Motion histogram quantification for human action recognition

Hedi Tabia, Michèle Gouiffès and Lionel Lacassagne
IEF, Institut d'Electronique Fondamentale
Bat 220, Campus Scientifique d'Orsay
91405 Orsay, FRANCE

Abstract

In this paper, we propose an approach for human activity categorizing based on the use of optical flow direction and magnitude features. The main contribution of this paper is the feature representation that mirrors the geometry of the human body and relationships between its moving regions when performing activities. The features are quantified using a quantization algorithm. We analyze the performance of two well-known classifiers: the Naïve Bayes and the SVM. The results show the effectiveness of our approach.

1. Introduction

Human action recognition and understanding has become quite popular and has gained a lot of interest during the past decade. Many applications of human action recognition from videos can be found such as video-surveillance, entertainment, user interfaces, sports and video annotation domains. The goal of an action recognition system is to identify simple actions of daily life (like walking, running, jumping ...) from referring videos. These actions correspond to models of simple movements performed by a single person in a short laps of time.

Over the recent years, many techniques have been proposed for human action recognition and understanding that are described in comprehensive surveys [11, 12]. Among the works of the state of the art which are directly related to this paper, we find the work of Ali et al. [1] who propose a set of kinematic features that are derived from the optical flow. The authors use a multiple instance learning (MIL) method to classify human actions. Kosmopoulos et al. [6] proposed a framework for visual behavior understanding. Their approach is based on the utilization of holistic visual behavior understanding methods, which perform modelling directly at the pixel level. Their system uses the information provided by multiple cameras. Dollár et al. [4] developed a behavior recognition system based on sparse spatio-temporal features which they called (*cuboid features*). Laptev and Lindeberg [7] presented a method for local spatio-temporal feature extraction and applied their method to human action recognition. They demonstrate how their velocity adapted features enable recognition of human actions in situations with unknown camera motion and complex, non stationary backgrounds. Most works are evaluated on the dataset KTH [7] available on the web[1].

In this paper, the analysis focuses on video sequences recorded by monocular camera because they are far less resource-intensive and more economical than a multiple camera system. The proposed method corresponds to a meta-algorithm which combines the human detection and the action recognition. For action recognition, shape and optical flow histograms are used together with *Bag of Feature* BoF classifiers. Its advantages are its genericity since both detection and recognition are performed, and its computational efficiency due to the simplicity of the descriptors. The remainder of the paper is organized as follows. The human shape and motion descriptors are detailed in Section 2 while the classification is the topics of section 3.1. Finally, the experiments are presented in Section 4. Conclusions and future developments end the paper.

2 The Method

Our method consists in representing a human action based on the motion features derived from each video sequence. The method encompasses four main steps. 1) The first step is the construction of motion histograms of a human action video. This

[1] http://www.nada.kth.se/cvap/actions/

step is divided into two stages namely human figure localizing and motion histogram computing. The proposed motion histogram encodes the geometry of the human body and relationships between its moving regions based on the optical flow vectors. 2) The second step consists in assigning motion histograms to a set of predetermined clusters (an alphabet set) with a vector quantization algorithm, 3) The third one is the construction of a bag of "keymotions", which takes account of the number of motion histograms assigned to each cluster and finally apply a multi-class classifier, treating the bag of "keymotions" as the feature vector, and thus determine which class or categories to assign to the human action to be classified. In order to increase classification accuracy and decrease the computational effort, the motion histograms constructed in the first step should be sufficiently rich to discriminate among the various classes at the category level. By analogy with "keywords" in text categorization, we refer to the quantized feature vectors (cluster centers) as "keymotions".

2.1 Human figure localizing

First of all, a video preprocessing step is needed in order to localize human figures in each frame of the video sequence.

Although most videos used for action recognition [7, 10] show the ideal case of a static background and one single human, most video surveillance applications require to first detect all the possible humans in the scene and filter out the possible undesired motions. In addition, when the camera is moving, when embedded on a vehicle for example, the background subtraction by motion analysis is inappropriate. By cons, the appearance-based detectors can answer this issue. Given a video sequence where human activities occur, the aim of our preprocessing step is to pick out the sequence of figure-centric bounding boxes. We apply an approach similar to [2, 3] for human figure stabilization.

Our human figure stabilizer is based on a linear SVM classifier which is trained with Histograms of Oriented Gradients (HOG) [3] descriptors extracted from manually cropped figure-centric bounding boxes and negative examples from random patches around the figures. A multi-scale human detection is performed in the local neighborhood defined by the previously detected location.

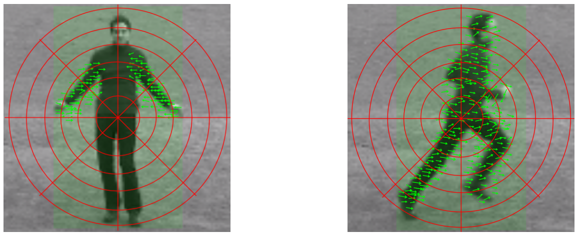
2.2 Motion histogram computing

Both shape and motion features are used to characterize human activities. The proposed representation encodes the geometry of the human body and the relationships between the body moving regions in the image space. Therefore, the proposed feature vector well reflects the motion behavior of specific regions of the human body. Formally as shown in Figure 1, we consider the center of the body-centric bounding box as the origin of a local coordinate system. The bounding box is split according to a polar coordinate system to 48 bins (6 magnitudes $\times$ 8 directions). A descriptor is constructed using the optical flow values in each bin, in a similar fashion as in [5]. For each i_{th} bin in our descriptor, we define optical flow histogram $h_i(\theta)$ such that

$$h_i(\theta) = \sum_{j \in B_i} f(u_\theta \cdot V_j) \tag{1}$$

where V_j represents the flow value in each pixel j, B_i is the set of pixels in the spatial bin i, u_θ is the unit vector in θ direction $\{0, 90, 180, 270\}$ and f function is defined as

$$f(x) = \left\{ \begin{array}{c} 0 \text{ if } x \leq 0 \\ x \text{ if } x > 0 \end{array} \right\} \tag{2}$$



(a) From Hand clapping

(b) From Running

Figure 1. Optical flow vectors from two different human action videos.

3 Motion alphabet

In BoF approach, the alphabet is obtained by quantification of the set of descriptors extracted in the training stage. The alphabet is used to construct discriminant representatives, with which any human action can be described. The most common method to build the action alphabet is to arrange histograms encountered in the training stage into a finite number of clusters using a clustering algorithm. The alphabet size is given by the number of the clusters. For this end, we chose to use the k-means algorithm. It proceeds by iterated assignments of points to their closest cluster centers and re-computation of the cluster centers. In our method, the distance between two points (motion histograms) is measured using the χ^2 distance. We run k-means several times with different number of desired representative vectors (k) and different sets of initial cluster centers. We select the final clustering giving the lowest empirical risk in categorization.

3.1 Action Recognition

After assigning each motion histogram to its closest cluster, the problem of action recognition can be scaled down to that of multi-class supervised learning. In order to make a decision about an action to be recognized the system performs two steps: training and testing. The purpose of the training is to achieve correct recognition of future action sequences. Based on knowledge learned on labeled data, the system make a decision rule for distinguishing categories of human actions. By applying this decision rule on the action to be recognized, the system predicts the class of that action. In this paper, we analyze the behavior of two well-known classifier: The Naïve Bayes and the Support Vector Machine.

3.1.1 Naïve Bayes classifier

The Naïve Bayes Classifier [9] is a probabilistic classifier based on the Bayesian theorem. To demonstrate the concept of action recognition using Naïve Bayes classifier, let us assume we have a set of labeled sequences of human actions $S = \{S_i\}$ and and an alphabet $A = \{a_t\}$ of representative keymotions. Each motion histogram extracted from a video sequence is labeled with the keymotion to which it lies closest in the motion space. We count the number $N(t, i)$ of times the keymotion a_t occurs in a video sequence S_i. To recognize a new action, we apply Bayes's rule and take the largest a posteriori score as the prediction:

$$P(C_j/S_i) \propto P(S_i/C_j)P(C_j) = P(C_j)\prod_{t=1}^{|A|} P(a_t/C_j)^{N(t,i)}. \tag{3}$$

It is evident in this formula that Naïve Bayes requires estimates of the class conditional probabilities of keymotion a_t given an action category C_j. In order to avoid probabilities of zero, these estimates are computed with Laplace smoothing:

$$P(a_t/C_j) = \frac{1 + \sum_{S_i \in C_j} N(t,i)}{|A| + \sum_{s=1}^{|A|} \sum_{S_i \in C_j} N(s,i)}. \tag{4}$$

3.1.2 SVM classifier

The SVM classifier is a classification method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels with maximal margin [13]. In order to apply the SVM to multi-class problems we take the one-against-all approach. Given an m class problem, we train m SVM's, each distinguishes video from some action category i from videos from all the other $m - 1$ categories j not equal to i. Given a video sequence of a human action to be classified, we assign it to the class with the largest SVM output.

4 Experiments and Results

In this section, we give results from two experiments. In the first experiment, we analyze the performance of the Naïve Bayes and the SVM classifier on classifying human actions. In the second experiment, we present results compared with related works. These experiments were conducted on a standard dataset containing a variety of daily life actions. The performance of our method using both classifiers was evaluated with 10-fold cross validation.

4.1 Dataset description

The KTH [7] is a dataset that contains low resolution videos (160×120 pixels) of six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 different subjects. This dataset is challenging because the sequences are recorded in different indoor and outdoor scenarios with scale variations and different clothes.

Boxing	Hand waving	Jogging	Walking	Hand clapping	Running
32	10	0	5	52	0
5	92	0	0	2	0
2	2	80	7	2	7
2	1	4	91	2	1
16	0	0	5	77	0
1	0	17	1	0	80

Table 1. Confusion Matrix (Naïve Bayes)

4.2 Results from Naïve Bayes classifier

Table 1 shows the performance of our method using the Naïve Bayes classifier. In this visualization, we display the values corresponding to the confusion matrix. The diagonal elements are the counts

Boxing	Hand waving	Jogging	Walking	Hand clapping	Running
54	1	0	0	44	0
3	95	0	0	2	1
0	0	86	5	1	8
0	0	2	98	1	0
19	0	1	0	79	0
0	0	16	1	1	83

Table 2. Confusion Matrix (SVM)

Method	Performance rate
Our method (SVM)	**82.36 %**
Space-time interest Points [8]	80 %
Our method (Naïve Bayes)	**75.83 %**
Velocity histories [10]	74 %
Spatio-temporal Cuboids [4]	66 %

Table 3. Comparison with related work

of the correct predictions. Naïve Bayes classifier gives 75.83% as a performance rate.

4.3 Results from SVM classifier

Results from applying the SVM are presented in Table 2. As awaited the SVM performance surpass the performance of Naïve Bayes classifier, reducing the overall error rate from 24.16% to 17.63%. We compared linear and rbf kernels SVM's and found that rbf kernels based method gave the best performance. One can notice from both confusion matrices shown in Table 1 and Table 2, that the two classifiers have a similar behavior.

4.4 Comparison with related work

In order to evaluate our approach, we compare its performance with some state-of-the-art human action recognition algorithms. The performance is evaluated in terms of classification rate (i.e. the percentage of action sequences which are correctly classified). Table 3 shows that our approach gives a comparable performance with other methods competing the contest KHT.

5 Conclusion

We have presented an effective action recognition system that relies on a direction and magnitude histogram of the dense optical flow vectors. A vector quantization technique is used to construct discriminant representatives, with which human activities are described. In this paper, we analyzed the behavior of two well-known classifier: The Naïve Bayes and the Support Vector Machine. First, the experimental results show that the SVM performances surpass the performances of Naïve Bayes classifier. Then, our representation of the action, although simple, outperforms some existing approaches based on some more heavy data structures.

Acknowledgment

This research is supported by the European Project ITEA2 SPY Surveillance imProved sYstem http://www.ppsl.asso.fr/spy.php.

References

[1] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:288–303, February 2010.

[2] C.-C. Chen and J. Aggarwal. Modeling human activities as speech. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, june 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, june 2005.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65 – 72, oct. 2005.

[5] N. Ikizler, R. Cinbis, and P. Duygulu. Human action recognition with line and flow histograms. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, dec. 2008.

[6] D. Kosmopoulos and S. Chatzis. Robust visual behavior recognition. *Signal Processing Magazine, IEEE*, 27(5):34 –45, sept. 2010.

[7] I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. 2004.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008.

[9] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. pages 4–15, 1998.

[10] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009.

[11] R. Poppe. A survey on vision-based human action recognition. *Image Vision Comput.*, 28:976–990, June 2010.

[12] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18, 2008.

[13] V. N. Vapnik. *Statistical learning theory.* Sept. 1998.