SPATIO-TEMPORAL PYRAMIDAL ACCORDION REPRESENTATION FOR HUMAN ACTION RECOGNITION

Manel Sekma, Mahmoud Mejdoub, Chokri Ben Amar

REGIM-Lab: REsearch Groups on Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS), 3038 Sfax, Tunisia

ABSTRACT

We propose in this paper a spatio-temporal pyramid representation (STPR) of the video based Accordion image. The Accordion image allows the pixels having a high temporal correlation to be put in space adjacency. The STPR introduces spatial and temporal layout information to the local SIFT features computed on the Accordion image. It consists in applying firstly, a temporal pyramid decomposition on the video to divide it into a sequence of increasingly finer temporal blocks and secondly in performing a spatial pyramid representation on the Accordion images relative to the temporal blocks. The Multiple Kernel Learning approach is used to combine the multi-histograms coming from different Spatio-Temporal Pyramid levels. Experiments using the human action recognition datasets (Hollywood2 and Olympic sports) show the effectiveness of the proposed approach.

Index Terms— Human Action Recognition, Accordion Image, Space-Time Descriptor, Motion, Spatio-Temporal Pyramid Representation, Multi Kernel Learning.

1. INTRODUCTION

Recognizing human actions in realistic uncontrolled video is a challenging problem in computer vision. Yet, 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, 5], 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 [6, 7, 8]. 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. [9] proposed the Motion Boundary Histograms (MBH) is a descriptor for human detection. The MBH descriptor describes the relative motion between pixels by computing the gradient of the optical flow. In [6], MBH is used as motion descriptor for dense trajectories. Considering that the spatial pyramid method [10] performs well in the image classification, it was adapted [4, 11, 12, 6], with the local space-time features to spatio-temporal domain. Indeed, in the context of action recognition, spatio-temporal pyramid (STP) is used to embed structure information of the video action. Ullah et al. [11] propose decomposing the video into region classes and increase local features with corresponding region-class labels. The local features are extracted at multiple scale levels in space-time video pyramid. Chakraborty et al. [12] introduce a novel vocabulary building strategy by combining spatial pyramid and vocabulary compression techniques to reduce the dimensionality of the feature space. Recently, Wang et al. [6] propose to use Spatio-Temporal Pyramid with the dense trajectory application and combination of different descriptors. In our previous work [14, 13], the presented descriptor is based on the Accordion representation that transforms the video into a plan in order to put pixels with temporal adjacency in a spatial neighbourhood. Then, the Accordion representation is applied separately on each elementary motion segment. The limitation inherent in this method is the lack of spatial and temporal locations in the video description. To surmount this problem, we propose in this work to apply the spatio-temporal pyramidal representation (STPR) on the Accordion image. Indeed, video action is described by many histograms of visual words obtained from the different spatio-temporal grids of STPR. Afterwards, we use Multi Kernel Learning approach (MKL) [15] to combine the different histograms. To describe the motion information, Harris3D interest points are detected on the video frames and projected onto the Accordion image. After that, SIFT descriptor [16, 17] is computed around the projected Harris3D interest point.

This paper is organized as follows: in Section 2 an overall description of our proposed motion descriptor is given. The experimental study and results are given in section 3 and section 4. Finally, concluding remarks are presented.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. OUR PROPOSED MOTION DESCRIPTOR

The graphical description of our motion descriptor computation is illustrated by Figure 1. Firstly, we proceed by the detection of Harris3D interest points in the video frames. These points are then projected on the Accordion representation of the video. Then, we compute SIFT descriptors around the projected Harris 3D interest points. Afterwards, the I_{ACC} is divided into the Spatio-Temporal Pyramid levels and a histogram of visual words is built for each level. Video action is then described by many histograms of visual words obtained from the different spatio-temporal grids of STPR. Finally, the various histograms are combined in a Multi Kernel Learning (MKL) framework to classify the videos into action classes.

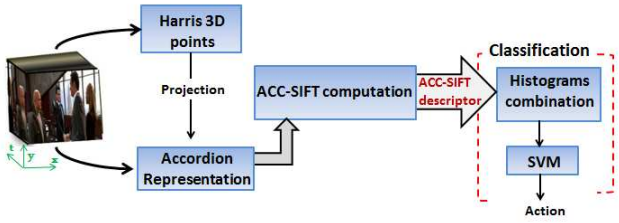

Fig. 1. Description of the proposed framework.

The Accordion representation is presented in section 2.1. In section 2.2, we present the computation steps of the proposed motion descriptor. The spatio-temporal pyramidal representation and the Multi Kernel Learning approach are described in section 2.3.

2.1. Accordion Representation

The accordion representation [13] 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

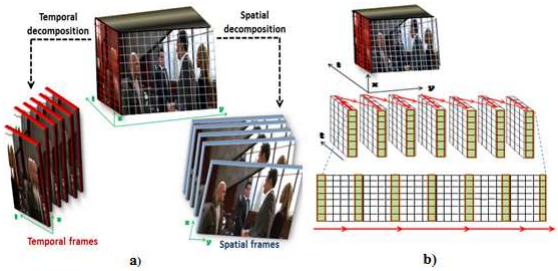

Fig. 2. The method of accordion representation: a) Video decomposition b) Video transformation into an Accordion image

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 Accordion image (I_{ACC}) throughout 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:

$$\begin{pmatrix} H\_acc & = & H \\ W\_acc & = & W * NbF \end{pmatrix} \quad (1)$$

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} .

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

2.2. Descriptor computation

The motion 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 video sequence into an I_{ACC}. After that, we project the detected Harris 3D interest points into the I_{ACC}. 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 patches. For that, every patch is sub-divided into 4×4 sub-regions (Figure 3). 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 resulting orientation histograms are transformed into $128d$ vector. Finally, the vector is normalized to unit length to achieve the invariance against illumination changes.

2.3. Histograms of Spatio-Temporal Pyramid and Multi Kernel Learning

In this section, we describe the general Spatio-Temporal Pyramidal framework and we present the MKL approach used in this work.

Spatio-Temporal Pyramidal Representation (STPR):

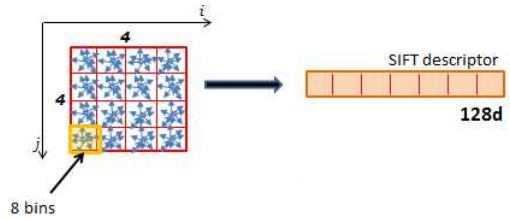

Fig. 3. SIFT computation.

The STPR adds spatial and temporal structural information to

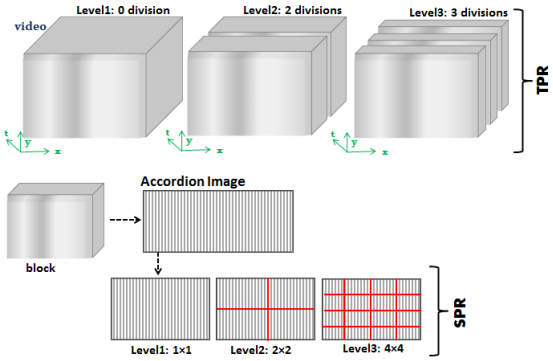

Fig. 4. Spatio-Temporal Pyramidal Representation (STPR).

the SIFT-ACC features. It consists in applying firstly a temporal pyramid representation (TPR) on the video to divide it into a sequence of increasingly finer temporal blocks and secondly in performing a spatial pyramid representation (SPR) on the Accordion images relative to the temporal blocks. For each temporal pyramid level, the video is divided into a sequence of temporal blocks (Level1: 0 division, Level2: 2 divisions and Level3: 3 divisions) and each temporal block is transformed into a I_{ACC}. Each I_{ACC} is then divided into a sequence of increasingly finer spatial grids on each spatial pyramid level (1×1; 2×2; 4×4), as shown in Figure 4. For each spatial level, histograms of SIFT-ACC features found inside the spatial grid cells of the I_{ACC} are concatenated to form the spatial histogram relative to this I_{ACC}. Finally, for each spatio-temporal pyramid level, the spatial histograms computed for each I_{ACC} are horizontally concatenated to form the spatio-temporal histogram.

Multi Kernel Learning (MKL):

The final representation of the video V is the set of histograms $h^V = \{h_l^V\}$, where each histogram h_l is computed for each level l. We use a MKL to combine the multi-histograms coming from different spatio-temporal pyramid levels. Our aim is to learn a SVM classifier [18] where, rather than using a pre-specified kernel, the kernel between the histograms of two videos V and U is learnt to be a linear combination of l kernel k_l:

$$K\left(h^V, h^U\right) = \sum_{l=1}^{9} \beta_l k_l\left(h_l^V, h_l^U\right), \; s.t\, \beta_l \geq 0, \; \sum_{l=1}^{9} \beta_l = 1$$

where β_l is the weight for each kernel optimized during the training process. In our implementation case, the kernel k_l corresponds to a chi-square RBF kernel of the form:

$$k_l\left(h_l^V, h_l^U\right) = exp^{-\gamma\chi\left(h_l^V, h_l^U\right)}$$

Where γ is fixed to the mean of the pairwise distances between histograms and χ is the chi-square distance.

3. EXPERIMENTAL STUDY

In our experiments, to implement the bag-of-features model, we use an identical pipeline as described in [19]. 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 [19] 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. The BOW [20, 21, 22] representation then assigns each feature to the closest vocabulary word features. The resulting histograms of visual word occurrences are used as video sequence representations. For classification, a learned multiple kernel SVM classifier (MKL) is used to combine the various histograms to give the actions classification.

3.1. Action recognition datasets

Fig. 5. Sample frames from action recognition sequences of Olympic Sports (top) and Hollywood2 (bottom) human action datasets.

Hollywood2 dataset: The Hollywood2 dataset [23] has been collected from 69 different Hollywood movies. There are 12 action classes: answering the phone, driving car, eating, fighting, getting out of car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up (Figure 5, bottom). 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 [23].

Olympic Sports dataset: The Olympic sport dataset [24] consists of athletes practicing different sports. There are 16 sports actions: high-jump, long-jump, triple-jump, pole-vault, basketball lay-up, bowling, tennis-serve, platform, discus, hammer, javelin, shot-put, springboard, snatch, clean-jerk and vault (Figure 5, top). 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. Results

In this section we present the experimental results using Hollywood2 and Olympic Sports datasets for our motion descriptor with and without Spatio-Temporal Pyramid Representation. Also we provide a comparison with the other approaches reported in the state-of-the art.

Hollywood2 results

In Table 1, we present the average precision of Hollywood2 action and we compare our SIFT-ACC(STPR+MKL) descriptor to SIFT-ACC obtained with the same method but without STPR+MKL method. We notice that SIFT-ACC(STPR+MKL) descriptor outperforms SIFT-ACC. Consequently, we can conclude that video description based on STPR+MKL improves the performance significantly. Qiang and Gang. [25] propose to use a spatio-temporal cuboid based on atomic actions, where atomic actions are basic units of human actions. The spatio-temporal pyramid is applied for each atomic action. This method gives 49.4%. In [26], 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%. Wang et al. [6] propose MBH+STP descriptor based on motion boundary histograms have achieved 57.6% using the Spatio-Temporal Pyramid. This descriptor outperforms to other descriptors, in particular in real videos that contain a significant amount of camera motion [6]. Despite dense features, we have obtained similar results with our motion descriptor SIFT-ACC(STPR+MKL).

An extension to the standard BoW approach is presented in [11] by locally applying BoW on regions that are spatially and temporally segmented. The method gives a mAP equal to 55.7%. In our previous work [14], based on temporal segmentation, the Accordion representation is applied separately on each elementary motion segment. We obtained a mAP equal to 55.9%. In this work, our motion descriptor achieves 57.5%. It outperforms the approaches proposed in [25, 26, 11, 14] and gives similar results with the MBH+STP descriptor [6].

Olympic Sports

A comparison of our descriptor with other approaches in the state-of-the-art on the Olympic Sports dataset is shown in table 2. We observe that SIFT-ACC(STPR+MKL) descriptor outperforms SIFT-ACC.

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

Action	[25]	[26]	MBH+ STP [6]	[11]	[14]	SIFT-ACC [14]	**Our**
AnswerPhone	-	40.20	-	26.30	29.9	28.1	31.6
DriveCar	-	75	-	86.5	88.2	87.2	89.4
Eat	-	51.5	-	59.2	67.1	66.8	69.4
FightPerson	-	77.1	-	78.2	75.4	71.9	76
GetOutCar	-	45.6	-	45.7	45.6	42.3	47.4
HandShake	-	28.9	-	49.7	32.9	29.7	34
HugPerson	-	49.4	-	45.4	45.8	41.8	46.5
Kiss	-	56.6	-	59.7	52.9	49.2	54.5
Run	-	47.5	-	72	77.2	75.5	78.8
SitDown	-	62	-	62.4	60.8	57.8	62.8
SitUp	-	26.8	-	27.5	35.4	33.4	37.3
StandUp	-	50.7	-	58.8	59.7	56.8	61.9
mAP	49.4	50.9	57.6	55.7	55.9	53.3	**57.5**

Laptev et al. [4] propose to generalize spatial pyramids to spatio-temporal domain and they suggest the HOG-HOF descriptor gives a mAP equal to 62%. Carlos et al [24] propose to use a modeling temporal structure of decomposable motion segments for activity classification. This method gives 72.1%. The atomic action approach [25] gives a mAP equal to 71.0%. The MBH+STP descriptor [6] achieves 74.9%. In our previous work [14] we obtained a mAP equal to 72.5%. As shown in table 2 our descriptor (mAP=75.6%) outperforms MBH+STP descriptor [6] as well as all methods proposed in [4, 24, 25, 14] and even in some cases by a significant margin.

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

Action	[4]	[24]	[25]	MBH+STP [6]	[14]	SIFT-ACC [14]	**Our**
mAP	62	72.1	71	74.9	72.5	70.1	**75.6**

4. CONCLUSION

In this work, we have presented a novel motion descriptor for human action recognition. Our descriptor relies on spatio-temporal pyramid representation (STPR) of the Accordion image. The Accordion image allows the pixels having a high temporal correlation to be put in space adjacency. The motion information is extracted by computing SIFT descriptor around Harris3D interest points projected onto the Accordion image. We apply the STPR on the Accordion image in order to introduce the spatial and temporal layout information into the local SIFT features computed. Experimental results on Hollywood2 and Olympic sports datasets have shown the efficiency of our descriptor.

5. REFERENCES

[1] Laptev, I. and Lindeberg, T., "Space-time interest points", In ICCV, 2003.

[2] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S., "Behavior recognition via sparse Spatio-Temporal features", In VS-PETS, 2005.

[3] Willems, G., Tuytelaars,T., and Van Gool, L.,"An effcient dense and scale-invariant Spatio-Temporal interest point detector", In ECCV, 2008.

[4] Laptev, I., Marsza, M., Schmid, C., and Rozenfeld,B., "Learning realistic human actions from movies", In CVPR, pp. 3265, 3266, 3268, 3271. 2008.

[5] Ben Aoun, N., Mejdoub, M., Ben Amar, C., "Graph-based approach for human action recognition using spatio-temporal features". Journal of Visual Communication and Image Representation, Elsevier, 2014.

[6] Heng, W., Alexander, K., Cordelia, S., and Cheng-Lin, L., "Dense trajectories and motion boundary descriptors for action recognition", In IJCV, 2013.

[7] Ben Aoun, N., Elaribi, M., and Ben Amar, C., "Multiresolution motion estimation and compensation for video coding", Proc. 10th IEEE International Conference on Signal Processing (ICSP'2010), Vol.2, pp. 1121-1124, 2010.

[8] Bouchrika, T. Zaied, M., Jemai, O. and Ben Amar, C., "Neural solutions to interact with computers by hand gesture recognition, International Journal "MULTIMEDIA TOOLS AND APPLICATIONS", MTAP, Springer Netherlands, DOI 10.1007/s11042-013-1557-y, 2013.

[9] Dalal, N., Triggs, B., and Cordelia, S., "Human detection using oriented histograms of flow and appearance", In ECCV, 2006.

[10] Lazebnik, S., Schmid, C., and Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", In Proc. of CVPR, 2006.

[11] Ullah, M., Parizi, S., Laptev, I., "Improving bag-of-features action recognition with non-local cue", In BMVC, pp. 1-11, 2010.

[12] Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzàlez, J. , "Selective Spatio-Temporal interest points", In CVTU 116 (3), pp. 396-410, 2012.

[13] Ben Ahmad, O., Mejdoub, M., Ben Amar, C., "SIFT Accordion : A space-time descriptor applied to human action recognition", In World Academy of Science, Engineering and Technology, 2011.

[14] Sekma, M., Mejdoub, M., Ben Amar, C., "Human Action Recognition Using Temporal Segmentation and Accordion Representation", In CAIP, Volume 8048, 2013, pp. 563-570, 2013.

[15] Gehler, P. and Nowozin S., "On feature combination methods for multiclass object classification". In Proc. of ICCV, 2009.

[16] Bouchrara, H., Chen, L, Ben Amar, C. and Chtourou M., "Face Recognition Under Varying Facial Expression Based on Perceived Facial Images and Local Feature Matching", International Conference on Information technology and e-Services, ICITeS 2012, DOI 10.1109/ICITeS.2012.6216663, pp. 1-6, 2012.

[17] Ben Aoun, N., Elghazel, Hacid, M. and Ben Amar, C., "Graph Aggregation Based Image Modeling and Indexing for Video Annotation", Lecture Notes in Computer Science, 6855, LNCS (PART 2), pp. 324-331., 2011

[18] Wali, A., Ben Aoun, N., Karray, H., Ben Amar, C., and Alimi, A. M., "A New System for Event Detection from Video Surveillance Sequences". Lecture Notes in Computer Science, Vol. 6475, LNCS (PART 2), pp. 110-120, 2010.

[19] Wang, H., Ullah, M., Klaser, A., Laptev, I., and Schmid, C., "Evaluation of local Spatio-Temporal features for action recognition", In BMVC, 2010.

[20] Mejdoub, M., Ben Amar, C. Classification improvement of local feature vectors over the KNN algorithm, Multimedia Tools Appl. 64(1), pp. 197-218, 2013.

[21] Dammak, M., Mejdoub, M., Zaied, M., Ben Amar, C., Feature Vector Approximation based on Wavelet Network.ICAART (1), pp. 394-399, 2012.

[22] Mejdoub, M., Fonteles, L., Ben Amar, C., Antonini, M., Fast indexing method for image retrieval using tree-structured lattices. CBMI, pp. 365-372, 2008.

[23] Marszalek, M., Laptev, I. and Schmid, C., "Actions in context", In CVPR, 2009.

[24] Niebles, J., Chen, C., and Fei-Fei, L., "Modeling temporal structure of decomposable motion segments for activity classification", In ECCV, 3265, 2010.

[25] Qiang, Z., and Gang, W., "Atomic Action Features: A New Feature for Action Recognition", in Computer Vision, Volume 7583, pp 291-300, 2012.

[26] Gilbert, A., Illingworth, J., Bowden, R., "Action Recognition using Mined Hierarchical Compound Features", IEEE Transactions on PAMI, 883-897, 2011.