 | 0 | 100 |

| Action | Acc | Sen | Spec | PPV | NVP |
|---|---|---|---|---|---|
| get car | 98.6 | 100 | 96.5 | 97.7 | 100 |
| leave Object | 98 | 96.7 | 100 | 100 | 95.2 |
| walk | 95 | 91.7 | 100 | 100 | 88.9 |
| run | 94.3 | 97.6 | 90.4 | 92.1 | 97.1 |
| hand shake | 97 | 100 | 92.2 | 95.2 | 100 |
| Average | 96.7 | 97.2 | 95.8 | 97 | 96.2 |

Table 2. Top: Confusion matrix for ViSOR dataset. w: walking, r: running, gc: get into a car, lo: leave an object, h: handshake. Every row represents a ground truth category, while every column represents a predicted category.

as expected, the major part of mistakes occurs when the motion descriptors is computed with few samples, specially during the transition between actions. Then, a time smoothing of the prediction (green line) was useful to improve the recognition rate. It consists in averaging the SVM votes for each class in a non causal interval $\Delta_t = [t-1, t+4]$ to get a more stable prediction. To take advantage of the stabilizations of the motion descriptor, the filter puts more weight in the future, a strategy whereby we achieved an average accuracy of 95.3%.

## 4. CONCLUSIONS AND PERSPECTIVES

This paper presented a novel approach for multiple human action recognition, using segmentation of the video flow in individual actions, multiple action representation using a multi-layer data structure, and action classification based on velocity orientation statistics. The motion descriptors can be computed on line and represented with a moderated amount of memory. Using a 96 dimension action descriptor, we achieved an averaged accuracy of 96.7% in human action classification of video sequences and 95.3% in frame level human action recognition. One advantage of the proposed motion descriptor is that it can be extended to mobile camera videos. In future works we will try to extend the recognition algorithm to mobile scenarios, and also build strategies that allow the recognition of more complex actions.

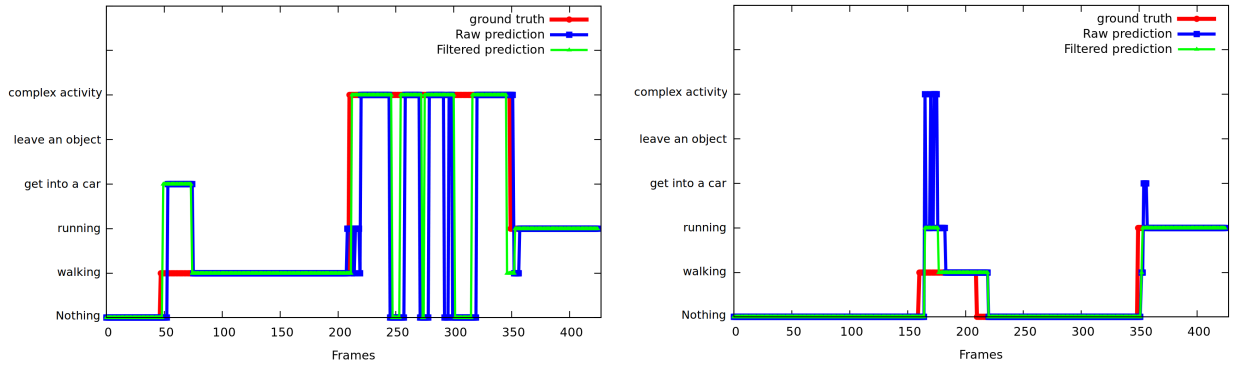


Figure 4. Action recognition prediction for a long video sequence with two actors (left: actor 1, right: actor 2). The red line represents the ground truth. The blue line stands for the raw prediction. The green line is the time filtered prediction. The "complex activity" corresponds to the moment when the 2 actors meet.

## REFERENCES


1. R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing* **28**(6), pp. 976 – 990, 2010.
2. J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys* **43**, pp. 16:1–16:43, Apr. 2011.
3. N. Ikizler, R. G. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms," in *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*, pp. 1–4, IEEE, 2008.
4. R. Chaudhry, A. Ravich, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *in In IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, 2009.
5. T. Cao, X. Wu, J. Guo, S. Yu, and Y. Xu, "Abnormal crowd motion analysis," in *Int. Conf. on Robotics and Biomimetics*, pp. 1709–1714, 2009.
6. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," pp. 357–360, 2007.
7. Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Int. Conf. on Computer Vision*, pp. 166–173, 2005.
8. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **29**, pp. 2247–2253, December 2007.
9. L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action categorization," in *Proc of ICCV. International Workshop on VOEC*, 2009.
10. A. Manzanera, "Local jet feature space framework for image processing and representation," in *Int. Conf. on Signal Image Technology & Internet-Based Systems*, 2011.
11. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Int. Conf. on Computer Vision & Pattern Recognition*, **2**, pp. 886–893, 2005.
12. C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks* **13**(2), pp. 415–425, 2002.
13. C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology* **2**, pp. 21–27, 2011.