International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:11, 2011

SIFT Accordion: A Space-Time Descriptor Applied to Human Action Recognition

Olfa.Ben Ahmed, Mahmoud. Mejdoub and Chokri. Ben Amar

***Abstract*—**Recognizing human action from videos is an active field of research in computer vision and pattern recognition. Human activity recognition has many potential applications such as video surveillance, human machine interaction, sport videos retrieval and robot navigation. Actually, local descriptors and bag of visuals words models achieve state-of-the-art performance for human action recognition. The main challenge in features description is how to represent efficiently the local motion information. Most of the previous works focus on the extension of 2D local descriptors on 3D ones to describe local information around every interest point. In this paper, we propose a new spatio-temporal descriptor based on a space-time description of moving points. Our description is focused on an Accordion representation of video which is well-suited to recognize human action from 2D local descriptors without the need to 3D extensions. We use the bag of words approach to represent videos. We quantify 2D local descriptor describing both temporal and spatial features with a good compromise between computational complexity and action recognition rates. We have reached impressive results on publicly available action data set

***Keywords*—**Accordion, Bag of Features, Human action, Motion, Moving point, Space-Time Descriptor, SIFT, Video.

I. Introduction

ACCORDING to several works, human action recognition methods in realistic uncontrolled video can be classified in three main categories: the model based approach, the holistic approach and the local features approach. The model based approach [1]-[2]-[3] tries to build 2D or 3D model from the human body parts such as hand, head and foot. This model is estimated by a motion capture operation. However, it is not beyond reproach as it is prone to variation of the degree of freedom of the moving person and the variability of human forms. Large varieties of parametric models were proposed such as elliptical cylinder model for human body representation and body model based on rectangular patches. Nevertheless, the latters are not suited to practical applications due to their complexity [1]-[3]. To overcome those problems, holistic (global) methods are put forth.

The global approaches [4]-[5]-[6]-[7] are based on the whole information of the image and do not require detection or labelling of human body parts. As a result, models extracted from the image are simpler, easier to build, more discriminatory, and much more robust than parametric models of human body. Traditional holistic methods, based on extracting silhouette information and shape masks, accumulate global motion information in a 2D representation to describe action. Among these methods, we can mention the method presented in [4]. In this method, the motion history images (MHI) and the motion energy images (MEI) are obtained by projecting 3D spatio-temporal volume into a 2D representation. The (MEI) describes the intensity of motion during the sequence and records its temporal evolution. (MHI) weights these regions according to the point in time when they occurred.

Without any preliminary knowledge about the location of the person in each video frame, the posture of a human body in video stream is recovered from a great number of characteristics extracted from frames. Human action recognition is carried using those estimations. Thus, Local features may be calculated in dense or sparse regions. Spatio-temporal points [5]-[6] were introduced to generalize the feature points and local descriptors that have been used in object recognition, image classification, action recognition and video indexing. Then, each point is described by a set of features called descriptor. After a classification stage, a vocabulary of keywords is built. Each video is described by a histogram called Bag of Features (BOF) [7]. In fact, a histogram represents the occurrences of keywords in a video. This model has recently been applied to action recognition [8, 9]. In this regard, several point detectors and descriptors have been proposed [5]-[6]-[8]-[9].

The spatio-temporal features are centered on an important issue which revolves around the way temporal information is represented. Thus, methods based on local space-time descriptors around local interest points are widely used to describe information on video. Each point is seen as a cuboid on a 3D volume to be described. The principle behind such approaches is that spatio-temporal volumes contain both temporal and spatial information. However, the 3D approaches have some disadvantages. On the one hand, it has been shown in [5] that when motion variations are not significant enough to be detected, motion detection is unavailing. On the other hand, the major drawbacks of space-time descriptors are high dimensionality and computational complexity [10]-[11]-[8]-[12].

In this work, we propose an efficient 2D spatio -temporal descriptor based on a local space-time description of moving points. Our description is based on an Accordion representation of video that allows to put pixels which have a very high temporal correlation in spatial neighbourhood [13]-[14]-[15]. The motion information is detected by collecting moving points from the 2D Accordion image. The spatial one is extracted by describing those points on spatial video frames. To describe points, we use the 2D SIFT descriptor [9]. We use the Bag of features approach to represent each video. We aim at including motion information into the feature description,

C. BEN AMAR is with the Department of Electrical and Computer Engineering, National Engineering School of Sfax, University of Sfax, Tunisia,e-mail: chokri.benamar@enis.rnu.tn.

O. Ben Ahmed and M. Mejdoub are with REsearch Group on Intelligent Machines, National Engineering School of Sfax, Road Sokra km 3, 3052, Sfax, Tunisia.

and at reducing the computational complexity without loss of relevant information. The remainder of this paper is organized as follows:

In section 2, some related works are presented. In section 3, the proposed approach is described. In section 4, the experimental results are given. Finally, concluding remarks and future works are presented in section 5.

II. Related Work

Methods based on feature descriptors around local interest points are widely used to describe information on video. Local space time features have recently become a popular video representation for actions recognition. Several methods for features detection and description have been proposed in the literature [5,6,8,9]. Most of the previous works in action recognition aim to combine motion and spatial information using the bag of word representation. Extending 2D local descriptors in 3D field was the most adapted solution.

The spatio-temporal local descriptors have recently been proved effective in human action recognition [16]. Feature extraction from video frames takes into account the temporal dimension. This is usually done by using optical flow vectors [17], spatial and temporal features such as cuboids [8].The selection of salient points is based on separable linear filters then cuboids are defined around these points [9]. In [18], local descriptors are based on responses to a 3D Gabor filter bank, followed by a Max-like operation. [19] calculates the descriptors by learning Ada-boost classifiers from low-level features, while [20] calculates the trajectories of moving points. The HOG/HOF [6] descriptor is based on concatenating histogram of optical flow and spatial histogram of oriented gradients HOG. [21] extends the SURF image descriptor to a video descriptor called extended SURF (ESURF)..

Building on the success of local descriptors based on histograms of oriented gradient (HOG) on image classification and objects categorization, existing space-time descriptors are usually based on the concept of HOG extended in 3D. Feature points are detected by the Spatio-Temporal Interest Points STIP algorithm [6], then 3D patches are described around selected points. More recently, in 3D local features point description, points are computed at random locations. This is different from 2D approach in that gradients are computed in polar coordinates. For orientation quantization gradient, histograms are built using meridians and parallels. SIFT 3D [11] can be seen as an example of those approaches. Scalavanner [11] extends 2D local SIFT descriptor on 3D spatio-temporal volume, for a given point, spatio-temporal gradients are computed, it leads to problems due to singularities at the poles and induces progressively smaller bins at poles.

To avoid this problem, Klaser [10] uses regular polyhedrons and spherical coordinates to quantize the orientation of spatio-temporal gradient. Therefore, this causes a high computational complexity and requires additional time. klaser et al [12] proposed an Harris 3D detector combined with HOG 3D descriptor. Otherwise, [22] assumes that motion information is located in the time axe so that it can be extracted by applying a 2D descriptor in a plan composed from one spatial dimension (x or y) and the time one. Space-time frames contain dynamic information, which makes them useful to describe action without the need to 3D extensions for the descriptors.

As discussed, most authors attempt to describe motion in addition to appearance information using 3D extension of 2D descriptors. However, all of these space time descriptors are extremely high dimensional, they often retain redundant information and high computational complexity.

III. Proposed Approach

The goal of the proposed approach is to represent efficiently actions by extracting local spatial and temporal features from videos.

A. Overview of the proposed approach

Our approach consists initially in extracting moving points and tracking them throughout the video frames. The displacement of those points is described by building trajectories. Each video is transformed into an Accordion image which makes the pixels having a great temporal correlation in the same spatial neighbourhood.

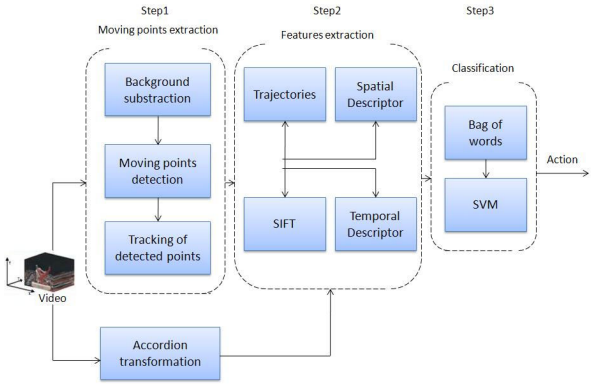


Fig. 1 Overview of our proposed framework

The moving points detected in the first step are projected into this image. Then, we extract 2D SIFT descriptors around every point. Moreover, local spatial information is described by extracting SIFT descriptor around moving points located on a selection of spatial frames. All of these descriptors are quantified. Feature fusion is achieved by concatenation of the resulting spatial and temporal descriptors. Later, we obtain one histogram per video sequence using the Bag of features models. A graphical overview of our approach is illustrated by Figure 1.

B. Moving points extraction

To detect moving points and to extract their displacements in video, we have been inspired by [23]. The idea is to combine two wide used techniques in computer vision: Background subtraction technique and Lucas Kanade optical flow algorithm.

Since the camera is static, salient points can be restricted inside a motion region by subtracting the background. According to a comparative study of methods of background subtraction [24], we adopt an adaptatif Gaussian Mixture Model (GMM) [25] to detect motion . Despite the simplicity of this model, the performances are largely satisfactory. In addition, this model has the advantage of requiring less computational time and memory resources.

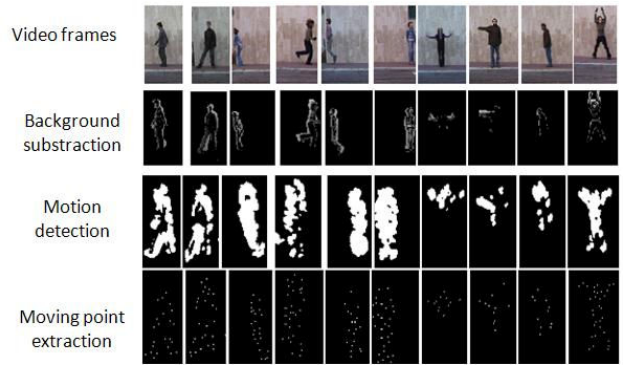


Fig. 2Moving points detection

Moving points detection is illustrated by Figure 2, each current video frame is subtracted from a reference frame and the result is thresholded to obtain the foreground or the motion regions. After motion detection, we extract salient points from the foreground by Shi and Tomasi corner detector [26]. Generally, in a video sequence, when a 3D point P moves to another point P , its corresponding pixel p moves in the image plan to another pixel p . We aim at finding its corresponding positions in every video frame. Corner features are tracked to obtain a sparse estimate of the optical flow field. In our work, Lucas Kanade pyramidal algorithm [27] is chosen for its known accuracy, simplicity and robustness. To compute points trajectories throughout the video, we calculate points displacements between every two consecutives frames i and i+1:

$$p_{i+1} = p_i + d_i = [x + dx_i;\ y_i + dy_i]^T \quad (1)$$

Where x_i and y_i are the coordinates of the point in the frame number \$i\$. The pixel displacement $d_i = [dx_i; dy_i]T$ is the image velocity at p_i also known as the optical flow. Hence, we obtain $T_p$the trajectories of p following the y and x axis:

$$Tp = \begin{bmatrix} X = (x_0; x_1 + dx_1; \ldots; x_i + dx_i; x_n + dx_n) \\ Y = (y_0; y_1 + dy_1; \ldots; y_i + dy_i; y_n + dy_n) \end{bmatrix} \quad (2)$$

n is the video frames number.

C. Accordion transformation

Several works proved that the video signal variation in 3D is much less in the temporal field compared to the space variation. Then, the pixels in 3D video volume are more correlated spatially than temporally. This was expressed as follows:

Every reference pixel I(x,y,t) is characterized by:

- I : Pixel intensity value
- x, y : Space coordinates of the pixel
- t : time

We generally have the following expression:

$$I(x,y,t) - I(x,y,t+1) < I(x,y,t) - I(x+1,y,t) \quad (3)$$

This assumption is the base of the representation in Accordion. Thus; it aims to put in space adjacency the pixels having a high temporal correlation. The video signal undergoes a space-time decomposition illustrated by the figure 3, the Accordion representation is built by carrying out the temporal decomposition of the video.

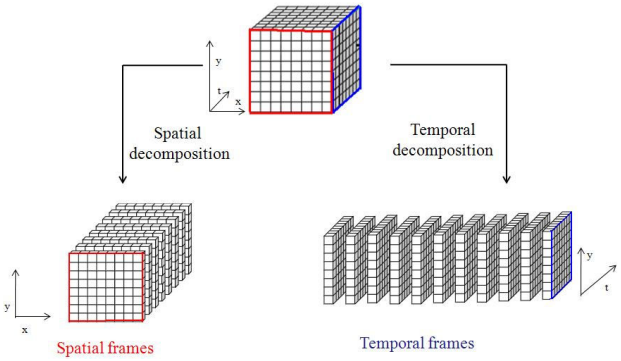


Fig. 3 Video decomposition

The resulting frames are called temporal frames which will be gathered into one 2D image that collects the video pixels having the same column rank referring to the 3D representation of the video.

As depicted in figure 4, the temporal frames are turned over horizontally (Mirror effect), the last stage consists of successively projecting this frame on a 2D plan called the IACC: Accordion. Indeed, this projection is obtained by traversing the temporal frames while reversing the direction of from one frame to another.

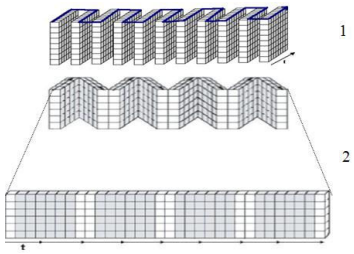


Fig. 4 Accordion representation

The purpose of changing the direction of course from a temporal frame to another is the use of spatial correlations in frames ends. However, it does not neglect the use of spatial correlations. It also minimizes the distances in the 2D representation between the pixels correlated in the source.

Thus, Accordion transformation tends to transform temporal correlation in the 3D original video source into a

high spatial correlation in the resulting 2D image (IACC). It aims to put in priority the exploitation of temporal correlation.

Let us consider the example of 3 video frames to which we would like to construct the IACC: The Accordion image size $(Xacc, Yacc)$ is given by:

$$\begin{bmatrix} X = X \\ Y = Y * NF \end{bmatrix} \quad (3)$$

With X and Y are the frame sizes; NF is the number of video frames.

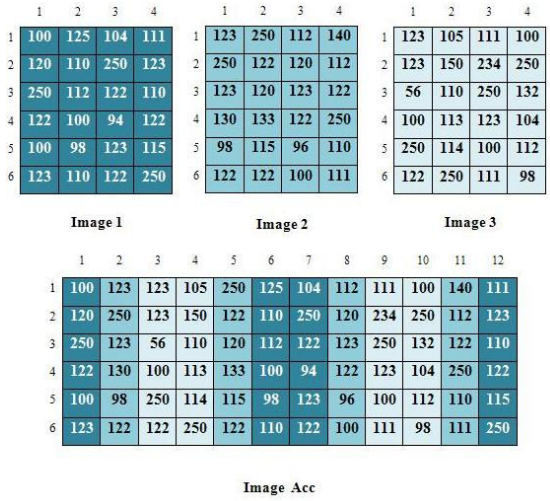


Fig. 5 Accordion transformation example

For instance, in Figure 5, columns 1 and 2 are adjacent in IACC and are also temporally adjacent to the video source. Columns 3 and 4 are adjacent in IACC and are spatially adjacent in the video source. Thus, the adjacent columns in the representation IACC are also adjacent to the video source either temporally or spatially.

D.Features extraction

After moving point extraction, we need to describe local information around them. We compute a visual descriptor on a normalized patch. We require descriptors to be highly distinctive for discrimination. Also, descriptors are claimed to be robust to lighting conditions as well as invariant to scale and rotation. Scale Invariant Feature Transform (SIFT) descriptor [9] remains the most popular and robust local invariant feature descriptor in computer vision. SIFT describes the gradient distribution in a local neighbourhood of a salient point. It delivers a descriptor of 128 dimensions. Here, we postulate two types of data to be extracted from video using SIFT: motion information that represents the change of a feature points along the time axis in a 3D space-time volume and spatial information that computes appearance around points. Hence, the resulting temporal and spatial descriptors are called, respectively, "Temporal SIFT Accordion" and "Spatial SIFT".

1) Temporal SIFT Accordion descriptor (T-SIFT-ACC)

The moving points trajectories are computed as explained in section B, each point position (x, y) on every video frame is projected into the Accordion image using the following projection function that calculates the new coordinates $(x\ Acc, yAcc)$ of each point in the Accordion image.

Projection: video3D → image2D
(x; y; i) → (xAcc; yAcc)

Once the feature point is extracted, it is described using the 2D SIFT descriptor.

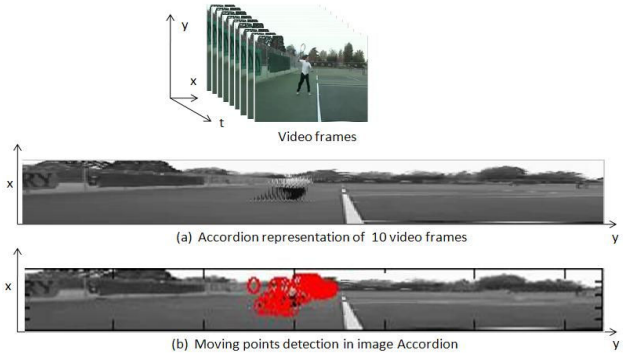


Fig. 6 Accordion representation example

As introduced in the previous section, the Accordion representation makes the pixels, temporally correlated, in spatial adjacency. Thus, we have to exploit this correlation by calculating gradient directions between those pixels. The obtained histograms are considered as local motion descriptors. Figure 6 illustrates an example of an Accordion transformation of ten video frames. This video shows a person playing tennis. Note that the background is stable and the camera is fixed. A look at the red points shows that they represent the moving points projected into the Accordion image. It follows, then, that the Accordion representation makes moving points in the same neighbourhood.

2) Spatial SIFT descriptor (S-SIFT)

The frames obtained following the video spatial decomposition are called spatial frames. After picking up M frames out of N video frames, SIFT descriptor is used to describe the spatial structure and the local orientation distribution of a patch surrounding the moving point detected on the M selected frames. The S-SIFT is the resulting appearance descriptor.

IV. Bag of Words Representation and Classification

The bag of visual words model (BOW) [1]has been very popular in recent years. It has been widely used thanks to its simplicity, perceptual and semantic property. Bag of visual features representation is inspired by bag of words in text categorization area in which it has been very successful. This technique was not only used for object recognition and image classification but also in video indexing and retrieval, event detection and human action recognition.

[1]Bag of Words (BoW), Bag of Features (BoF) are the same concepts

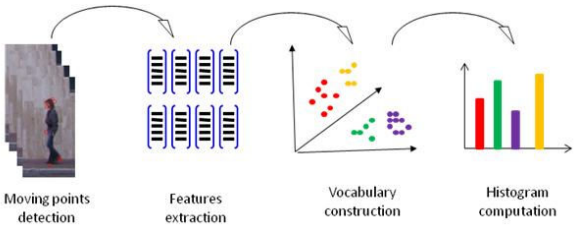


Fig. 7 Descriptor computation

The general steps for building a BOW representation is as follows: Firstly, after describing regions around moving points by SIFT descriptor, we cluster the descriptors by k-means to obtain a visual vocabulary and to quantize descriptors. Then, every descriptor is associated with one word from this codebook. Finally, we obtain one histogram per video sequence.

In this paper, we propose to combine appearance and motion information for human action classification in video. So, after building visual histogram for T-SIFT-ACC and S-SIFT descriptors, we combine them to obtain one histogram per video sequence describing space-time information. In order to evaluate our descriptors, we need to apply a classifier for actions. In our work, we use the supervised learning platform SVM (Support Vector Machines)

V. Experimental Results

A set of experiments were performed on a videos database. The Weizmann database is focused on human actions and provides a common benchmark to evaluate and compare human action recognition methods.

Fig. 8Weizmann dataset

The Weizmann dataset contains 93 video sequences with a homogeneous and static background. It consists in ten types of action classes: *bending downwards, running, walking, skipping, jumping-jack, jumping forward, jumping in place, galloping sideways, waving with two hands, and waving with one hand*. Each action class is performed once (sometimes twice) by 9 subjects. Action classification is usually done by extracting descriptors from a training subset and comparing them to descriptors extracted from the testing videos.

A. S-SIFT

SIFT descriptors are extracted from detected points in the selected video frames, in our case, we have chosen M=1 and N=3.

Building the visual vocabulary means quantifying extracted local descriptors. The vocabulary is generated by clustering SIFT features using the k-means algorithm. Once we have the vocabulary, we make signature, witch is a frequency histogram of the visual words. Each bin of the histogram corresponds to a visual word in the dictionary and the value of bin is the occurrence number of this word in the video.

We experiment with different codebook sizes varying from 50 to 900. The variation of performance as a function of codebook size is plotted in figure 9(a).

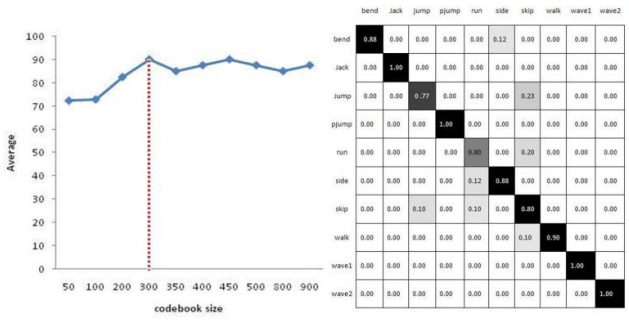


Fig 9 Results for the Weizmann dataset,(a) shows the average accuracy as a function of codebook size (b) shows results for the SVM classifier

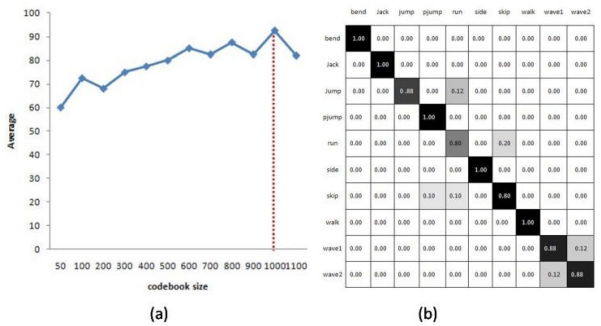
It is shown that in most cases the classifier became much more discriminative with increase in size of the codebook.

A recognition process is usually performed by comparing observations to examples. Hence, to evaluate the performance of the S-SIFT descriptor we gather the results in the matrix form showing the confusions among various activities. The best recognition rate for all tested values of vocabulary size is 90,3\% its correspondent confusion matrix for the ten classes is shown in figure 9(b), it presents the precision values for a vocabulary size equal to 300.

B. T-SIFT-ACC

The 3D video volume is transformed in 2D representation. The detected points are projected into the Accordion image using the projection function introduced in the previous section. Every point coordinates x and y in the original frame are transformed into new coordinates on this image. The same steps described on the section of S-SIFT descriptor are followed to represent every video by a histogram of visual words.

We test this descriptor with different codebook sizes varying from 50 to 1100. The variation of performance as a function of codebook size is plotted in figure 10 the best recognition average for all tested values of vocabulary size is 92,4% illustrated by the below confusion matrix 10 (b).



(a) (b)

Fig 10 Results for the Wiezman dataset (a) shows the average accuracy as a function of codebook size, (b), shows results for an SVM classifier

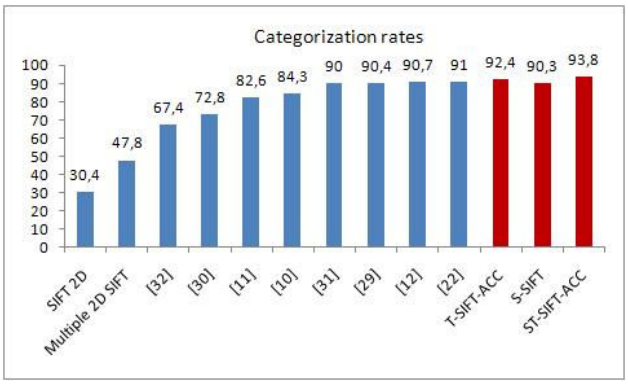


Fig. 11Comparison of our proposed descriptors with other works

1) Spatio-temporal SIFT Accordion (ST-SIFT-ACC)

The ST-SIFT-ACC is built by concatenating the tow preview descriptors. We obtain an average by 93,8 %. The T-SIFT-ACC addition improves accuracy of the spatial one. Table 1 shows the Categorization rates of the proposed descriptors

TABLE I
CATEGORIZATION RATES OF THE PROPOSED DESCRIPTORS

Descriptor	Accuracy
S-SIFT	90,3%
T-SIFT-ACC	92,4%
ST-SIFT-ACC	93,8%

These results provide a strong indication that 2D interest points descriptors can indeed be used to capture motion information in videos, when applied to the IACC. The description of moving points on the spatial frames improve recognition rate when combing with the temporal description illustrated by the T-SIFT-ACC descriptor.

VI. Comparison with the State-of-the-Art

In this section we aim to evaluate the performance of our work and compare it with state-of-art.

A. Categorization performances

The premise here is that a descriptor which better captures motion information is also better suited for actions recognition, we compare our descriptors to the state-of-the art descriptors evaluated in a bag of words framework and applied on the same videos data set. Categorization rates are shown in figure 11.

SIFT 2D describes a random selection of points from a single video frame. Using multiple 2D SIFT, descriptors are extracted around interest points in consecutive frames to describe the spatio-temporal region around each point. The motion information is missed in those representations. In [22, 21, 28, 11], the authors propose a variety of 3D descriptors, which combines the temporal and spatial information in a single representation.

Liu [29] combine multiple feature type. Therefore, we outperform his reported result by using a single descriptor (T-SIFT-ACC). In [22], the motion information is described using 2D SIFT descriptors and included in a bag of visual words representation without the need to extend information into 3D representation.

Typically 3D based descriptors have a large dimension. However, the dimensions of our descriptors (128 for The S-SIFT, 128 for the T-SIFT-ACC and 256 for the ST-SIFT-ACC) are much lower compared to 3D ones, for example, to the SIFT 3D [11] 4096 dimensional descriptor.

We can see that our proposed descriptors exceed the other representations. In Fact, we realize better average with small descriptors sizes, a reduced codebook dimensions and much less computational complexity.

B. Computational complexity

We test the time complexity of our proposed descriptors comparing it to The STIP [5] and the ST-SIFT [22]. As shown in table 2, point detection and description running time is much faster than other records [5,22]. The time space of histograms computation is smaller than STIP and closer to ST-SIFT. Also, our detection algorithm generates less features points than [22], in which points are collected from 2D frames. Detecting only moving points performs better than a random selection. This helps to reduce the computational complexity.

VII. Conclusion and Future Work

In this work, we proved that 2D moving points descriptors can indeed be used to capture both motion and spatial information. 2D spatio-temporal video representation, notably when using the Accordion representation, provides better recognition rates and lower time complexity than those

provided by 3D descriptors. Future work includes the validation of these results on other action databases and the application of our descriptors in other contexts such as video indexing and event detection. It seems that it will be very applicable to those tasks.

TABLE II
TIME COMPARISON

Descriptor	STIP	ST-SIFT	T-SIFT-ACC	S-SIFT
Detection+Description	582 s	1327 s	15,59 s	6,72s
Moving points detection	------	------	3,28 s	3,28s
Histogram computation	113 s	1395 s	104 s	38 s
Number of points	10886	504766	214	214

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB

REFERENCES

[1] D. Weinland, R.Ronfard and E. " A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition", *Computer Vision and Image Understanding 2010*
[2] K.Aggarwal and S.Park, "Human motion: Modeling and recognition of actions and interactions", *in 3DPVT'04 Washington, DC, USA: IEEE Computer Society, 2004*, pp. 640647
[3] T.B.Moeslund, A.Hilton, and V.Kruger, " A survey of advances in vision-based human motion capture and analysis", *CVIU* 2006, 90-126
[4] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates", *IEEE T-PAMI*, 257-267, 2001
[5] I. Laptev and T. Lindeberg, "Space-time interest points", *In ICCV,* 2003
[6] I. Laptev, M. Marsza lek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", *In CVPR*, 2008
[7] M. Mejdoub, L. Fonteles, C. BenAmar, and Marc Antonini. "Embedded lattices tree: An Efficient indexing scheme for content based retrieval on image databases", *Journal of Visual Communication and mage Representation, Elsevier*, 2009.
[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, *In VS-PETS*, 2005
[9] D. Lowe,"Distinctive image features from scale-invariant keypoints", *IJCV*, 91-110,2004
[10] A. Klaser, M. Marsza lek, and C. Schmid, "A spatio-temporal descriptor based on 3Dgradients", *In BMVC*, 2008
[11] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition", *In MULTIMEDIA*, 2007
[12] A. Klaser, M. Marsza lek, C. Schmid, and A. Zisserman,"Human Focused Action Localization in Video", *in International Workshop on Sign, Gesture, Activity* 2010
[13] T.Ouni, W.Ayedi and M.Abid, " New low complexity DCT based video compression method", *In Proceedings of the 16th International Conference on Telecommunications (ICT'09),* 202-207, Piscataway, NJ, USA, 2009, IEEE Press
[14] T.Ouni, W.Ayedi and Mohamed Abid, "New Non Predictive WaveletBased Video Coder: Performances Analysis", *In Proceedings of International Conference on Image Analysis and Recognition*. Volume 6111 of LNCS, pages 344-353, Berlin, Heidelberg, 2010. Springer-Verlag
[15] T.Ouni, W.Ayedi et M.Abid, "A Complete Non predictive VideoCompression Scheme Based on a 3D to 2D Geometric transform",*International Journal Signal and Imaging Systems Engineering (IJSISE), Inderscience Publisher*, 2011
[16] J.Wang, H. Lu, L.Duan and J.S. Jin, "Commercial Video Retrieval with Video-based Bag of Words", *Fifth International Conference on Intelligent Multimedia Computing and Networking* 2007, July.22, 2007. Salt Lake City, Utah, USA
[17] S.Ali, and M.Shah, "Human action recognition in videos using kinematic features and multiple instance learning", in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)28830, 2010
[18] H. Ning, Y. Hu, T. Huang, "Searching human behaviors using spatialtemporal words", *in Proceedings of IEEE ICIP* 07, 2007, pp. 337340
[19] A. Fathi and G. Mori. Action recognition by learning mid-level motion features, *In CVPR*, 2008
[20] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", *In ICCV*, 2009
[21] G. Willems, T. Tuytelaars, and L. Van Gool, "An effcient dense and scale-invariant spatio-temporal interest point detector", *In ECCV,* 2008
[22] A.P.B.Lopes, R.S. Oliveira, J.M. de Almeida, and A.de Albuquerque Araujo, Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition, *in SIBGRAPI 09. IEEE Computer Society*, 2009
[23] Y. Kawai, M. Takahashi, M. Fujii, M. Naemura, S. Sato, "NHK STRL at TRECVID 2010: Semantic Indexing and Surveillance Event Detection", *Proc. TRECVID Workshop, Gaithersburg, MD, USA*, November 2010
[24] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, C.Rosenberger,Review and evaluation of commonly-implemented background subtraction algorithms, *in Proc. of the International Conference on Pattern Recognition*, 2008
[25] C.Stauffer, W. Grimson, "Learning patterns of activity using real-time tracking", *in IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000, pp. 747757
[26] C.Tomasi and T.Kanade, Detection and tracking of Point Features, Carnegie Mellon University TeChnical Report *CMU-CS-91-132*, April 1991
[27] J.Y Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm", *Intel Corporation, Microprocessor Research Labs*,1999
[28] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos : Primal estimated sub-gradient solver for svm, *ICML*, pages 807814, 2007
[29] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features", *In CVPR*, 2008
[30] J. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classiffcation", *In CVPR*, 2007
[31] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words*", IJCV*, 299-318, 2008
[32] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words", *In BMVC*, 2006

Olfa Ben Ahmed IEEE Graduate Student member (M'09) Born in Tunisia in 17/02/1987, PhD student in REsearch Group on Intelligent Machines, National Engineering School of Sfax, Road Sokra km 3, 3052, Sfax, Tunisia. Tresorier of « Signal Processing Society » de ENIS « IEEE Student Branch » 2011.