A Motion Descriptor Based on Statistics of Optical Flow Orientations for Action Classification in Video-Surveillance

Fabio Martínez[1,2], Antoine Manzanera[1], and Eduardo Romero[2]

[1] Unité d'Informatique et d'Ingénierie des Systèmes, ENSTA-ParisTech
antoine.manzanera@ensta-paristech.fr
[2] CIM&Lab, Universidad Nacional de Colombia, Bogota, Colombia
{fmartinezc,edromero}@unal.edu.co

Abstract. This work introduces a novel motion descriptor that enables human activity classification in video-surveillance applications. The method starts by computing a dense optical flow, providing instantaneous velocity information for every pixel. The obtained flow is then characterized by a per-frame-orientation histogram, weighted by the norm, with orientations quantized to 32 principal directions. Finally, a set of global characteristics is determined from the temporal series obtained from each histogram bin, forming a descriptor vector. The method was evaluated using a 192-dimensional descriptor with the classical Weizmann action dataset, obtaining an average accuracy of 95 %. For more complex surveillance scenarios, the method was assessed with the VISOR dataset, achieving a 96.7 % of accuracy in a classification task performed using a Support Vector Machine (SVM) classifier.

Keywords: video surveillance, motion analysis, dense optical flow, histogram of orientations.

1 Introduction

Classification of human actions is a very challenging task in different video applications such as surveillance, image understanding, video retrieval and human computer interaction [1, 2]. Such task aims to automatically categorize activities in a video and determine which kind of movement is going on. The problem is complex because of the multiple variation sources that may deteriorate the method performance, such as the particular recording settings or the inter-personal differences, particularly important in video surveillance, case in which illumination and occlusion are uncontrolled.
Many methods have been proposed, coarsely grouped into two categories: (1) the silhouette based methods, and (2) the global motion descriptors (GMDs). The silhouette based methods aim to interpret the temporal variation of the human shape during a specific activity. They extract the most relevant silhouette shape variations that may represent a specific activity [7, 8]. These approaches achieve high performance in data sets recorded with static camera and simple background, since they

F.L. Wang et al. (Eds.): CMSP 2012, CCIS 346, pp. 267–274, 2012.

need an accurate silhouette segmentation, but they are limited in scenarios with complex background, illumination changes, noise, and obviously, moving camera.

On the other hand, the GMDs are commonly used in surveillance applications to detect abnormal movements or to characterize human activities by computing relevant features that highlight and summarize motion. For example, *3d* spatio temporal Haar features have been proposed to build volumetric descriptors in pedestrian applications [6]. GMDs are also frequently based on the apparent motion field (optical flow), fully justified because it is relatively independent of the visual appearance. For instance, Ikizler *et al* [3] used histograms of orientations of (block-based) optical flow combined with contour orientations. This method can distinguish simple periodic actions but its temporal integration is too limited to address more complex activities. Guangyu et al [5] use dense optical flow and histogram descriptors but their representation based on human-centric spatial pattern variations limits their approach to specific applications. Chaudhry *et al* [4] proposed histograms of oriented optical Flow (HOOF) to describe human activities. Our descriptor for instantaneous velocity field is very close from the HOOF descriptor, with significant differences that will be highlighted later, and the temporal part of their descriptor is based on time series of HOOFs, which is very different from our approach.

The main contribution of this work is a motion descriptor which is both entirely based on dense optical flow information and usable for recognition of actions or events occurring in surveillance video sequences. The instantaneous movement information, represented by the optical flow field at every frame, is summarized by orientation histograms, weighted by the norm of the velocity. The temporal sequence of orientation histograms is characterized at every histogram bin as some temporal statistics computed during the sequence. The resultant motion descriptor achieves a compact human activity description, which is used as the input of a SVM binary classifier. Evaluation is performed with the Weizmann [8] dataset, from which 10 natural actions are picked, and also with the ViSOR video-surveillance dataset [9], from which 5 different activities are used. This paper is organized as follows: Section 2 introduces the proposed descriptor, section 3 demonstrates the effectiveness of the method and the last section concludes with a discussion and possible future works.

2 The Proposed Approach

The method is summarized on Figure 1. It starts by computing a dense optical flow using the local jet feature space approach [10]. The dense optical flow allows to segment the region with more coherent motion in a RoI. A motion orientation histogram is then calculated, using typically 32 directions. Every direction count is weighted by the norm of the flow vector, so an important motion direction can be due to many vectors or to vectors with large norms. Finally, the motion descriptor groups up the characteristics of each direction by simple statistics on the temporal series, whose purpose is to capture the motion nature.

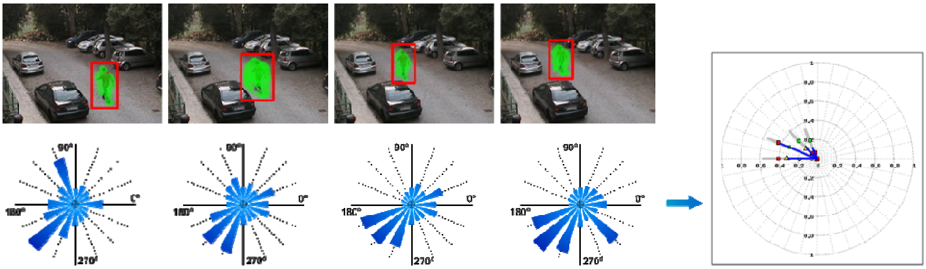


Fig. 1. General framework of the proposed method. First row: calculation of a dense optical flow. Second row: Orientation histograms representing the instantaneous velocities for every frame. Finally on the right, it is shown the descriptor made of temporal statistics of every histogram bin.

2.1 Optical Flow Estimation Using Local Jet Features

Several optical flow algorithms can be used within our method. They need to be dense and globally consistent, but not necessarily error-free, nor particularly accurate in terms of localization. In our implementation, we used the optical flow estimation based on the nearest neighbor search in the local jet feature space, proposed in [10]. It consists in projecting every pixel to a feature space composed of spatial derivatives of different orders and computed at several scales (the local jet): $f_{ij}^{\sigma} = f * \frac{\partial^{i+j} G_{\sigma}}{\partial x^{i} \partial y^{j}}$, where σ, the standard deviation of the $2d$ Gaussian function G_{σ} represents the scale, and $i + j$ the order of derivation. For each frame t and every pixel x, the apparent velocity vector $\boldsymbol{V}_t(x)$ is estimated by searching the pixel associated to the nearest neighbor in the space of local jet vectors calculated at frame $t - 1$. The interest of this method is to provide a dense optical flow field without explicit spatial regularization, and an implicit multi-scale estimation by using a descriptor of moderate dimension for which the Euclidean distance is naturally related to visual similarity. In our experiments, we used 5 scales, with $\sigma_{n+1} = 2\,\sigma_n$, and derivatives up to order 1, resulting in a descriptor vector of dimension 15.

2.2 Motion RoI Segmentation

The dense optical flow can be used for a coarse spatial segmentation of potential human actions at each frame. First a binary morphological closing operation is performed on pixels whose velocity norm is above a certain threshold, to connect close motion regions. The resulting connected components may also be grouped according to a distance criterion, and the bounding boxes of the remaining connected components form the motion RoIs. We use this simple segmentation to eliminate noisy measurements outside the moving character (Single actions are considered in these experiments).

2.3 Velocity Orientations Histogram

The next step consists in coding the distribution of instantaneous motion orientations. For a non-zero flow vector $\boldsymbol{V}$, let $\phi(\boldsymbol{V})$ denotes its quantized orientation. Based on the HOG descriptor [11], we compute the motion orientation histogram of each frame as the relative occurrence of flow vectors within a given orientation, weighted by the vector norm:

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

where $\omega \,\epsilon\, \{\omega_o \dots \omega_{N-1}\}$. ω_N the number of orientations was set to 32 in our experiments. This part of our descriptor, dealing with instantaneous velocity information, is almost identical to the HOOF descriptor of Chaudhry *et al* [4], except that the HOOF descriptor is invariant under vertical symmetry, i.e. it does not distinguish the left from the right directions. This property makes the HOOF descriptor independent to the main direction of transverse motions, but it also reduces its representation power, missing some crucial motion information, like antagonist motions of the limbs. For this reason, we chose to differentiate every direction of the plane, the invariance w.r.t. the global motion direction being addressed at the classification level.

2.4 Motion Descriptor

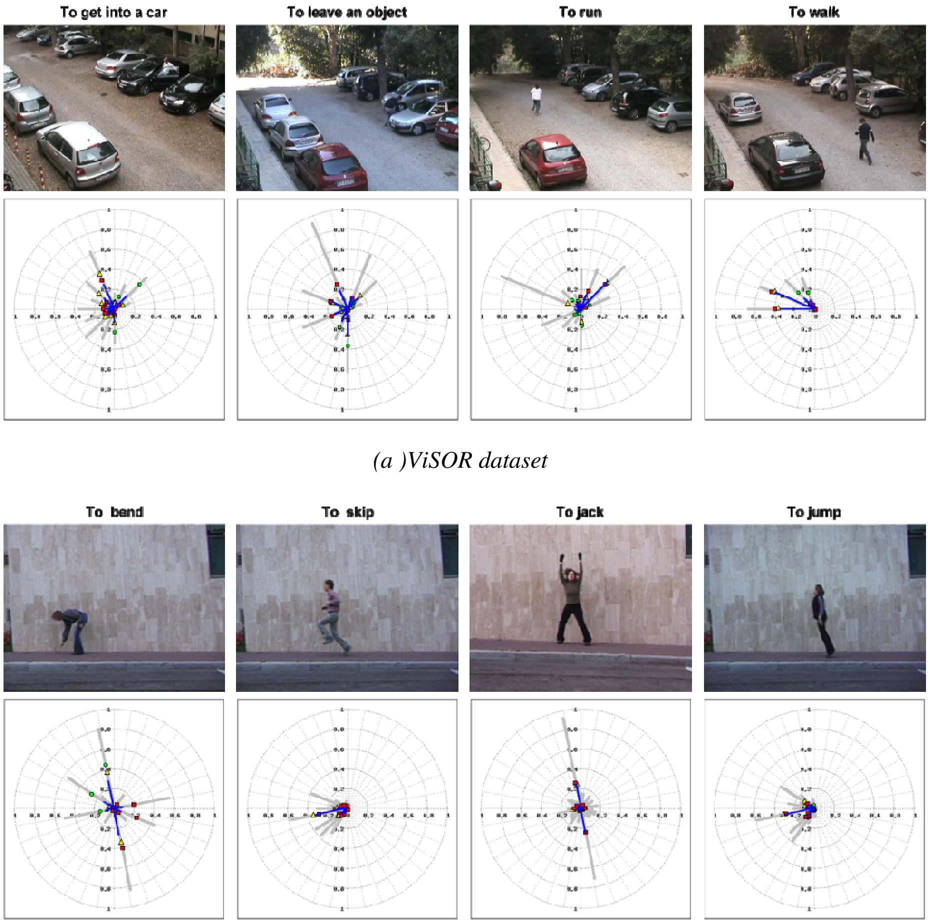
Finally, a description vector is computed to capture the relevant motion features. For n frames, it consists in a set of temporal statistics computed from the time series of histogram bins $H_t(\omega)$, as follows:

1. *Maximum*: $M(\omega) = \max_{\{0\le t<n\}}\{H_t(\omega)\}$
2. *Mean:* $\mu(\omega) = \sum_{\{0\le t<n\}} \frac{H_t(\omega)}{n}$
3. *Standard deviation:* $\sigma(\omega) = \sqrt{\sum_{\{0\le t<n\}} \frac{{H_t}^2(\omega)}{n} - \mu(\omega)^2}$

We also split the sequence into 3 intervals of equal durations and compute the corresponding means as follows:

4. *Mean Begin:* $\mu_b(\omega) = \sum_{\{0\le t<\frac{n}{3}\}} \frac{H_t(\omega)}{n/3}$
5. *Mean Middle:* $\mu_m(\omega) = \sum_{\{\frac{n}{3}\le t<\frac{2n}{3}\}} \frac{H_t(\omega)}{n/3}$
6. *Mean End:* $\mu_e(\omega) = \sum_{\{\frac{2n}{3}\le t<n\}} \frac{H_t(\omega)}{n/3}$

Some examples of human activities and their associated motion descriptor in the two datasets are shown in Figure 2. For each motion descriptor, the blue and gray lines respectively represent the maximum and mean values. The red square, yellow triangle and green disk represent the mean values for the beginning, middle and end portion of the sequence respectively. For readability purposes, the standard deviation is not displayed here. It turns out from our experiments that the aspect of the descriptor is visibly different for distinct human activities.



(a)ViSOR dataset

(b) Weizmann dataset

Fig. 2. Example of motion descriptors for human activities

2.5 SVM Classification

Classification was performed using a bank of binary SVM classifiers. The SVM classifier has been successfully used in many pattern recognition problems given its robustness, applicable results and efficient time machine. In our approach, we use the *one-against-one SVM multiclass classification* [12], where given k motion classes, $\frac{k(k-1)}{2}$ classifiers are built and the best class is selected by a voting strategy. The SVM model was trained with a set of motion descriptors, extracted from hand labeled human activity sequences (see next section). The Radial Basis Function (RBF) kernel was used [13].

3 Evaluation and Results

Our approach was evaluated in two datasets: The Weizmann dataset [14] that it is commonly used for human action recognition and the VISOR dataset [9], which is a real world surveillance dataset. Performance on each dataset was assessed using a leave-one-out cross validation scheme, each time selecting a different single action sequence, as described in the literature by previous human action approaches [15, 16]. A first evaluation was done over the Weizmann dataset [14]. This dataset is composed of 9 subjects and 10 actions recorded in 93 sequences. The classes of actions are "run", "walk", "skip", "jumping-jack" (jack), "jump-forward-on-two-legs" (jump), "jump-in-place-on-two-legs" (pjump), "gallop-sideway" (side), "wave-two-hands" (wav2), "wave-one-hand"(wav1) and "bend". The corresponding confusion matrix for the Weizmann dataset is shown in Table 1. Our approach achieves a 95 % of accuracy, which is comparable to results reported in the literature.

Table 1. Confusion matrix for the Weizmann dataset. Every row represents a ground truth category, while every column represents a predicted category.

Category	bend	jack	jump	pjump	run	side	skip	walk	wav1	wav2
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
skip	0	0	0	0	0	20	80	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wav1	0	0	0	0	0	0	0	0	100	0
wav2	0	0	0	0	0	0	0	0	0	100

A second test was carried out with a dataset for human action recognition from a real world surveillance system (ViSOR: the Video Surveillance Online Repository) [9], which has been less used in the literature, but is more representative for video-surveillance applications. This dataset is composed of 150 videos, captured with a stationary camera, showing 5 different human activities: walking, running, getting into a car, leaving an object and people shaking hands, four of them shown in Figure 2. The high variability of this dataset is challenging: each activity is performed by several actors with different appearance, the background scene is usually different, and the motion direction, starting and halting points locations may be different for every video sequence. Evaluation was performed with 32 directions, corresponding to a descriptor dimension of 192, and obtaining an averaged accuracy of 96.7 %. Results obtained are shown in the confusion matrix (Table 2, top).

Table 2. Top: Confusion matrix. Row: ground truth / Column: predicted category. For example, row 4 means that out of all the “run” sequences, 96.43 % were classified correctly, and 3.57% were classified as “get into a car”. Bottom: Statistical indices measured for each category.

Category	get car	leave object	walk	run	hand shake
get car	100	0	0	0	0
leave object	0	95	0	0	5
walk	0	0	92	8	0
run	3.57	0	0	96.43	0
hand shake	0	0	0	0	100

Action	Accuracy	Sensitivity	specificity	PPV	NPV
get car	98	100	94.9	96.6	100
Leave object	97	95	100	100	93
walk	95	91.7	100	100	88.9
run	94	96.4	90.4	92	95.6
hand shake	97	100	92.2	95.2	100
Average	96.2	96.6	95.5	96.8	95.5

The performance was also evaluated in terms of classical statistical indices (Table 2, bottom). Let *TP, TN, FP* and *FN* be the number of true positive, true negative, false positive and false negative, respectively, associated to each label. The *accuracy* is $Acc = \frac{TP+TN}{TP+TN+FP+FN}$, the *Sensitivity* is $Sen = \frac{TP}{TP+FN}$, the *specificity* is $Spec = \frac{TN}{TN+FP}$, the *Positive Predictive Value* is $PPV = \frac{TP}{TN+FP}$ and the *Negative Predictive Value* is $NPV = \frac{TN}{TN+FN}$. The obtained results demonstrate both good performance and a significant degree of confidence, using a very compact action descriptor of dimension 192. Our approach was also tested on the KTH dataset [14] but the results were significantly worse, with accuracy around 90 %. This is mainly due to a limitation of the local jet based dense optical flow, which needs enough scale levels to be effective and then provides poor results when the resolution is too low.

4 Conclusions and Perspectives

A novel motion descriptor for activity classification in surveillance datasets was proposed. A dense optical flow is computed and globally characterized by per-frame-orientation histograms. Then a global descriptor is obtained using temporal statistics of the histogram bins. Such descriptor of 192 characteristics to represent a video sequence was plugged into a bank of SVM binary classifiers, obtaining an average accuracy of 96.7 % in a real world surveillance dataset (ViSOR). For the classical human action dataset (Weizmann) our approach achieves a 95 % of accuracy.

A great advantage of the presented approach is that it can be used in sequences captured with a mobile camera. Future work includes evaluation on more complex scenarios. We also plan to adapt this method to perform on line action recognition system, by coupling it with an algorithm able to segment the video in space × time boxes containing coherent motion.

Acknowledgements. This work takes part of a EUREKA-ITEA2 project and was partially funded by the French Ministry of Economy (General Directorate for Competitiveness, Industry and Services).

References

1. Aggarwal, et al.: Human activity analysis: A review. ACM Computing Surveys 43, 1–43 (16), 3 (2011)
2. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28, 976–990, 6 (2010)
3. Ikizler, N., et al.: Human Action Recognition with Line and Flow Histograms. In: 19th International Conference on Pattern Recognition (ICPR), Tampa, FL (2008)
4. Chaudhry, et al.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1932–1939 (2009)
5. Zhu, et al.: Action recognition in broadcast tennis video using optical flow and support vector machine, pp. 89–98 (2006)
6. Ke, Y., et al.: Efficient Visual Event Detection Using Volumetric Features. In: Int. Conf. on Computer Vision, pp. 166–173 (2005)
7. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 249–257, 104 (2006)
8. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2247–2253, 12 (2007)
9. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Effective Codebooks for Human Action Categorization. In: Proc. of ICCV. International Workshop on VOEC (2009)
10. Manzanera, A.: Local Jet Feature Space Framework for Image Processing and Representation. In: Int. Conf. on Signal Image Technology & Internet-Based Systems (2011)
11. Dalal, N., et al.: Histograms of Oriented Gradients for Human Detection. In: Int. Conf. on Computer Vision & Pattern Recognition, pp. 886–893, 2 (2005)
12. Hsu, et al.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13, 415–425, 2 (2002)
13. Chang, C., et al.: LIBSVM: A library for support vector machines. ACM Trans. on Intelligent Systems and Technology 2, 21–27, 3 (2011)
14. Schuldt, et al.: Recognizing Human Actions: A Local SVM Approach. In: Proceedings of the 17th International Conference on Pattern Recognition, pp. 32–36, 3 (2004)
15. Wong, S.F., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: Proc. IEEE Int. Conf. Computer Vision, pp. 1–8, 2 (2007)
16. Ballan, et al.: Human Action Recognition and Localization using Spatio-temporal Descriptors and Tracking (2009)