




# Human actions recognition: an approach based on stable motion boundary fields

Imen Lassoued<sup>1</sup>  · Ezzeddine Zagrouba<sup>1</sup>

Received: 17 January 2017 / Revised: 23 November 2017 / Accepted: 29 November 2017 /

Published online: 19 December 2017

© Springer Science+Business Media, LLC, part of Springer Nature 2017

**Abstract** Automatic video action recognition have been a long-standing problem in computer vision. To obtain a scalable solution for actions recognition, it is important to have efficient visual representation of motions. In this paper, we propose a new visual representation for actions based in the body motion boundaries. The first step, a set of optical flow frames highlighting the principal motions in the poses is substracted. Then, the motion boundaries are computed from the previous optical flow frames. Maximum Stable Extremal Regions are then applied to motion boundaries maps in order to obtain Motion Stable Shape (MSS) features. Local descriptors were computed based on each detected MSS to capture motion patterns. To predict the classes of the different human actions, we have represented different descriptors with a bag-of-words (BOW) model and for classification, we use a non-linear support vector machine. We have performed a set of experiments on different datasets: Weizmann, KTH, UCF sport, UCF50 and Hollywood to prove the efficiency of our developed model. The achieved results improve the state-of-the-art on the KTH and Weizmann datasets and are comparable to state-of-the-art for UCF sport and UCF50 datasets.

**Keywords** Video sequences · Human action recognition · Motion boundary · Stable regions

## 1 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.

---

✉ Imen Lassoued  
lassoued.imen@yahoo.fr

Ezzeddine Zagrouba  
e.zagrouba@gmail.com

<sup>1</sup> Limtic Laboratory, Institut Supérieur d'Informatique (ISI), Ariana, Tunisia

In fact, the state-of-the-art methods related to action recognition can be divided into two categories: global and local representation. Indeed, the global representation methods [6, 15, 20] 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 [2, 14, 33, 39]. 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 [1, 4, 34, 42]. 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 each 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) [21] feature detector. Local descriptors are subsequently extracted from each detected MSS to characterize the local motion patterns. 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, we have presented an overview of existing methods for action classification and recognition. In Section 3, we have introduced our new developed method based on MSS features for modeling 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.

## 2 State of the art

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

### 2.1 Global action representation

The global methods are based on the structure and dynamics of the whole body to represent human actions. In fact, they 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. In order to characterize general motion and appearance of actions, many information derived from silhouettes can be used. In this case, the global dynamics of body are supposed to be discriminative enough to recognize human actions. In [3], 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 [20]. 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 represents global motion of body and limbs. Note that human silhouettes are, in general, computed using background subtraction techniques. Moreover, Efros et al. [6] 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, the 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. [35] 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. [36] 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. [18] 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. A different way to model human actions was proposed by Ali et al. [21]. 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. [1] and Laptev et al. [9] 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. [9] 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. [2] propose a generic pre-filtering approach to detect and track humans in video sequences. In this case, action localization is done with a temporal sliding window classifier on the human tracks. For the description of actions, the authors introduce a spatio-temporal extension of histograms of oriented gradients (HOG) [24], which extracts appearance and motion information. Jiang et al. [13] focus on general information of the event, and use it to locate human activity or human action.

Global models representations depend on the silhouette extraction or structured grid that represents the area covered by the person performing an action. This 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.

## 2.2 Local action representation

The local spatio-temporal characteristics describe shape 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. For instance, Laptev et al. [17] 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. For Dollar et al. [5] descriptors are based on normalized brightness, gradient and optical flow information. Ones et al. [12] use detector proposed by Dollar et al. [5] to detect interest points and use k-means to cluster them. The novelty is that they integrated the relevance feedback mechanism using SVM-ABRS for action classification. Bregonzio et al. [4] have extended this approach with 2D Gabor filters of different orientations. Hessian et al. [38] have made a spatio-temporal detector based on the determinant of Hessian matrix. Scovanner et al. [31] extend the popular SIFT descriptor to the spatio-temporal domain. Willems et al. [38] 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. [1] propose Local Trinary Patterns for videos as extension of Local Binary Patterns (LBP). Klazer et al. [14] 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 time. On the contrary, trajectories tracks spatial point over time and captures motion information. Messing et al. [23] 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. [33] 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. [32] proposed a way to learn the spatiotemporal relationship, where the authors factored 3D convolution in a space of spatial temporal convolution 2D and 1D. More precisely, their temporal convolution is a 2D convolution over time and the function channels. Raptis and al. [26] track feature points in regions of interest. They compute tracklet descriptors as concatenation of HOG and HOF descriptors along the trajectories. Jain et al. [10] decompose the visual motion into residual dominant movements. It uses this decomposition both in the extraction of space-time trajectories and for the descriptors computing. This has significantly improved the action recognition algorithms. They design a new motion descriptor, based on differential motion scalar quantities, divergence, curl and shear features. This descriptor captures additional information on the local motion patterns enhancing results. Jain and al apply the recent VLAD coding technique proposed in image retrieval. It 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. Geng et al. [7] and ji et al. [11] propose a deep model convolutional neural network (CNN) for human action recognition that can act directly on the raw inputs. In [27], Reddy Kishore et al. perform a combination of early and late fusion on multiple features to handle the very large number of categories. they use also the scene context as a feature to perform action recognition on very large datasets. Sadanand et al.in [29] present Action Bank, a new high-level representation of video. Action bank is

comprised of many individual action detectors sampled broadly in semantic space as well as viewpoint space.

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 bag-of-word representation for the actions recognition.

### 3 Proposed method

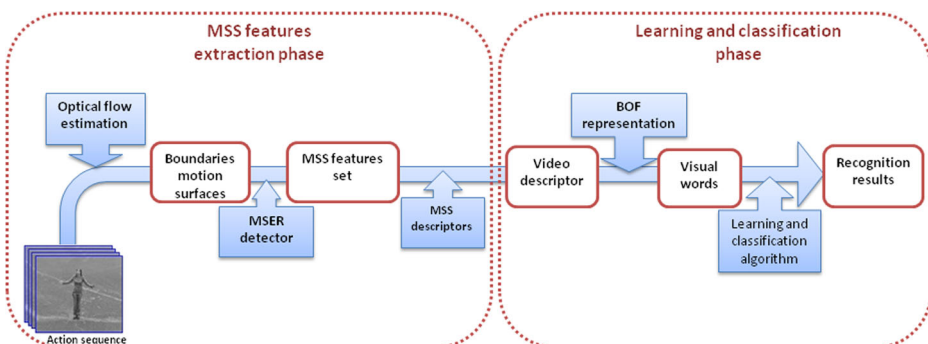
The main idea of the proposed method is to use motion boundaries surfaces to represent human actions. More precisely, the temporal evolution of the different parts of the human body is modeled by a set of particular regions named Motion Stable Shape(MSS). These regions are localized on the boundaries surfaces. Figure 1 summarises the different steps of our developed method. Indeed, our method is composed of several steps. The first step consists in computing 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 second step, MSER regions are detected from the motions boundaries surfaces to obtain MSS sets. The final step is dedicated to the classification of the different image sequences using MSS descriptors combined to bag-of-words model.

#### 3.1 Motion boundaries field extraction

Motion boundaries of different images are obtained by computing the 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 boundaries is more discriminant for action classification than optical flow fields. The boundary motion function 'Bound' for each frame I is defined as follows:

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

Where  $\Omega$  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



**Fig. 1** Flowchart of the proposed method

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. This information 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 [44]. 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 [44]. 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 [44], 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.

### 3.2 MSS implementation

In this step, the MSS is detected from motion boundary fields using the maximally stable extremal regions (MSER) detector [21].

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. Mser method is based principally [21] on detecting covariant regions from images. It allows assembling some gray levels in the image using a large range of thresholds. Then, obtained pixels are labeled into two categories, those below the threshold  $t$  will be considered white and the others become black. After that, images will be transformed to a sequence of black and white spots with different thresholds. These connected components will be representative of all extremal regions. Finally, ellipses forms will be fitted to the regions and combined to MSER to create devoted descriptors. Particularly, MSER has been widely used in image matching and object recognition, and present a better recognition performance [31] in several applications. Most of the detected MSS are localized on the boundary of silhouettes where motions are more frequent. Figure 2 shows the different steps of MSS features implementation.

Each video  $\beta$  is represented by ' $Rep(\beta)$ ' defined as follows:

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

Where

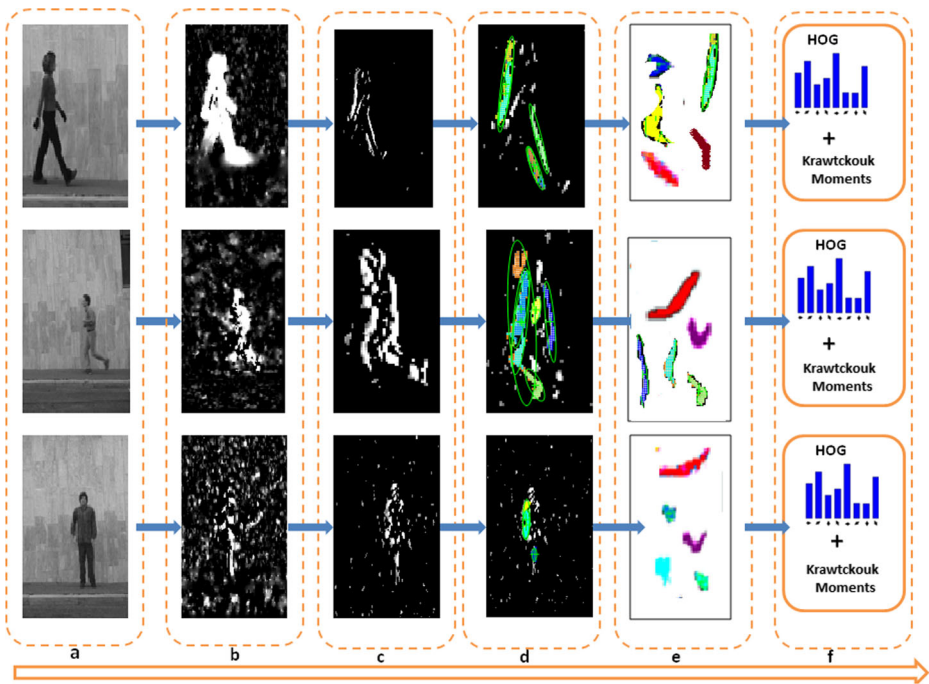
$$I_i (i = 1 \dots N) \text{ is the set of frame from } \beta$$

and

$$feat(I_i) = \bigcup_j^{j=1..s} MSS_j = MSER(Bound(\Omega)) \tag{3}$$

Where ' $s$ ' is the number of detected MSER regions from the image  $I_i$ .

In each detected region  $MSS_j$ , we calculate a histogram of the orientations of the optical flow (HOF) vector and a discrete polynomial krawtchouk moments defined by yap et al. [41]. The orientations of optical flow are computed from ' $u'$ ' and ' $v'$ '. All the orientations



**Fig. 2** Motion Stable Shape (MSS) extraction process. **a** the first frame of an image pair, **b** the u and v component of the estimated optical flow, **c** the boundary motion field, **d** MSER detection on the boundaries motion in **c**, **e** MSS forms, **f** calculating krawtchouk moments and a histogram of optical flow orientations with 8 bins from each MSS form

are then quantized and aggregated into discrete bins with their magnitudes as weights. Each histogram has 8 bins and normalized to have unit L1-norm.

The choice of histograms of flow orientations as one of our local descriptors is explained by the fact that the speed of actions varies widely, particularly among different humans. A good actions recognition algorithm therefore should be relatively insensitive to the speed with which actions are performed. We focus in the orientations of body movement as one of the significant measures for recognition. We chose, also, Krawtchouk moments as a second shape descriptor because they have the interesting property of extracting the local characteristics of the regions. 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. [41], the krawtchouk moments of the order  $(n, m)$  for  $MSS_{ij}$  is defined as:

$$\tilde{Q}_{nm} = \sum_{x=0}^{N_x} \sum_{y=0}^{N_y} \tilde{K}_n \tilde{K}_m MSS_{ij}(x, y) \tag{4}$$

Where  $MSS_{ij}(x, y)$  is the function intensity and  $\tilde{K}_n$  is the polynomial orthonormal krawtchouk 1D proposed by Yap et al. [41].

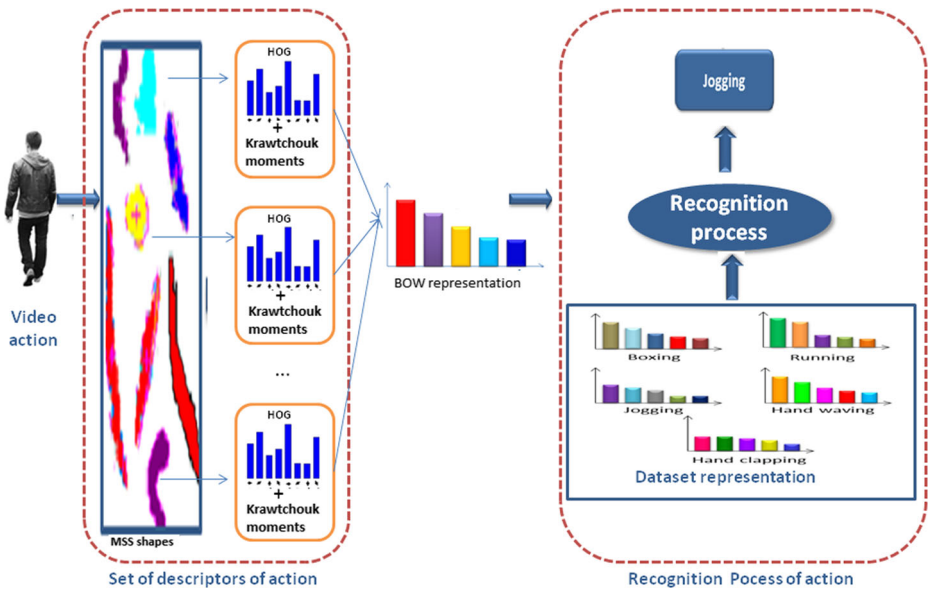
We use a vector  $K = [\tilde{Q}_{11}, \tilde{Q}_{22}, \dots, \tilde{Q}_{ij}]$  composed by Krawtchouk moments computed in different order as a second descriptor of MSS.

### 3.3 Bag of MSS features

As mentioned above, each input video is presented by a set of extracted (MSS). Each MSS is described by combined a histogram of the orientations of the optical flow and Krawtchouk moments vectors. Based on these descriptors, videos are presented using a bag-of-words model [25]. 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, we use a non-linear support vector machine with a  $\chi^2$ kernel [17] to train and classify visual word signatures is applied.

### 4 Experimental results

The experimentation process is composed of six steps. The first one consists in estimating the optical flow in the 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) which depends on the complexity of surfaces. once the computing step is down, each video have generates a number of MSS features between 500 and 700. Next, the MSS features are described using the krawtchouk moments [18] and a histogram of the



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



orientations of the optical flow. During the fifth step, we use hierarchical k-means to cluster descriptors into 278 visual words. In fact, we use these visual words to classify action sequences. For classification, we use a non-linear support vector machine with a  $\chi^2$ kernel [17], we apply the “one – against – rest” approach and select the class with the highest score.

**Fig. 4** Examples of video actions from datasets: **a** UCF sports, **b** KTH, **c** Hollywood, **d** Weizmann and **e** UCF50



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. Curves of Fig. 4 show clearly that the best classification rate in different datasets has been obtained for the order 60.

To assess objectively the proposed work, a set of experiment was performed in different datasets: KTH, Weizman, Hollywood, UCF50 and UCF sports (Fig. 4). KTH dataset [30] consists in six human action classes: walking, jogging, running, boxing, waving and clapping. The Weizman actions dataset [6] 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 the processing. The choice of KTH and Weizman datasets is explained by the fact that they are the most used in similar works. Furthermore, Hollywood [42], UCF50 [27] and UCF sports [28] datasets are characterized 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. In our experiments, we used the clean training dataset (the authors also provide an automatic, noisy dataset). UCF sport dataset contains ten different types of human actions: swinging, diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging at the high bar, golf swinging and walking. The dataset consists of 150 video samples. UCF50 [27] Dataset includes unconstrained web videos of 50 action classes with more than 100 videos for each class.

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

In Table 1, four different feature representations and descriptor combinations taken from the literature are compared with our developed 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 with HOF and krawtchouk moments descriptors. Moreover, 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. In KTH dataset, our results improve those obtained in

**Table 1** 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
MSS+Kawtchouk moments	97.2
MSS+HOF+Krawthouk moments	98.2

[7] and [11]. In fact, Geng et al. [7] and Ji et al. [11] use a convolutional Neural Networks (CNN) primitives . they are based on complicated hierarchical features via convolutional operations with sub-sampling of input images and they are very time consuming since the convolution use kernels/weights for training. Table 2 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. Indeed, we have obtained a classification rate around 97.8% for the two action datasets Weizmann and KTH and this result is considered among the best compared to other methods sited in the Table 2.

Nevertheless, classification rate of our method is becoming lower for UCF and Hollywood dataset(86% in UCF sport, 75.9 % in UCF50 and 58% in Hollywood). However, these

**Table 2** 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. [17]	68
	Bregonzio et al. [4]	72.8
	Raptis et al. [26]	96.67
	Matikainen et al. [22]	82.6
	Klaeser et al. [14]	84.3
	Our method	<b>98.2</b>
KTH	Kovashka et al. [16]	94.5
	Youan et al. [43]	93.7
	Le et al. [19]	93.9
	Gilbert et al. [8]	94.5
	Xin et al. [40]	95.2
	Geng et al. [7]	92.49
	Ji et al. [11]	90.2
	Our method	<b>97.3</b>
UCF sports	Wang and Suter [37]	85.6
	Klaeser et al. [14]	86.7
	Kovashka et al. [16]	82.27
	Le et al.	86.5
	Our method	<b>86.8</b>
UCF50	Laptev et al. [17]	47.9
	Reddy et al. [27]	76.9
	Sadanand et al. [29]	76.4
	Our method	<b>75.9</b>
Hollywood	Wang et suter [37]	47.7
	Taylor et al. [34]	46.6
	Guilbert et al. [8]	50.9
	Li et al. [19]	53.3
	Xin et al. [40]	63.1
	Jain et al. [10]	62.5
	Our method	<b>58</b>

The bold text emphasis our results in order to compare with others

results remain higher than most of the results presented in Table 2. This can be explained by the fact that MSS features capture different boundaries motions. Furthermore, they also capture some details from background which may provide useful context information. Motion boundaries of different objects in scene context, such as ball or horse for example, may be helpful for sports actions which often involve specific equipment and scene types. The dataset KTH and Weizmann are characterized both by their homogeneous background with a pre-defined number of action. These criteria make the classification rate higher than those of UCF and Hollywood datasets. In fact, both UCF and Hollywood datasets represent realistic scenes where background are heterogeneous and actions are not explicitly represented. Moreover, they contain complex actions with inter and intra classes differences. For these reasons, the recognition rate is lower not only for our approach but also for the different state of the art methods. The obtained classification improvement is more than 3% in Weizmann and KTH datasets. Particularly, in UCF sport dataset, the obtained results remains the same as the state-of-the-art methods. It is important to notice here that the quality of results depends on the features modelling complexity.

## 5 Conclusion

This paper introduces a new approach based on Motion Stable Shape (MSS) for video action recognition. To obtain MSS features, each video is transformed into a set of motion boundaries volume. MSERs regions are then detected based on the computed motion boundaries. We have used, also, MSS signatures along with the bag-of-words model to classify different actions for image sequences in several datasets namely KTH, Weizmain, UCF and Hollywood. The experimental results show that the proposed method captures, efficiently, action information in videos and demonstrate good performances compared to state-of-the-art approaches for action classification. Future work will focus on studying the impact of using a deep learning method instead of the SVM classifier in order to perform action recognition and action change detection.

## References

1. Alexander K, Marcin M, Cordelia S (2008) A spatio-temporal descriptor based on 3d-gradients. In: Proceedings of the British machine vision conference. Leeds, pp 995–1004
2. Alexander K, Marcin M, Cordelia S, Andrew Z (2010) Human focused action localization in video. In: International workshop on sign, gesture, and activity (SGA) in conjunction with ECCV, vol 21. Crete, pp 219–233
3. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. In: IEEE Transactions on pattern analysis and machine intelligence, vol 23. Atlanta, pp 257–267
4. Bregonzio M, Gong S, Xiang T (2009) Recognising action as clouds of space-time interest points. In: IEEE Conference on computer vision and pattern recognition, pp 1948–1955
5. Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: IEEE International workshop on performance evaluation of tracking and surveillance (PETS), vol 3. China, pp 65–72
6. Efros AA, Berg AC, Greg M, Jitendra M (2003) Recognizing action at a distance. In: IEEE International conference on computer vision, vol 3. Nice, , pp 726–733
7. Geng C, JianXin S (2015) Human action recognition based on convolutional neural networks with a convolutional auto-encoder. In: 5th International conference on computer sciences and automation engineering (ICCSAE 2015)
8. Gilbert A, Illingworth J, Bowden R (2011) Action recognition using mined hierarchical compound features. In: IEEE Transactions on pattern analysis and machine intelligence. Guildford, pp 883–897

9. Ivan L, Patrick P (2007) Retrieving actions in movies. In: Proceedings of the eleventh IEEE international conference on computer vision, vol 2. Rio de Janeiro, pp 1–8
10. Jain M, Jégou H, Bouthemy P (2013) Better exploiting motion for better action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 21, pp 2555–2562
11. Ji S, Yang WM, Yu K (2013) 3D convolutional neural networks for human action recognition. In: IEEE Transactions on pattern analysis and machine intelligence, vol 35, pp 221–231
12. Jiang Y, Bhattacharya S, Chang S, Shah M (2013) High-level event recognition in unconstrained videos. In: International journal of multimedia information retrieval, vol 2, pp 73–101
13. Jones S, Shao L, Zhang J, Liu Y (2012) Relevance feedback for real-world human action retrieval. In: Pattern recognition letters, vol 33, pp 444–452
14. Klaser A, Marszafek M, Laptev I, Schmid C (2010) Will person detection help bag-of-features action recognition. In: Rapport de recherche 00514828 NRIA
15. Konrad S, Luc J (2008) Action snippets: how many frames does human action recognition require. In: Proceedings of the IEEE international conference on computer vision and pattern recognition. Alaska, pp 1–8
16. Kovashka A, Grauman K (2010) Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: IEEE Conference on computer vision and pattern recognition, pp 2046–2053
17. Laptev I, Marszafek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1–8
18. Lassoued I, Zagrouba E, Chahir Y (2011) Video action classification: a new approach combining Spatio-temporal Krawtchouk moments and Laplacian Eigenmaps. In: 7th IEEE International conference on signal image technology and internet-based systems. Dijon, pp 291–301
19. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE Conference on computer vision and pattern recognition, vol 3. Springs, pp 3361–3368
20. Lena G, Moshe B, Eli S, Michal I, Ronen B (2005) Actions as space time shapes. In: Proceedings of the tenth IEEE international conference on computer vision, vol 2. Washington, DC, pp 1395–1402
21. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide baseline stereo from maximally stable extremal regions. In: Image and vision computing journal, vol 22, pp 761–767
22. Matikainen P, Hebert M, Sukthankar R (2009) Trajectons: action recognition through the motion analysis of tracked features. In: ICCV Workshops on video-oriented object and event classification. Japan, pp 514–521
23. Messing R, Pal C, Kautz H (2009) Activity recognition using the velocity histories of tracked keypoints. In: IEEE International conference on computer vision, vol 21. Japan, pp 104–111
24. Navneet D, Bill T (2005) Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition, vol 1. San Diego, pp 886–893
25. Nistér D, Stewénius H (2006) Scalable recognition with a vocabulary tree. In: Proceedings of conference on computer vision and pattern recognition, vol 2, pp 2161–2168
26. Raptis M, Soatto S (2010) Tracklet descriptors for action modeling and video analysis. In: European conference on computer vision. Crete, pp 577–590
27. Reddy Kishore K, Shah M (2013) Recognizing 50 human action categories of web videos. In: Machine vision and applications, vol 24, pp 971–981
28. Rodriguez M, Ahmed J, Shah M (2008) Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR, pp 85–91
29. Sadanand S, Corso J (2012) Action bank: a high-level representation of activity in video. In: Computer vision and pattern recognition (CVPR), pp 1234–1241
30. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition, vol 3, pp 332–336
31. Scovanner P, Ali S, Shah M (2007) A 3-dimensional SIFT descriptor and its application to action recognition. In: ACM Conference on multimedia. Germany, pp 23–29
32. Sun J, Mu Y, Yan S, Cheong LF (2010) Activity recognition using dense long-duration trajectories. In: IEEE International conference on multimedia and expo. Singapore, pp 322–327
33. Sun L, Jia K, Yeung D, Shi B (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: IEEE International conference on computer vision, pp 4597–4605
34. Taylor GW, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. In: European conference on computer vision. Greece, pp 140–153
35. Thomas S, Lior W, Stanley B, Maximilian R, Tomaso P (2007) A biologically inspired system for action recognition. In: Proceedings of the eleventh IEEE international conference on computer vision. Rio de Janeiro, pp 1–8

36. Thomas S, Lior W, Stanley B, Maximilian R, Tomaso P (2007) Robust object recognition with cortex-like mechanisms. In: *IEEE Transactions on pattern analysis and machine intelligence*, vol 23, pp 411–426
37. Wang L, Suter D (2007) Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In: *IEEE Conference on computer vision and pattern recognition*, pp 1–8
38. Willems G, Tuytelaars T, Gool L (2008) An efficient dense and scaleinvariant spatio-temporal interest point detector. In: *European conference on computer vision*, vol 4. Heidelberg, pp 650–663
39. Wong SF, Cipolla R (2007) Extracting spatiotemporal interest points using global information. In: *IEEE International conference on computer vision*. Rio de Janeiro, pp 1–8
40. Xin M, Zhang H, Wang H, Sun M, Yuan D (2016) ARCH: adaptive recurrent-convolutional hybrid networks for long-term action recognition. In: *Neurocomputing*, vol 178, pp 87–102
41. Yap PT, Paramesran R, Ong SH (2003) Image analysis by Krawtchouk moments. In: *IEEE Transactions on image processing*, vol 12, pp 1367–1376
42. Yeffet L, Wolf L (2009) Local trinary patterns for human action recognition. In: *IEEE International conference on computer vision*. Japan, pp 492–497
43. Yuan J, Liu Z, Wu Y (2011) Discriminative video pattern search for efficient action detection. In: *IEEE Transactions on pattern analysis and machine intelligence*, vol 33, pp 1728–1743
44. Zach C, Pock T, Bischof H (2007) A duality based approach for realtime tvl1 optical flow. In: *Procedure of pattern recognition*, p 214–223



**Imen Lassoued** received an engineering degree in computer science and a master's degree in intelligent systems of imaging and artificial vision from the Higher Institute of Computer Science, Tunisia, in 2007 and 2009, respectively. She is a lecturer at the Higher Institute of Education and Training in Tunisia. Her research interests include computer vision and intelligent imaging.



**Ezzeddine Zagrouba** received his HDR from FST/University Tunis ElManar and his PhD and engineering degree from the Polytechnic National Institute of Toulouse (ENSEEIH/INPT) in France. He is a Professor at the Higher Institute of Computer Science (ISI). He is Vice President of Virtual University of Tunis and the Director of LimTic Research Laboratory at ISI. His main activity is focused on intelligent imaging and computer vision and he is vice president of the research association ArtsPi.