Human Action Recognition Using Temporal Segmentation and Accordion Representation

Manel Sekma, Mahmoud Mejdoub, and Chokri Ben Amar

REGIM: Research Group on Intelligent Machines Engineering
National School of Sfax (ENIS), 3038 Sfax, Tunisia
{manel_sekma,mah.mejdoub,chokri.benamar}@ieee.org

Abstract. In this paper, we propose a novel motion descriptor Seg-SIFT-ACC for human action recognition. The proposed descriptor is based both on the accordion representation of the video and its temporal segmentation into elementary motion segments. The accordion representation aims to put in space adjacency the columns of the video frames having a high temporal correlation. For complex videos containing many different elementary actions, the accordion representation may put in spatial adjacency temporally correlated pixels that belong to different elementary actions. To surmount this problem, we divide the video into elementary motions segments and we apply the accordion representation on each one separately.

Keywords: Human Action Recognition, Accordion Image, Motion Segment, Motion Descriptor.

1 Introduction

Recognizing human actions in realistic uncontrolled video is an important and challenging topic in computer vision. However, in recent years, many different space-time feature detectors (Harris3D [1], Cuboids [2] and Hessian [3]) and descriptors (HOG (Histogram of Oriented Gradients)/HOF (Histogram of Optical Flow) [4], Cuboids [2] and Extended SURF [3]) have been proposed in the state-of-the art. Feature detectors usually select the most salient Spatio-Temporal locations. Feature descriptors detect shape and motion information in the neighborhoods of selected points using usually spatial and temporal image gradients as well as optical flow. The motion descriptors are well suited to describe the human actions [5]. HOF descriptors characterize local motions. They are computed by dividing the space time neigborhood of the Harris3D interest points into small space-time regions and accumulating a local 1-D histogram of optic flow over the pixels of each sub-region. Dalal et al.[6] proposed the motion boundary histograms (MBH) descriptor for human detection. The MBH descriptor describes the relative motion between pixels by computing the gradient of the optical flow. In [5], MBH is used as motion descriptor for dense trajectories.

Recently there was a growing trend of using temporal video segmentation as preprocessing for action recognition [7,8]. It was hoped that segmentation

R. Wilson et al. (Eds.): CAIP 2013, Part II, LNCS 8048, pp. 563–570, 2013.

methods could partition videos into coherent constituent parts, and recognition could then be simply carried out based on the obtained segments. Carlos and al [7] propose a modelling temporal structure of decomposable motion segments for activity classification. They use a discriminative model that encodes a temporal decomposition of video sequences, and appearance models for each motion segment. In [8] Qiang and Gang propose a new representation of local spatio-temporal cuboids based on atomic actions that represent the basic units of human actions.

In our previous work [9], we presented a motion descriptor based on the accordion representation of the video frames. The descriptor was computed around moving points extracted with the KLT tracker. Experiments were carried out on the Weizman dataset in which each video contains only one simple action. The accordion representation transforms the video in an image that allows to put pixels which have a temporal adjacency in spatial neighbourhood. But for complex videos containing many different elementary actions, the accordion representation risk to put in spatial adjacency temporally adjacent pixels that belong to two semantically different motion segments. To surmount this problem, we propose in this work to split the video into elementary motion segments using the EHD [10,11] histogram comparison between successive frames in the video. Afterwards, the accordion representation is applied separately on each elementary motion segment. To describe the motion information, Harris3D interest points are detected on the frames of the motion segment and projected onto the image accordion relative to this segment. After that, SIFT [12] descriptor is computed around the projected Harris3D interest point. Experiments are carried out into two action recognitions datasets (hollywood2 and Olympic sports) formed by complex videos that contain different elementary actions. This paper is organized as follows: in Section 2 an overall description of the proposed motion descriptor is given. The experimental validation and results are given in section 3 and section 4. Finally, concluding remarks are presented.

2 Description of the Proposed Motion Descriptor

The graphical description of our motion descriptor (Seg-SIFT-ACC) computation is illustrated by Figure 1. Firstly, we decompose the video sequence into motion segments. After that, we create an Accordion image for every motion segment. For each motion segment, we proceed by the detection of Harris3D interest points in the video frames. Harris 3D interest points detected on a motion segment are projected onto the I_ACC relative to this motion segment. Then, for each I_ACC, we compute SIFT descriptors around Harris 3D interest points. Afterwards, we employ the bag-of-features model [13,14,15,16] on the Seg-SIFT-ACC descriptors in order to obtain a histogram of visual word occurrences for each video sequence. For classification, we use the supervised learning algorithm SVM (Support Vector Machines) [17]. The Accordion representation is presented in section 2.1. The temporal segmentation method is described in section 2.2. In section 2.3, we present the computation steps of the proposed motion descriptor.

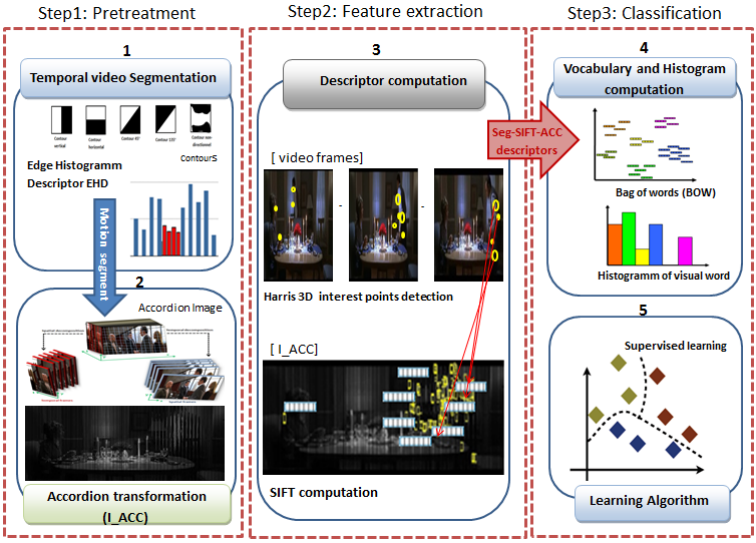

Fig. 1. Description of the proposed framework

2.1 Accordion Representation

The accordion representation [9] aims to put in spatial adjacency the pixels having a high temporal correlation. It is built by carrying out the temporal decomposition of the signal video. In a first stage, the video is transformed into temporal frames (Figure 2a). Each one represents a 2D image that collects the video pixels having the same column rank in all video frames. In a second stage (Figure 2b), the temporal frames are successively projected onto a plane called the Accodrion Image (I_ACC) througout this work. Hence, Accordion transformation tends to transform temporal correlation in the original video source into a spatial correlation in the resulting 2D image (I_ACC). The dimension($H\_acc \times W\_acc$) of I_ACC is:

$$\begin{pmatrix} H\_acc = & H \\ W\_acc = W * & NbF \end{pmatrix} \tag{1}$$

where H_acc (height) and W_acc (width) are the frame sizes; NbF is the number of video frames. Each point position (x, y) on every video frame i is projected onto the I_ACC using the Equation 2 that calculates the I_ACC coordinates (x_ACC, y_ACC) of the projected point. (x_ACC, y_ACC) is obtained such as x_ACC is equal to x and y_ACC is equal to the position given to the frame column y in the I_ACC.

$$Projection : \begin{vmatrix} video & \rightarrow & \text{I\_ACC} \\ (x, y, i) & \mapsto & (x\_ACC = x, y\_ACC = i + NbF * (y - 1)) \end{vmatrix} \tag{2}$$

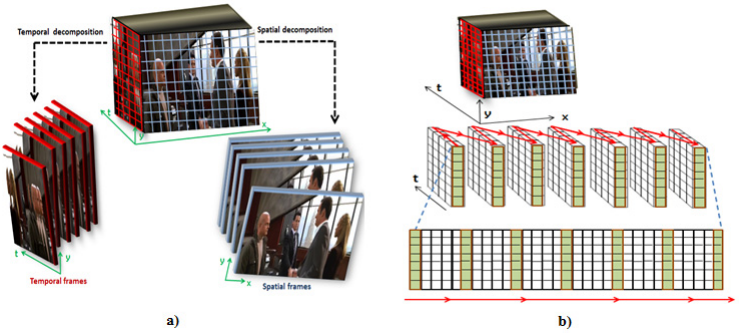

Fig. 2. The method of accordion representation: a) Video decomposition b) Video transformation into an Accordion image

2.2 Temporal Segmentaion

Video Segmentation Method. For complex videos containing different elementary actions, the accordion representation risk to put in spatial adjacency temporal adjacent columns that belongs to two adjacent elementary actions. To surmount this problem, we segmented the video into elementary segments. Then, an I_ACC is created for each segment separately (Figure 3). To segment the video, we compute the difference between the histogram of successive frames [18]. If the distance is above a threshold T, a cut is declared. This cut specifies start and end frames of actions in the video sequence (Figure 3). To compute the histogram, we use the Edge Histogram Descriptor (EHD) [10,11] since it can be efficiently utilized for image matching. The EHD descriptor represents the local edge distribution in the image.

As described in section 2.1, the size of the I_ACC is given by $W\_ACC = NbF \times W$, the I_ACC construction needs a high memory consumption especialley for large videos. Splitting the video into elementary units also provides a memory usage reduction.

2.3 Descriptor Computation

The Seg-SIFT-ACC descriptor is based on the computation of the histogram of gradient orientations in every local patch of the I_ACC. It reflects the motion variation along the temporal axis of the video. In a first step, we transform each motion segment into an I_ACC. After that, we project the detected Harris 3D interest points into the I_ACC (Figure 3). Afterwards, we define 16×16 patches in the I_ACC on the spatial neighbourhood of the projected Harris3D interest points. To capture the motion information from the I_ACC, SIFT descriptors are computed from the 16×16 patchs. For that, every patch is sub-divided into 4×4 sub-regions. From each sub-region, an orientation histogram with 8 bins is computed, where each bin covers 45 degrees. Each sample in the sub-region is added to the histogram bin and weighted by its gradient magnitude. The 16

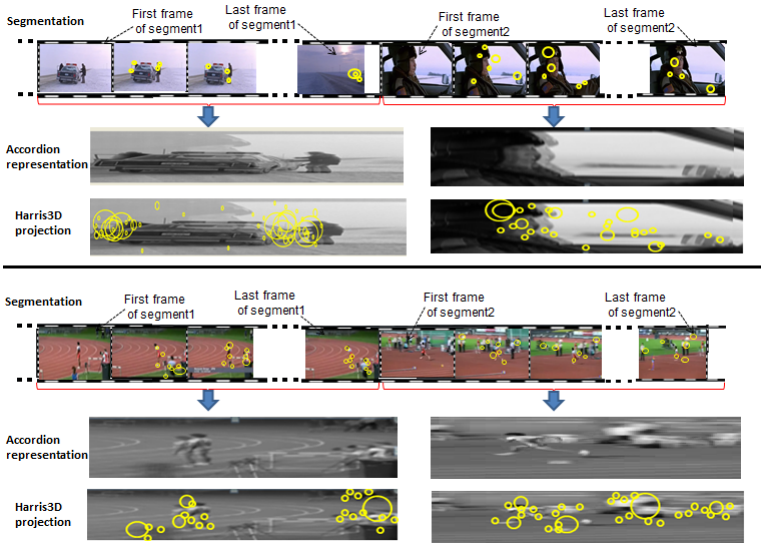

Fig. 3. Projection of the Harris3D interest points on the I_ACCs of the motion segments for two video examples

resulting orientation histograms are transformed into $128d$ vector. Finally, the vector is normalized to unit length to achieve the invariance against illumination changes. Thus we obtain our proposed Seg-SIFT-ACC descriptor.

3 Experimental Validation

3.1 Action Recognition Datasets

Hollywood2 Dataset: The Hollywood2 dataset [19] has been collected from 69 different Hollywood movies. There are 12 action classes. In total, there are 1707 action samples divided into a training set and a test set. The performance is evaluated by computing the average precision (AP) for each of the action classes and reporting the mean AP over all classes (mAP) as suggested in [19].

Olympic Sports Dataset: The Olumpic sport dataset [7] consists of athletes practicing different sports. There are 16 sports actions represented by a total of 783 video sequences, divided into a training set and a test set. Mean average precision over all classes is reported.

3.2 Implementation Details

The threshold T used to segment the video into motion segments is computed applying the cross validation method on the training set. We obtain the best recognition accuracy with value of T equal to 200, 120 respectively for the Hollywood2 and Olympic sports. To implement the bag of features model, we use

an identical pipeline as described in [20]. For that, we cluster a subset of 100, 000 randomly selected training features with the k-means algorithm. We fix the number of visual words per descriptor to 4000 which has shown [20] to empirically give good results for a wide range of datasets and descriptors. To increase precision, we initialize k-means 8 times and keep the result with the lowest error. For classification, we use a non-linear support vector machine (SVM) with a multi-channel χ^2 kernel [17].

4 Results

In table 1, we present our result per class of Hollywood2 action and we compare our descriptor to SIFT-ACC (obtained with the same method but without segmentation) and the state-of-the-art methods. Heng and al.[5] have achieved 55.1% using the Motion Boundary Histograms (MBH) to describe the video with a dense trajectory information. Qiang and Gang in [8] propose to use a spatio-temporal cuboid based on atomic actions. where atomic actions are basic units of humain actions. They obtain mAP equal to 49.4%. Gilbert et al. propose a hierarchical approach for constructing and selecting discriminative compound features of 2D Harris corners which gives a mAP equal to 50.9%. In [21], Quoc et al present a human action recognition method that learns features from Spatio-Temporal data using independent subspace analysis. This method gives 53.3%. An extension to the standard BoW approach is presented in [22] by locally applying BoW on regions that are spatially and temporally segmented. The method gives a mAP equal to 55.7%. Our descriptor (mAP=55.9%) outperforms the approaches proposed in [8,21,23] and gives comparable results with [22,5]. A

Table 1. Comparison with the state-of-the-art: Hollywood2 dataset

Action	MBH[5]	[8]	[23]	[21]	[22]	**SIFT-ACC**	**Seg-SIFT-ACC**
AnswerPhone	-	-	40.20	-	26.30	28.1	29.9
DriveCar	-	-	75	-	86.50	87.2	88.2
Eat	-	-	51.50	-	59.20	66.8	67.1
FightPerson	-	-	77.10	-	76.20	71.9	75.4
GetOutCar	-	-	45.6	-	45.7	42.3	45.6
HandShake	-	-	28.90	-	49.7	29.7	32.9
HugPerson	-	-	49.4	-	45.40	41.8	45.8
Kiss	-	-	56.6	-	59.7	49.2	52.9
Run	-	-	47.50	-	72	75.5	77.2
SitDown	-	-	62	-	62.40	57.8	60.8
SitUp	-	-	26.80	-	27.50	33.4	35.4
StandUp	-	-	50.7	-	58.8	56.8	59.7
mAP	55.1	49.4	50.9	53.3	55.7	53.3	55.9

comparaison of our descriptor with other approaches in the state-of-the-art on the Olympic Sports dataset is shown in table 2. The HOG-HOF descriptor proposed by Laptev et al.in [4] gives a mAP equal to 62.0%. Heng and al. [5] obtain mAP equal to 71.6%. with the MBH descriptors. Carlos and al [7] propose to use a modelling temporal structure of decomposable motion segments for activity classification. This method gives 72.1%. The atomic actions approach [8] give a

mAP equal to 71.0%. Our descriptor (mAP=72.5%) achieve gives a significantly better performance than [4,5,8] and it is comparable to the modelling temporal structure of decomposable motion segments reported in [7].

Table 2. Comparison with the state-of-the-art: Olympic sports dataset

Action	MBH[5]	[4]	[7]	[8]	**SIFT-ACC**	**Seg-SIFT-ACC**
mAP	71.6	62.0	72.1	71.0	70.1	72.5

Impact of the Threshold Value. We vary the threshold value and we report in table 3 the recognition accuracy on the test set. We remark that, as found by cross validation, the best threshold value used for the Hollywood2 dataset is equal to 200 and greater than the best threshold value used for the olympic sports dataset that is equal to 120. This can be explained by the fact that the gradual transitions between segments occur more frequently in Olympic sports than in Hollywood2.

Table 3. MAP of Seg-ACC-SIFT descriptor with different thresholds on the olympic sports and Hollywood2 datasets

	T=100	T=120	T=150	T=200	T=250	T=350
Olympic Sports	71.1	**72.5**	72.01	71.8	71.4	71
Hollywood2	53	54.9	55.2	**55.9**	55.1	55

5 Conclusion

In this paper, we propose a novel motion descriptor Seg-SIFT-ACC for human action recognition. Our descriptor is focused both on the accordion representation of the video and its temporal segmentation. The Accordion representation is applied on every elementary motion segment. It transforms the temporal correlation between pixels into a spatial one. The motion information is extracted by computing SIFT descriptor around Harris3D interest points projected onto the accordion representation of each motion segment. The proposed descriptor shows better and comparable performances with the state-of-the-art methods.

References

1. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV (2003)
2. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse Spatio-Temporal features. In: VS-PETS (2005)
3. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)

4. Laptev, I., Marsza, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, pp. 3265–3271 (2008)
5. Heng, W., Alexander, K., Cordelia, S., Cheng-Lin, L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision (2013)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
7. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
8. Zhou, Q., Wang, G.: Atomic Action Features: A New Feature for Action Recognition. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 291–300. Springer, Heidelberg (2012)
9. Ahmed, O.B., Mejdoub, M., Amar, C.B.: SIFT Accordion: A space-time descriptor applied to human action recognition. In: ICMVIPPA 2011, Venice, Italy (2011)
10. Kwon Park, D., Seok Jeon, Y., Sun Won, C.: Efficient use of local edge histogram descriptor. ACM Multimedia Conference-MM, 51–54 (2000)
11. Sekma, M., Ben Abdelali, A., Mtibaa, A.: Application d'un descripteur MPEG7 de texture pour la segmentation temporelle de la vidéo. Sciences of Electronics of Information and Telecommunications (2012)
12. Mejdoub, M., Fonteles, L., BenAmar, C., Antonini, M.: Embedded lattices tree: An efficient indexing scheme for content based retrieval on image databases. Journal of Visual Communication and Image Representation 20(2), 145–156 (2009)
13. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of CVPR (2006)
14. Mejdoub, M., Ben Amar, C.: Classification improvement of local feature vectors over the KNN algorithm. Multimedia Tools Appl. 64(1), 197–218 (2013)
15. Dammak, M., Mejdoub, M., Zaied, M., Ben Amar, C.: Feature Vector Approximation based on Wavelet Network. ICAART (1), 394–399 (2012)
16. Mejdoub, M., Fonteles, L., Ben Amar, C., Antonini, M.: Fast indexing method for image retrieval using tree-structured lattices. CBMI, pp. 365–372 (2008)
17. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001)
18. Petersohn, C.: Temporal video segmentation. Berlin Institute of Technology, pp. 1–272 (2010) ISBN 978-3-938860-39-7
19. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
20. Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local Spatio-Temporal features for action recognition. In: BMVC (2010)
21. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical invariant Spatio-Temporal features for action recognition with independent subspace analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), pp. 3361–3368 (2011)
22. Ullah, M., Parizi, S., Laptev, I.: Improving bag-of-features action recognition with non-local cues. In: Proceedings of the British Machine Vision Conference (BMVC 2010), pp. 1–11 (2010)
23. Gilbert, A., Illingworth, J., Bowden, R.: Action Recognition using Mined Hierarchical Compound Features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 883–897 (2011)