A new approach of action recognition based on Motion Stable Shape (MSS) features

Imen Lassoued, Ezzedine Zagrouba, Youssef Chahir

A new approach of action recognition based on Motion Stable Shape (MSS) features

Imen Lassoued
Team of research SIIVA - RIADI laboratory
Institut supérieur d'informatique (ISI),Tunisia,
lassoued.imen@yahoo.fr

Ezzeddine Zagrouba
Team of research SIIVA - RIADI laboratory
Institut supérieur d'informatique (ISI),Tunisia,
ezzeddine.zagrouba@fms.rnu.tn

Youssef chahir
Team of research IMAGE- GREYC laboratory
université de Caen Basse Normondie, 14000 Caen, France
chahir@unicaen.fr

***Abstract*—Action recognition is actually considered as one of the most challenging areas in computer vision domain. In this paper, we propose a new approach based on utilization of motion boundaries to generate Motion Stable Shape (MSS) features to describe human actions in videos. In fact, we have considered actions as a set of human poses. Temporal evolution of each human pose is modeled by a set of new MSS feature's. Motion stable shapes of considered poses are defined by specific regions located at the borders of movements. Our modelisation is composed of different steps. First, a volume of optical flow frames highlighting the principal motions in poses is substracted. Then, motion boundaries are computed from the previous optical flow frames. Finally, maximally Stable Extremal Regions (MSER) are applied to motion boundaries frames in order to obtain MSS features. To predict classes of different human actions, the MSS features are combined with a standard bag-of-words representation. To prove the efficiency of our developed model, we have performed a set of experiments on four datasets: Weizmann, KTH, UFC and Hollywood. Obtained experimental results show that the proposed approach significantly outperforms state-of-the-art methods.**

I. Introduction

Human action recognition is an important component of video analysis with potential applications in video indexing, surveillance, gesture recognition and analysis of sport events. In fact, the state-of-the-art methods related to action recognition can be divided into two categories: global and local representation. Indeed, global representation methods [2,3] are generally restricted to some specific video actions because they rely on exact human localization or silhouette extraction. Therefore, many difficulties should be treated such as camera motion, dynamic and cluttered backgrounds, lighting changes, etc. Some other works focus on local features representation [11, 20, 23]. The local features methods allow to recognize a rich set of actions ranging from simple periodic motion (running, waving) to interactions (shaking hands, kissing), even in difficult realistic conditions [9, 15, 19, 31]. The principal contribution of this work is a development of a novel visual representation of human actions based on motion boundaries of the optical flow. The temporal evolution of different parts of human body is modeled with a set of particular regions of motion boundaries named motion stable shape (MSS). Given a human action sequence, the optical flow between any two consecutive frames is extracted. Their motion boundaries are derived and normalized to gray-scale images. Then, salient regions are detected using a Maximally Stable Extremal Regions (MSER) [8] feature detector. For each MSS, a local descriptor is computed. These descriptors are used as visual words that contain local optical flow and boundaries motions information. The remaining of this paper is structured as follows: In Section 2, an overview of existing methods for action classification and recognition are presented. In Section 3, we have introduced our new developed method based on MSS features for representing different actions in videos. In Section 4, different experiments have been performed on several actions datasets and results have been presented. Conclusion and ideas for further work are summarised in Section 5.

II. State of the art

In the literature, two main categories for action representation can be distinguished. The first one is global representation, witch aims to Utilize knowledge of the human location in the video and therefore learn a pattern of actions that capture the characteristics and overall body movements without any idea of the body parts. The second category is local representation where is based entirely on the descriptors of the local areas in a video, without any prior knowledge about the human positioning nor its members.

A. Global action representation

Global methods are based on the structure and dynamics of the whole body to represent human actions. Global models represent an action using descriptors of appearance and movements of either the whole body or the region of interest surrounding an actor. These representations generally depend on the silhouette extraction or structured grid that represents the area covered by the person performing an action. Global models are widely applicable because they do not rely on the identification and monitoring of various body parts. To characterize general motion and appearance of actions, many

information derived from silhouettes can be used. In this case, global dynamics of body are supposed to be discriminative enough to recognize human actions. In [1], silhouette information are used to represent actions. In this context, Motion Energy Image (MEI) and Motion History Image (MHI) are introduced to extract temporal information from video sequences. In the binary MEI, silhouettes are extracted from a single view and the difference between consecutive frames are aggregated. The MEI indicates therefore where motion occurs. At the same time, MHIs are used in combination with MEIs to weight regions that occurred more recently in time. Another way to model actions consists to use Space-time Shapes [2]. A space-time shape encodes both, spatial and dynamic information of a given human action. More precisely, the spatial information describe location and orientation of torso and limbs while dynamic information represent global motion of body and limbs. Note that human silhouettes are, in general, computed using background substraction techniques. Gorelick et al. [2] propose to compute local shape properties from the solution of Poisson equation. Those properties include local saliency, action dynamics, shape structure and orientation. A sliding temporal window is then used to extract space-time chunks of 10 frames length with an overlap of 5 frames. After that, those chunks are described by a high-dimensional feature vector and are matched against space-time shapes of test-sequences. The authors employ a nearest-neighbor classifier to vote for the associated class. In almost all cases, global action representation is based on a combination between the optical flow with the shape information. Moreover, Efros et al. [3] propose a method to recognize human actions in low-resolution videos. In the beginning, human-centered tracks are obtained from sports footage. As a second step, motion information are encoded by blurred optical flow. The horizontal and vertical optical flow as well as positive and negative components yielding four different motion channels are separated. To classify a human action, the test sequence is aligned to a labeled data set of actions. Their method has shown promising results on different sport video datasets such as: ballet, tennis, and soccer video sequences. The drawback of this approach is that they only consider full actions of completely visible people in simple scenarios (i.e., no occlusion and simple backgrounds). Furthermore, Jhuang et al. [4] propose a biologically-inspired system for action recognition. Their approach is based on extension of the static object recognition method proposed by Serre et al. [5] in the spatial-temporal domain. The original form features are replaced by motion-direction ones obtained from: gradient-based information, optical flow and space-time oriented filters. Lassoued et al. [6] represent actions by 3D silhouettes and describe it by different types of spatio-temporal moments. Multi-class SVM is then used to classify different actions. Schindler et al. [7] propose a method that combines both motion and appearance information to characterize human actions. For each frame, appearance information are extracted from the responses of Gabor filters. Motion information are extracted from optical flow filters. Finally, a linear multi-class SVM estimates the final class label for a given test sequence. A different way to model human actions was proposed by Ali et al. [8]. In their approach, they use concepts from the theory of chaotic systems to model and analyze non-linear dynamics of human actions. Klaser et al. [9] and Laptev et al. [10] proposed two promising global approaches for action localization in realistic video. Both methods propose an initial filtering to identify possible action localizations and to reduce the computational complexity. To avoid an exhaustive spatio-temporal search for localizing actions, Laptev et al. [10] use a human key-pose detector trained on keyframes. In a second step, actions are generated and represented as cuboids with different temporal extents and aligned to the detected keyframes. The cuboid region is represented by a set of appearance (histograms of oriented spatial gradients) and motion (histograms of optical flow) features which are learned in an AdaBoost classification scheme. These previous features can be organized in different spatial and temporal layouts within the cuboid search window. Klaser et al. [11] propose a generic pre-filtering approach to detect and track humans in video sequences. Action localization is done with a temporal sliding window classifier on the human tracks. For the desription of actions, the authors introduce a spatio-temporal extension of histograms of oriented gradients (HOG) [12], which that extracts appearance and motion information. Jiang et al. [32] focus on general information of the event, and use it to locate human activity or human action.

B. Local action representation

The local spatio-temporal characteristics describe form and movement for local video area. they offer a relatively independent representation of events with respect to spatio-temporal scales as well as the confusion of the background with the different movement in the scene. These characteristics are generally extracted directly from the video which avoids possible failures of other pretreatment methods such as segmentation or human movement detection. The literature propose many approaches for local spatio-temporal features extraction in videos. Laptev et al. [13] extend the Harris corner detector to 3D domain to determine the space-time interest points corresponding to local regions characterized by significant spatial and temporal changes. Dollar et al.[14] descriptors are based on normalized brightness, gradient and optical flow information. Ones et al. [33] use detector proposed by Dollar et al. [14] to detect interest points and use k-means to cluster them. The novelty is that it integrated the relevance feedback mechanism using SVM-ABRS for action classification. Bregonzio et al. [15] have extended this approach with 2D Gabor filters of different orientations. Hessian et al. [16] have made a spatio-temporal detector based on the determinant of Hessian matrix. Wong and Cipolla et al. [17] have added global information to the interest point detection by applying non-negative matrix factorization (NNMF) on the entire video sequence. The locations extracted by all these approaches are sparse and detect salient motion patterns. To describe spatio-temporal points, Schldt et al. [6] use higher order derivatives (local

jets). Scovanner et al. [18] extend the popular SIFT descriptor to the spatio-temporal domain. Willems et al. [16] generalize the image SURF (Speeded-Up Robust Features) descriptor to the video domain by computing weighted sums of uniformly sampled responses of spatio-temporal Haar wavelets. Yeffet et al. [9] propose Local Trinary Patterns for videos as extension of Local Binary Patterns (LBP). Klazer et al. [20] combine histograms of oriented gradients (HOG) and histograms of optical flow (HOF). Spatio-temporal interest points encode video information at a given location in space and tim. On the contrary, trajectories tracks spatial point over time and captures motion information. Messing et al. [21] extract features trajectories by tracking Harris3D interest points. Trajectories are represented by a sequence of log-polar quantized velocities and used it for action classification. Matikainen et al. [22] extract trajectories using a standard KLT tracker, cluster the trajectories and compute an affine transformation matrix for each cluster center. The elements of the matrix are then used to represent the trajectories. Sun et al. [23] compute trajectories by matching SIFT descriptors between two consecutive frames. They impose a unique match constraint among the descriptors and discarded matches that are too far. Actions are described with intra- and inter-trajectory statistics. Sun et al. [24] combine both KLT tracker and SIFT descriptor matching to extract long-duration trajectories. To assure a dense coverage with trajectories, random points are sampled for tracking within the region of existing trajectories. Spatio-temporal statistics of the trajectories are then used to discriminate different actions. Raptis and al. [25] track feature points in regions of interest. They compute tracket descriptors as concatenation of HOG and HOF descriptors along the trajectories. Jain and al. [39] decompos visual motion into dominant and residual motions, both in the extraction of the space-time trajectories and for the computation of descriptors, significantly improves action recognition algorithms. they design a new motion descriptor, the DCS descriptor, based on differential motion scalar quantities, divergence, curl and shear features. It captures additional information on the local motion patterns enhancing results. Jain and al apply the recent VLAD coding technique proposed in image retrieval provides a substantial improvement for action recognition. Xin and al [40] propose a learning framework using static, dynamic and sequential mixed features to solve different fundamental problems such as spatial domain variation and temporal domain polytrope. They utilise a cognitive-based data reduction method and a hybrid network up on networks architecture and extract human action representations for spatial and temporal interferences and adaptive to variations in both action speed and duration.

Methods of video actions representation are divided into two classes. Global action representation that are particularly effective when used to recognize the videos aligned spatially and temporally. These methods are not robust to occlusions (eg truncated actors), significant perspective changes, and changes in duration as they focus on the overall structure. The second class of methods is based on local primitives. The key advantage of local primitives based approaches is their flexibility regarding the type of video data. they can be applied to videos with the location of people where parts of their bodies are invisible. More recent work shows their successful application to the video data of the real world, like Hollywood movies and YouTube videos. The method proposed in this paper is based on local features and a representation bag to words for the actions of recognition

III. Proposed method

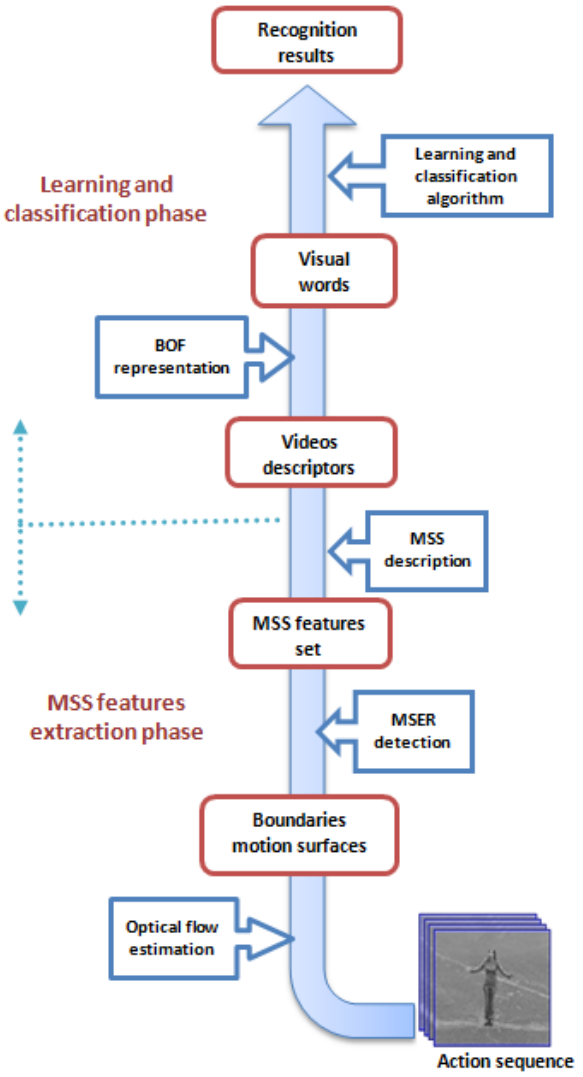

Figure 1. Flowchart of the proposed method

The main idea of the proposed method is to use motion boundaries surfaces to represent human action. More precisely, the temporal evolution of different parts of human body is modeled by a set of particular regions named Motion Stable Shape(MSS). These regions are localized on the boundaries surfaces. Figure 1 summarises the different steps of our developed method. Our method is composed of different steps. The first step consists to estimate the optical flow between every two consecutive frames in the human action sequence.

Thereafter, the motion boundaries are derived and identified from the optical flow fields. In the third step, MSER regions are detected from the motions boundaries surfaces to obtain MSS sets. The final step is dedicated to classification of different image sequences using MSSs descriptors combined to bag-of-words model.

A. Motion boundaries field extraction

Motion boundaries of different image are obtained by computing spatial derivatives of each optical flow by applying the gradient function. This processing consists in eliminating the stable motions of camera and conserving the motion boundaries and changes in the optical flow fields. In fact, motion boundary is more discriminant for action classification than optical flow fields. The boundary motion function 'Bound' for each frame I is defined as follows:

$$Bound(\Omega) = div[u, v] = \frac{d(u)}{dx} + \frac{d(v)}{dy} \qquad \text{(III.1)}$$

Where Ω is the optical flow variation for each frame I. u and v are respectively the horizontal and vertical components of optical flow at position (x,y). Indeed, surface boundaries are an excellent and simple source of visual motion information. As known, optical flow became discontinuous due to independent motion of different objects located in videos. As a result, we get motion boundaries between adjacent image regions having different velocities. These motion boundaries provide information about position and orientation of surface boundaries in the scene. Moreover, analysis of occlusion or disocclusion of pixels at motion boundaries can provide information about relative depth ordering of neighboring surfaces. Information about surface boundaries and depth ordering can be useful for different tasks like navigation, video compression and object recognition. The optical flow is calculated using the proposed TV-L1 variational method [34]. TV-L1 is a very efficient algorithm. In fact, variational methods are among the most popular and successful approaches for computing optical flow between two frames[34]. Among the reasons of popularity of the TV-L1 method are: a very appealing properties of the two terms in the energy formulation of the problem, the robust L1 norm in terms of data fidelity and the total variation (TV) regularization that smoothes the flow while preserving strong discontinuities. Specifically in [34], a very clean and efficient algorithm for calculating TV-L1 optical flows between gray scale images is provided. This algorithm can maintain discontinuities in the flow field and affords greater robustness against illumination changes, occlusion and noise.

B. MSS implementation

In this step, the MSS is detected from motion boundary fields using the maximally stable extremal regions (MSER) detector [8]. This former extracts a set of stable connected regions from a gray scale image. These regions are defined by an extremal property of intensity function in the region and on its outer boundary. The set of MSERs is closed under continuous geometric transformations. Particularly, MSER has

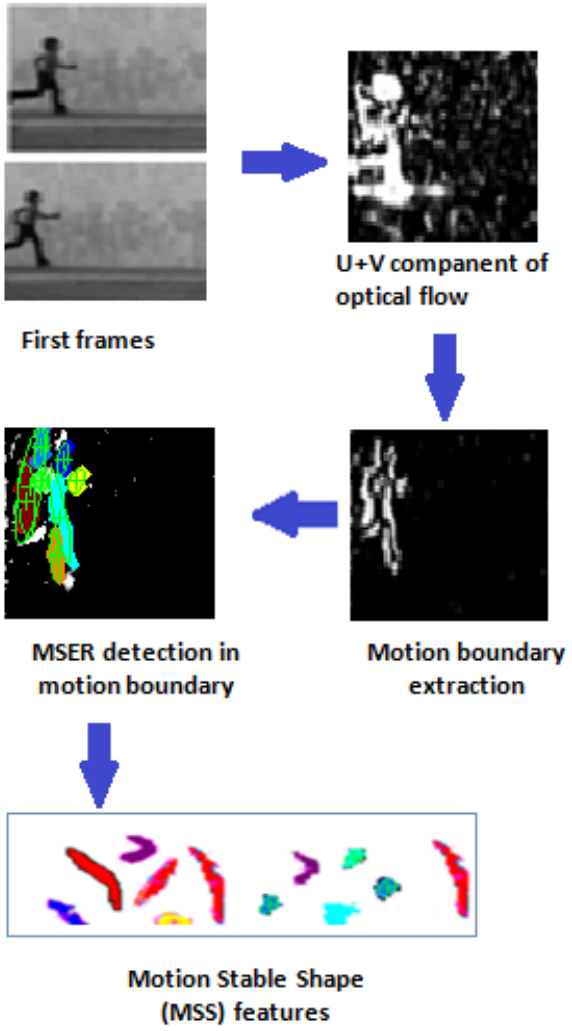

Figure 2. MSS feature of human motion

been widely used in image matching and object recognition, and present a better recognition performance [18] in several applications. MSER depends on the threshold of the image ξ , if we give them some threshold value the pixels below that threshold value are white and all those above or equal are black.it also depends on the Threshold of maximum area variation between extremal regions ρ. ρ is a threshold when an extremal region is maximally stable. Typical values of ρ range from 0.1 to 1.0.

Most of the detected MSS are localized on the boundary of silhouettes where motions are more frequent. Figure2 shows the different steps of MSS features implementation.

Given a set of MSSs, for each video β and $I_i(i = 1 \ldots N)$ is the set of frame from β, the representative ofβ ' Rep(β)' is defined as follows

$$Rep(\beta) = \bigcup_{I_i \in \beta}^{i=1..N} feat(I_i) \qquad \text{(III.2)}$$

Where

$$feat(I_i) = \bigcup_{j}^{j=1....s} MSS_{ij} = MSER(Bound(\Omega)) \quad \text{(III.3)}$$

where $'s'$ is the number of detected MSER regions in the image I_i.

Below, we present the MSER algorithm used to detect a set of MSS features using motion boundaries.

Algorithm 1: MSS implementation

Require: $Bound(\Omega)$
Ensure: $setofMSS$
ξ = find(Threshold intensity)
ER=Extremal Regions(ξ)
ρ = $find(Threshold\_max\_area\_variation)$
$setofMSS$ = Maximally Stable Extremal Regions(ρ,ER)

We describe each feature MSS_{ij} by the discrete polynomial Krawtchouk moments defined by yap et al. [37]. Krawtchouk moments have the interesting property of extracting the local characteristics of the image.they can recognize non-frontal images although it is trained up with frontal images. In fact, a good recognition rate is obtained using these moments when images are corrupted by noise. Based on Krawtchouk weighted polynomials defined by yap et al. [37], The Krawtchouk moments of the order $(n+m)$ for MSS_{ij} is defines as :

$$\varsigma_{nm}^{\tilde{v}} = \sum_{x=0}^{N_x}\sum_{y=0}^{N_y} K_{n,m}^{\tilde{}} MSS_{ij}(x,y) \quad \text{(III.4)}$$

Where $MSS_{ij}(x,y)$ is the function intensity and $\tilde{K_n}$ is the polynomial orthonormal Krawtchouk 2D proposed by Yap et al. [37] defined as:

$$K_{n,m}^{\tilde{}} = \tilde{K_n}(x;p_x,N_x)\tilde{K_m}(y;p_y,N_y) \quad \text{(III.5)}$$

C. Bag of MSS features

As mentioned above, each input video is described by a set of (MSSs) descriptors vectors. Based on these MSSs, the image sequences are classified using a bag-of-words model[26]. This model aims to represent videos by the occurrence of features or descriptors named visual words. Thus, videos are identified and compared using histograms of the visual word occurrences. Using this technique, various levels of abstraction are obtained after building a tree structure by repetitive clustering descriptors. We consider a class the one with the most similar visual word in each level.

The proposed approach aims to group descriptors of the detected MSS features into important visual words containing local optical flow and boundary motions. In particular, signatures are constructed by counting the number of visual word (MSS descriptors) in the video sequence. Figure 3 illustrates a diagram of video representation and classification with MSS features descriptors. Finally, a machine learning algorithm to train and classify visual word signatures is applied.

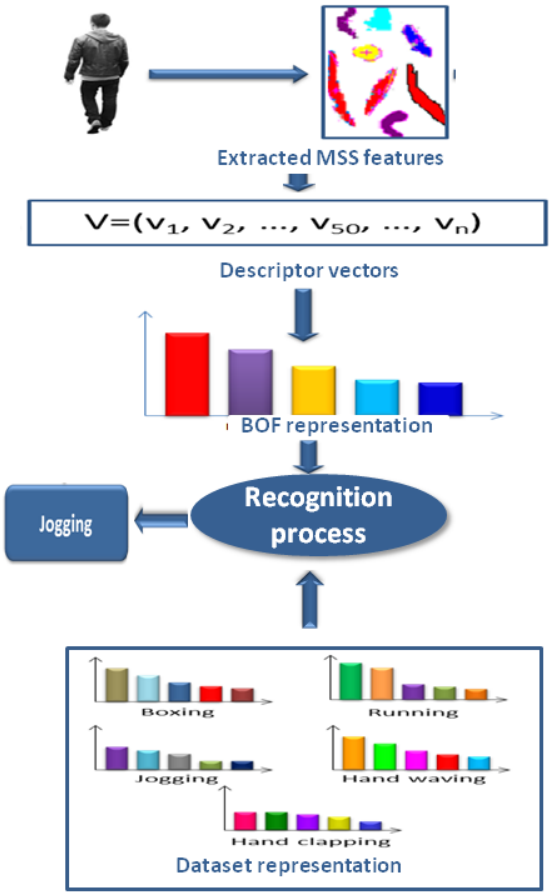

Figure 3. Illustration for video representation and classification with MSS features descriptors

IV. Experimental results

The experimentation process is composed of six steps. The first one consists in estimating the optical flow in image sequences. After that, the motion boundary for each frame is computed. In the following step, we apply the MSER detector to obtain Motion Stable Shape (MSS)witch depends on the complexity of surfaces. After computing, each videos have generated a number of MSS features between 500 and 700. Fourth, the MSS feature are described using Krawtchouk moments[6]. Fifth, we use hierarchical k-means to cluster descriptors into 278 visual words. We use these visual words signatures to classify acion sequences. Sixth, random ferns process is trained with 150 trees each one contains 1024 leaf nodes witch are used to make the last classification. Moments order is chosen after a series of experiments to study its influence on the classification result. Several tests of classification were performed on different datasets for several orders. Figure 4 shows the variation of average classification rate for different Krawtchouk moment orders. Curvesof figure 4 show clearly that the best classification rate in different datasets has been obtained for the order 60.

To asses objectively the proposed work, a set of experiment

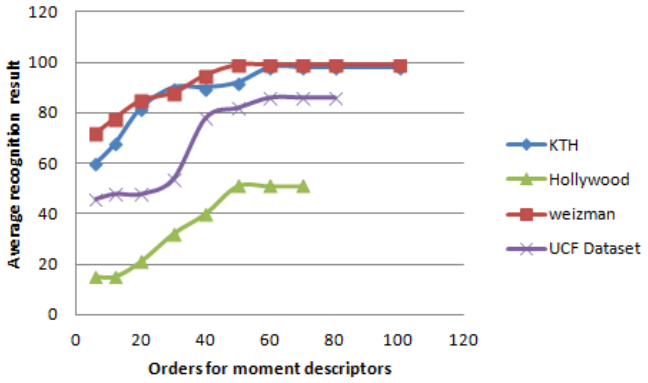

Figure 4. Average of classification rate vs different Krawtchouk moments orders

was performed in different datasets KTH, WEIZMAN, UCF sports and Hollywood (Figure5). KTH dataset [36] consists in six human action classes: walking, jogging, running, boxing, waving and clapping. The Weizmann actions dataset [3] is composed of nine different 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. For these two datasets, video backgrounds are homogeneous and static which simplifies computing. The choice of KTH and WEIZMAN datasets is explained by the fact that they are the most used in similar works. Furthermore, Hollywood [19] and UCF sports[38] datasets are caracterized by higher action complexity. We remind that Hollywood dataset videos has been collected from 69 Hollywood movies and representing 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). Training and test sequences come from different movies.

Table 1 shows the average classification rate for weizmann dataset compared to other work using different descriptors in bag-of-words classifier. We note that, the average classification rate of our proposed method is higher than other works witch prove its good effencicy. Table 1 illustrates an average recognition results. Indeed, the use of MSERs regions placed on the surface of boundary motion volumes improves the results by 3%. This demonstrates the advantages of analyzing stable regions of motion boundary.

Table I
ACTION RECOGNITION PERFORMANCE COMPARISON ON THE WEIZMANN DATASET FOR DIFFERENT COMBINATIONS OF FEATURES

Methods	Average classification rate (%)
Gray+ 3D SIFT	90.22
Flow+ 3D sift	93.11
Binary MSV+ 3D shape context	96.67
Proposed method	98.7

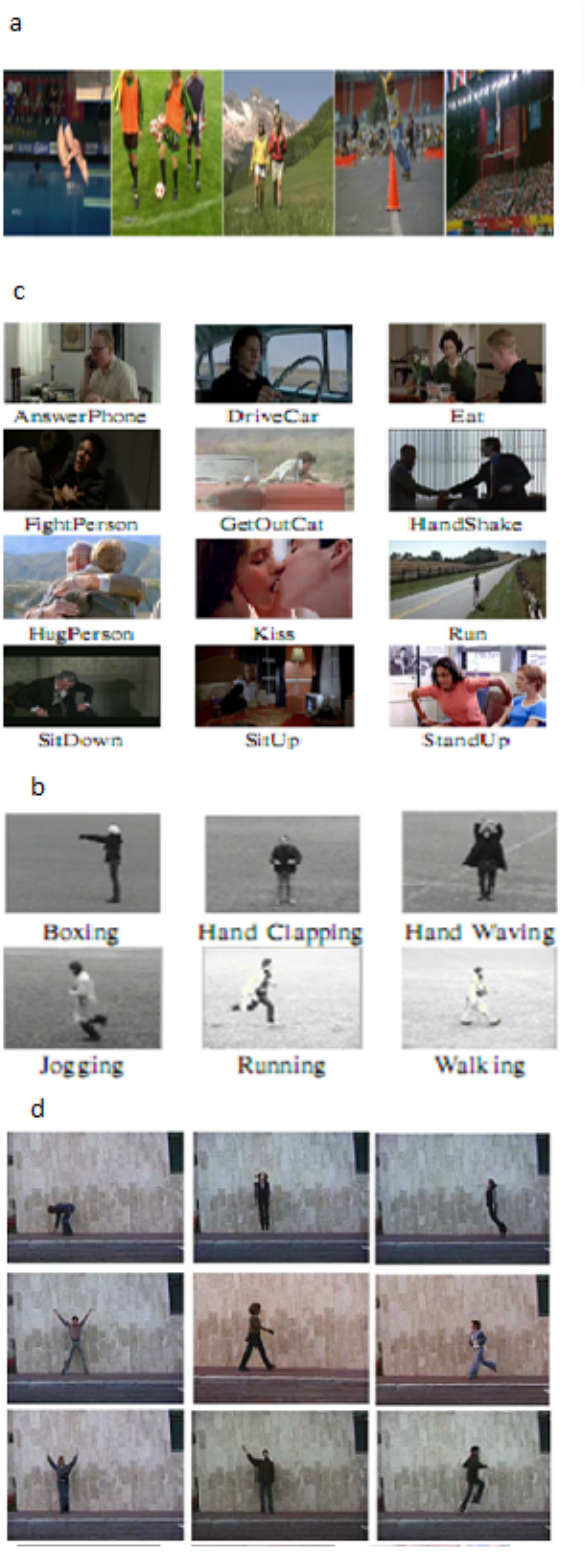

Figure 5. Examples of video actions from datasets: (a) UCF sports , (b)KTH, (c)hollywood and (d) Weizmann

In table I, four different feature representations and descriptor combinations taken from the literature are compared with our developped method. For the two first features : gray-scale image data and optical flow magnitude field, both are described by 3D SIFT. Concerning the binary volumes feature, they are described by a 3D shape context. Finally, our method uses MSS feature and described by krawtchouk moments. The best result is achieved by our MSS features with a classification rate of 98.7%. This highlighted the benefit of using flow volumes and stable motion boundaries as feature representation. This proves that boundary surfaces of motions contain enough information to properly classify actions in videos.

Table II shows experimental result of our developed method using bag-of-word classifier and comparison with state-of-art works using other different classifiers for different video actions datasets. This comparison proves that using MSS features with bag-of-words classifier improves results in the tested datasets. Inded, we have obtained a classification rate arround 98% for the two action datasets Weizmanna and KTH and this result is considered among the best compared to other methods sited in the table2. Nevertheless, classification rate of our method is becoming lower for UCF and Holluwood dataset(86% in UCF and 58% in Holluwood). But these last results remain higher than those obtained by other methods shown in table2. We can explain the decrease in classification rate UCF and Holluwood datasets by the fact that they contain more complex actions besides background mobility. we note that, our results are globally similar to those presented by XIN et al. [40] and Jain et al [39].

In order to study computation complexity, we have investigated theoretical and practical complexity of our method. We have computed MSS temporal feature and their Krawtchouk moments for each video.

The theorical complexity Cx can be written as

$$Cx = Nbr * (O(2 * nlog(n) + n^2 + nb_m(ns * log(log(ns))))) \quad \text{(IV.1)}$$

Where (nbr) is the number of frame considered in video,(n) is the number of pixels in each frame, (nb_m) is the number of MSS feature in each frame and ns is a dimensionality of each MSS.

Nine training video clips from the KTH dataset were used To analyze the computatio complexity of MSS extraction. The nine previous video clips have resolution of 160 $*$ 120 pixels and approximately 1000 frames. The computation time is devided as follows : 35% of stable boundary motion computation, 30% of the optical flow computing and the rest of the time is consumed equally between moments descriptors and motion boundary computation.

Table II
COMPARISON OF THE MSS DESCRIPTOR TO THE-STATE-OF-THE-ART, AS REPORTED IN THE CITED PUBLICATIONS IN DIFFERENT DATASETS

Dataset	***Methods***	***Average classification rate (%)***
Weizmann	Laptev et al. [13]	68.4
	Bregonzio et al. [15]	72.8
	Raptis et al. [25]	96.67
	Matikainen et al. [22]	82.6
	Klaeser et al. [20]	84.3
	Our method	**98.7**
KTH	Kovashka et al. [28]	94.5
	Youan et al. [29]	93.7
	Le at al. [27]	93.9
	Gilbert et al. [30]	94.5
	xin et al. [40]	95.2
	Our method	**98.3**
UCF sports	Wang et suter[35]	85.6
	Klaeser et al. [20]	86.7
	Kovashka et al. [28]	82.27
	Le et al.	86.5
	Our method	**86.8**
Hollywood	Wang et suter[35]	47.7
	Taylor et al. [31]	46.6
	Guilbert et al. [30]	50.9
	Li ey al[27]	53.3
	xin et al. [40]	63.1
	Jain et al. [39]	62.5
	proposed method	**53**

V. CONCLUSION

This paper introduces a new approach based on Motion Stable Shape (MSS) for video action recognition. To obtain MSS feature, each video is transformed to a set of motion boundaries volume. MSERs regions are then detected in computed motion boundaries. We have used MSS signatures with a bag-of-words model to classify different actions for image sequences in datasets KTH, Wezmainn, UFC and Hollywwod. The experimental results show that this method captures effeciently action information in videos and show good performances compared to state-of-the-art approaches for action classification. Future work will focus on the combination of our MSS features with other features based on appearance and spatial destribution to further improve recognition performance.

REFERENCES

[1] Bobick Aaron F. and Davis James W. "The recognition of human movement using temporal Templates". In IEEE Transactions on Pattern Analysis and Machine Intelligence, page 257-267,Atlanta,USA, 2001

[2] Lena G. Moshe B. Eli S. Michal I. and Ronen B. "Actions as space time shapes". In Proceedings of the Tenth IEEE International Conference on Computer Vision, pages 1395-1402, Washington, DC, USA, 2005.

[3] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. "Recognizing action at a distance". In IEEE International Conference on Computer Vision, pages 726-733, Nice, France, 2003.

[4] Hueihan J. Thomas S. Lior W. and Tomaso P. "A biologically inspired system for action recognition". In Proceedings of the Eleventh IEEE International Conference on Computer Vision, pages 1-8, Rio de Janeiro, Brazil, 2007.

[5] Thomas S. Lior W. Stanley B. Maximilian R. and Tomaso P." Robust object recognition with cortex-like mechanisms". In IEEE Transactions on Pattern Analysis and Machine Intelligence, page 411-426, 2007.

[6] Lassoued I. Zagrouba E. and Chahir Y. "Video Action Classification: A New Approach combining Spatio-teporal Krawtchouk Moments and Laplacian Eiginmaps ", In 7th IEEE International Conference on Signal Image Technology and Internet-Based Systems, page 291-301, Dijon, FRANCE, 2011.

[7] Konrad S. and Luc J." Action snippets: How many frames does human action recognition require". In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition", pages 1-8, Alaska, USA, 2008.

[8] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Proceeding of British Machine Vision Conference, pages 384396,

[9] Alexander K. Marcin M. and Cordelia S." A spatio-temporal descriptor based on 3d-gradients". In Proceedings of the British Machine Vision Conference, pages 995-1004, Leeds,UK British, 2008.

[10] Ivan L. and Patrick P. "Retrieving actions in movies". In Proceedings of the Eleventh IEEE International Conference on Computer Vision, pages 1-8, Rio de Janeiro, Brazil,2007.

[11] Alexander K. Marcin M. Cordelia Sc. and Andrew Z. "Human focused action localization in video". In International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV, pages 219-233,Crete, Greece, 2010.

[12] Navneet D. and Bill T ." Histograms of oriented gradients for human detection". In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition , pages 886-893, San Diego, CA,2005.

[13] Laptev I. Marszaek M. Schmid C. and Rozenfeld B." Learning realistic human actions from movies". In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages. 1-8,Anchorage, Alaska, USA, 2008.

[14] Dollr P, Rabaud V, Cottrell G, and Belongie S. "Behavior recognition via sparse spatio-temporal features". In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), Beijing, China, 2005.

[15] Bregonzio M. Gong S. Xiang T. "Recognising action as clouds of space-time interest points" In IEEE Conference on Computer Vision and Pattern Recognition,pages 1948-1955,Miami,USA,2009.

[16] Willems G. Tuytelaars T. Gool L. "An efficient dense and scaleinvariant spatio-temporal interest point detector". In European Conference on Computer Vision, pages 650-663,Heidelberg,2008

[17] Wong SF, Cipolla R. "Extracting spatiotemporal interest points using global information". IEEE International Conference on Computer Vision. pages 18, Rio de Janiero, Brazil,2007.

[18] Scovanner P. Ali S. Shah M. "A 3-dimensional SIFT descriptor and its application to action recognition" ACM Conference on Multimedia. page 23-29,Augsburg, Germany,2007.

[19] Yeffet L. Wolf L. "Local trinary patterns for human action recognition". IEEE International Conference on Computer Vision.pages 492-497,Kyoto, japan, 2009.

[20] Klaser A. Marszafek M. Laptev I. Schmid C. "Will person detection help bag-of-features action recognition". Rapport de recherche 00514828 NRIA. 2010.

[21] Messing R. Pal C. Kautz H. "Activity recognition using the velocity histories of tracked keypoints". IEEE International Conference on Computer Vision,pages 313-324,Kyoto,jpan,2009.

[22] Matikainen P, Hebert M, Sukthankar R "Trajectons: Action recognition through the motion analysis of tracked features". ICCV Workshops on Video-Oriented Object and Event Classification.Kyoto,japan,2009.

[23] Sun J. Wu X. Yan S. Cheong LF. Chua TS. Li J. "Hierarchical spatio-temporal context modeling for action recognition". IEEE Conference on Computer Vision and Pattern Recognition,pages 2004-2001,Miami, USA,2009

[24] Sun J. Mu Y.Yan S. Cheong LF "Activity recognition using dense long-duration trajectories".IEEE International Conference on Multimedia and Expo.pages 322-327, singapore,2010.

[25] M. Raptis, S. Soatto S, "Tracklet descriptors for action modeling and video analysis", In European Conference on Computer Vision,pages pp 577590, Crete,2010.

[26] Nistr D. Stewnius H. "Scalable Recognition with a Vocabulary Tree". In Proceedings of Conference on Computer Vision and Pattern Recognition, pages 2161-2168,New York, USA, 2006.

[27] Le Q. V. Zou W. Y. Yeung S. Y. Ng A. Y. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis". IEEE Conference on Computer Vision and Pattern Recognition,pages 3361-3368,Springs, USA, 2011.

[28] Kovashka A. Grauman K. "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition". IEEE Conference on Computer Vision and Pattern Recognition. pages 2046 - 2053, San Francisco, CA , 2010.

[29] Yuan J. Liu Z. Wu Y. "Discriminative video pattern search for efficient action detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1728-1743, 2011.

[30] Gilbert A. Illingworth J. and Bowden R. "Action recognition using mined hierarchical compound features," IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 883-897,Guildford, UK, 2011.

[31] Taylor G. W. Fergus R. LeCun Y. Bregler C. "Convolutional learning of spatio-temporal features," in European Conference on Computer Vision,pages 140-153, Crete, Greece 2010.

[32] Jones.S, Shao .L, Zhang .J, Liu .Y , "Relevance feedback for real-world human action retrieval". Pattern Recognition Letters, pages 446452. 2012.

[33] Jiang Y. Bhattacharya, S.; Chang, S.; Shah, M. High-level event recognition in unconstrained videos. In International Journal of Multimedia Information Retrieval, pp. 73 -101, 2013,

[34] Zach C., Pock T. Bischof H.," A duality based approach for realtime tvl1 optical flow". In procedure of Pattern Recognition, p. 214223, Heidelberg, Germany,pp 34-43, 2007

[35] L. Wang and D. Suter, "Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model",In IEEE Conference on Computer Vision and Pattern Recognition, pages 1-8, New York,USA,2007.

[36] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions: A local SVM approach." In ICPR, pp.332-36, 2004.

[37] P. T. Yap, R. Paramesran and S. H. Ong, Image Analysis by Krawtchouk Moments,. IEEE Transactions on Image Processing, Vol. 12, No. 11, pp. 13671376, November 2003

[38] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height
lter for action recognition. In CVPR, 2008.

[39] Jain, M., Jgou, H. and Bouthemy, P. Better exploiting motion for better action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2555-2562, 2013.

[40] Xin, M., Zhang, H., Wang, H., Sun, M. and Yuan, D. ARCH: Adaptive recurrent-convolutional hybrid networks for long-term action recognition. Neurocomputing, 178, pp.87-102, 2016.