A FAST ACTION RECOGNITION STRATEGY BASED ON MOTION TRAJECTORY OCCURRENCES

G. Garzón*, F. Martínez**

Biomedical Imaging, Vision and Learning Laboratory (BIVL2ab). Motion Analysis and Computer Vision (MACV). Universidad Industrial de Santander, Bucaramanga, Colombia

**e-mail: gustavo.garzon@saber.uis.edu.co*

***e-mail: famarcar@saber.uis.edu.co*

Abstract – A few light stimuli coherently distributed in the space and time are the essential input that a visual system needs to perceive motion. Inspired in such fact, a compact motion descriptor is herein proposed to describe patterns of neighboring trajectories for human action recognition. The proposed method introduces a strategy that models the local distribution of neighboring points by defining a spatial point process around motion trajectories. Particularly, a two-level occurrence analysis is carried out to discover motion patterns that underlying on trajectory points representation. Firstly, local occurrence words are computed over a circular grid layout that is centered in a fixed position for each trajectory. Then, a regional occurrence description is achieved by representing actions as the most frequent local words that occur in a particular video. This second occurrence layer could be computed for the entire video or by each frame to achieve an online recognition. This compact descriptor, with local size of 72 and sequence descriptor size of 400, acquires importance in real-time applications and environments with hardware restrictions. The proposed strategy was evaluated on KTH and Weizmann dataset, achieving an average accuracy of 91.2% and 78%, respectively. Moreover, a further online recognition was performed over UT-Interaction achieving an accuracy of 67% by using only the first 25% of video sequ*ences.*

Keywords: action recognition, active point counting, circular grid, local descriptor, visual words

1. INTRODUCTION

Recognizing human actions is a fundamental task in many areas and applications, such as surveillance and crowd control [17], automatic annotation of human actions in videos [10] and video indexing [16], analysis of sports videos [13], HCI applications [21] and gesture-based video games interaction [22], among other examples [16] [11]. Nevertheless, such applications hardly offer ideal conditions with respect to environmental factors which difficult the action characterization. The typical challenges are reported because of scene variations such as different shapes and clothing, scale changes and movement on the background. Strong variations of the object of interest with respect to the geometry, appearance and motion patterns can also difficult the task of identification and recognition.

Representations of human actions have been studied in the last decades, including global shape approaches and local-interest point representations. Global shape strategies involve the temporal segmentation of regions of interest and the association of the geometrical shape characterization with particular actions.

For instance, Bobick et al. [3] studied the motion presence in video sequences by matching temporal templates against stored shapes of fixed actions. This method reported to be robust to linear changes in speed, and performs well in real-time applications but with limitations to perspective of capture. Also, Gorelick et al. [5] proposed to analyze actions as volumetric shapes computed from silhouettes along video. Such volumes are characterized analytically by using the Poisson equation to obtain saliency, dynamics and structure features. Such proposed approach achieves relative invariance to scale-viewpoint changes but it is dependent of a proper silhouette computation to describe the volumes. Junejo et al. [7] introduced a silhouette representation that combines time series and Symbolic Aggregate approXimation (SAX), while in [1] is proposed a combination of Cartesian Coordinate features, Fourier Descriptors among others that complemented with appearance-based histograms achieved an action recognition. Such approaches are relatively fast but dependent of controlled scenarios of capture.

On the other hand, local-interest point representations have focused on statistical frameworks that combine appearance and inter-frame-based features to recognize actions. For instance, Schuldt et al. [15] used scale-space representations in which Gaussian derivatives were applied to compute gradient keypoints and its orientations were quantized for representation as histogram bins. Also, Laptev [8] proposed a spatio-temporal invariant interest points that maximizes a normalized spatio-temporal Laplacian operator over spatial and temporal scales. Later, Laptev and Lindeberg [9] proposed a space-time interest point representation that is projected to a lower-dimensional space to recover a main vector of representation. These approaches outperform occlusion problems but remain limited to motion direction and appearance variability.

Motion primitives, such as optical flow and long trajectories, have been incorporated in local statistical frameworks to develop strategies robust to appearance changes. For instance, Wang et. al. [19] proposed motion trajectories as sample feature points tracked according to a local optical flow directions. Around the computed motion trajectories were computed local volumes described by Histogram of gradients (HoG), histograms of optical flow (HoF) and histograms of motion boundaries (MBH). Such volumes were used to built a Bag-of-words (BoW) representation of actions in video sequences. This strategy achieved action recognition in uncontrolled videos but with some restrictions for

abrupt camera motions. To outperform such problems, in [20] was proposed a set of improved motion trajectories that filter background and camera motions by using additional matching restrictions from SURF descriptors and taking into account the homography parameters from RANSAC algorithm. Results showed a significant accuracy but additional restrictions increased the computational cost with respect to the dense field strategy.

Additionally, recent action recognition methods are based on deep learning and recurrent architectures that codify features according to the correlation of thousands of video samples. Such strategies have demonstrated high capability to recognize human actions at different scenarios. Nevertheless, these approaches demand large training sets and the adjustment of hyper-parametric functions that require special computational demands to achieve proper results [2][12].

This work introduces a compact motion descriptor that model the local spatial distribution of a set of interest points that are computed along motion trajectories. For doing so, a set of coherent spatio-temporal interest points are computed along the video as a set of motion trajectories. Then, a spatial point process is defined around of each trajectory to measure the spatial distribution of neighborhood point trajectories. Such distribution is measured following a circular counting distribution scheme at several radii and angles. The set of point processes are spatio-temporal features of the action, that are codified from a BoW representation. The proposed motion descriptor achieve a very compact description of actions by using only 72 scalar features to represent each trajectory. A video descriptor with size 400 was computed from BoW to perform the action recognition. Such compact descriptor result fundamental in real-time or applications with hardware limitations. The rest of the paper is organized as follows: in section 2 is introduced the proposed approach to compute a compact descriptor by only using few spatio-temporal points of interest. In this section is also described the modeling of motion trajectories and the strategy of classification. Then, section 3 presents the results and a quantitative evaluation of the method. Section 4 concludes and presents prospective works.

2. PROPOSED METHOD

Human visual system has developed robust mechanisms to perceive and recognize complex actions by coding coherent information. Such fact has been widely demonstrated in Johansson experiments,

where actions were properly identified by analyzing few bright spots spatially distributed that changed over time [6]. In such experiments it was demonstrated that visual systems coded the spatial distribution of interest points that have been moving coherently through time. Inspired in such natural mechanisms, in this work is introduced a novel action recognition strategy based on the local spatial characterization of motion of points of interest.

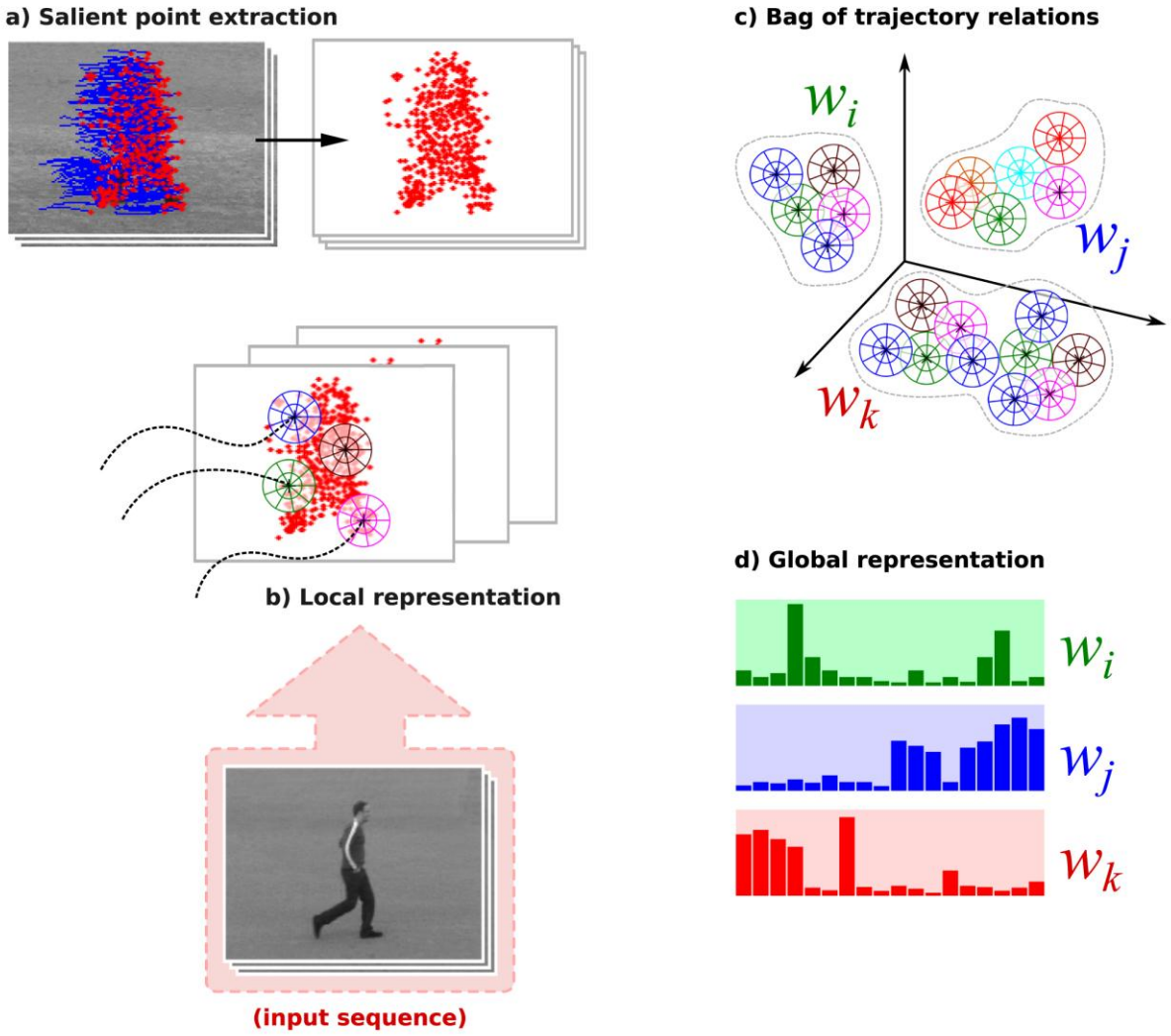

Fig. 1. Pipeline of the proposed method: a) extracting points of interest on a framewise basis, b) characterization for each trajectory (local representation), c) codifying relations between trajectories using a BoW model, and d) histogram representation for the entire video sequence (global representation).

Such motion of interest are herein computed from motion trajectories that remain coherent in a temporal interval of time Δt (see in Figure 1-a and Figure 2). The first occurrence layer is locally computed into a circular grid fixed at each

frame and around of each active trajectory. Then, from each circular grid is carried out a spatial counting process of neighborhood motion trajectories (Figure 1-b) which is further explained in section 2.2. Then each point is marked with the local spatial distribution of motion trajectories and represent atoms of local occurrences in our representation. A second regional occurrence layer is defined as a counting process of local atom descriptors. For doing so, from a set of training videos is firstly computed a dictionary of local occurrence atoms. Then any action can be globally represented in the video sequences by projecting the spatial distribution atoms to the learned dictionary and obtaining a global histogram representation (see in Figure 1-d). This representation can be also computed at frame level, allowing an online action recognition. Finally, such global/frame histograms are mapped to a previously trained Support vector machine to obtain an action label. The pipeline of the proposed approach is illustrated in Figure 1.

2.1. Motion trajectories representation

The herein proposed work starts by computing a set of dense trajectories, allowing a primary motion representation of the action present in the video. As visual perception, such motion trajectories are a set of salient spots that are tracked in a set of frames and recover temporal coherence of local motion. In this work was implemented the improved motion trajectories proposed by Wang et. al. [19]. These dense motion trajectories have demonstrated a proper representation of actions, by following points of interest from a dense optical flow field. The spatio-temporal points of interest are tracked, regarding the velocity field direction, and are smoothed by using a median filter at different spatial scales. Formally, the set of trajectories $T = \{\tau_1, \tau_2, \dots, \tau_N\}$ can be considered as paths that appear during the course of a video, and describe the motion of objects. Each path trajectory $\tau_i = \{\tau_{t_1}, \tau_{t_2}, \dots, \tau_{t_k}\}$ is a set of k spatial points in certain consecutive frames within the video. Once the raw trajectories are computed, some statistical motion filters are implemented to remove trajectories with insignificant or incoherent motion information, according to the local variation of norm velocity. The estimation of these trajectories was also improved by taking into account a camera motion correction assuming a homography relationship between consecutive frames.

A per-frame action representation is defined as the set of spatial points that correspond to active trajectories, described

as the spatial position τ_{t_k} of each trajectory τ_i on the current frame.

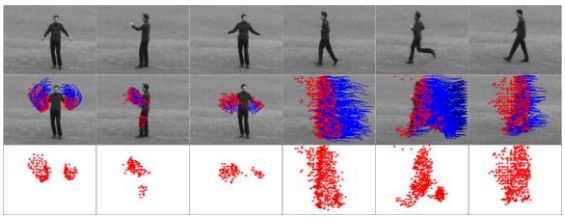

Fig. 2. First row: a set of frames of six different actions. Second row: motion trajectories for six different actions. Third row: red pixels indicating active trajectories for each frame.

An example of computed motion trajectories are illustrated in Figure 2 for different activities recorded in open scenarios. As illustrated in the second row, the red points indicate the frame representation from motion trajectories, while the blue lines code the motion history of each red point. In the third row is shown the point representation that globally can describe the actions by the spatial coding configuration. The local distribution of salient points is defined as a spatial point process around motion trajectories.

2.2. First occurrence layer: a local trajectory representation

Motion trajectories tend to form spatial clusters in regions with a significant number of active points. This fact implies the existence of motion patterns that can be measured in local regions centered around active points. A local point representation is herein achieved by representing each active point trajectory in frame t as the spatial distribution of neighborhood trajectories. In such case the signature of each point is defined by the spatial density of points in its neighborhood. Hence neighborhood positions describe events at random locations, formulated as a finite-set-valued random variable $u = \{u_1(\tau), u_2(\tau), \dots, u_n(\tau)\}$. Those atomic point distributions represent a local spatial point process of motion trajectories contained in a bounded local region.

In this work, a counting scheme is obtained by locally counting the occurrences of neighboring motion trajectories that fall inside a proposed circular bound region. For doing so, a counting process at several circles and angles is defined around of active point trajectories. The circular grid is split in different sub-regions defined by a set of γ concentric circles and α angular divisions. The circular distribution is then achieved by counting motion neighborhood trajectory points that fall into each subregion of the circular grid. A general scheme of the circular counting process is depicted in Figure 3. In this illustration, the spatial distribution is computed as a signature of the red point.

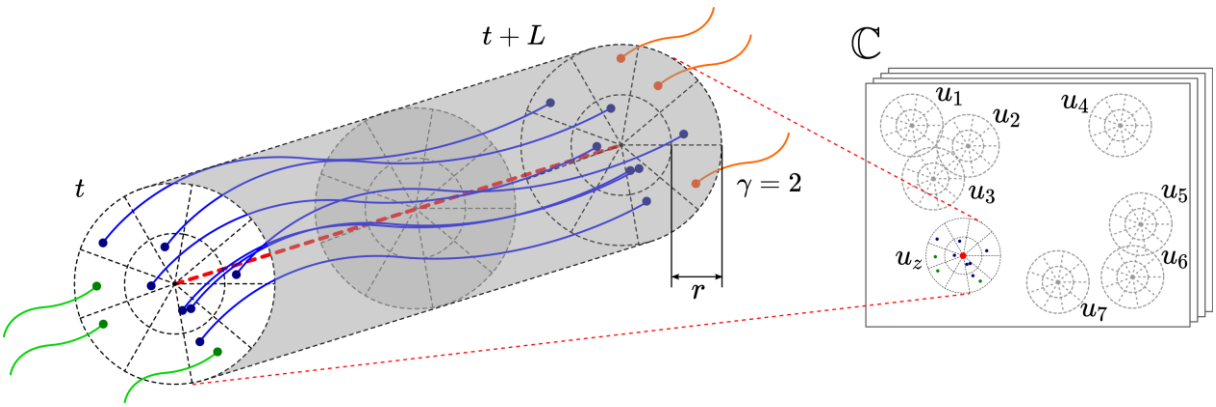

Fig. 3. Bounded region detail: position of neighboring trajectories is characterized using a circular grid layout corresponding to bounded local region **C**. Trajectories that end (green), go through (blue) and start (orange) in the tracking interval L are taken into account.

As illustrated in Figure 4, because the density of important motion regions, some of the point signatures can share counting information, producing redundant data in the representation. This fact result relevant in our description because important motion patterns can be represented in several atomic signatures. Then, any particular spatio-temporal boundary **C** of action sequence is represented by a set of l-dimensional occurrence points $\{u_1, u_2, \ldots, u_z\}$; $u_i \in R^l$, coded from a circular grid. The boundary **C** could be considered as total video sequence in classification tasks or at frame level for online action recognition. This local representation of actions constitutes the input for a mid-level representation, that will describe the spatial point distribution in a higher level.

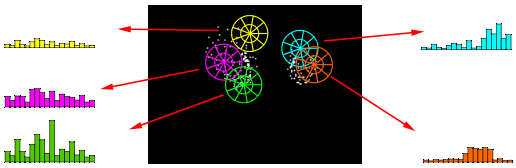

Fig. 4. Grid overlapping on a per-frame representation. Relations between trajectories are efficiently characterized and represent a robust signature for some actions in a local perspective.

In such way, the proposed strategy comprises a hierarchical model that allows to characterize spatial point distributions from a local to a global level to represent actions like observed in natural visual perception systems.

2.3. Second occurrence layer: a Bag of spatial representation words

As is well known, visual mechanisms operate from fine to coarse different scale of analysis to understand and define the

world around. The proposed work completes the local spatial representation by coding the circular distribution points in a mid-level representation. This intermediate regional representation can operate from two possible perspectives: at each frame or for the complete video sequence, according to the considered points distributed into a boundary **C**. A Bag of Spatial Representation Words (BoSRW) is then herein defined to compute a global descriptor of the temporal region by measuring the number of occurrences of each learned centroid in the processed video.

Firstly, a dictionary of representative spatial words is computed from a non supervised strategy over a set of training videos. For doing so, a k-means was herein implemented to compute k representative circular distribution points, defined as: $D = [d_1, d_2, \dots, d_k] \in^{1 \times K}$. Each of these centroids are recovered from an objective function:

$$D(k) = \min_{d_k} = \sum_{m=1}^{M} \sum_{k=1}^{K} \left\| u(\tau)_m - d_k \right\|_2^2, \quad \text{as}$$

the K points closer to each of the members of resulting groups. In this sense, the actions are assumed that can be globally described by the distribution of these centroids d_i in each particular video.

Once the dictionary is computed, a mid level action representation is obtained by coding the set of spatial words in global histogram occurrences w.r.t the learned centroids. This representation could be defined into a particular time interval Δt such as the complete video or even at level of a single frame. For each temporal interval, each computed word u_n contributes with the count occurrence only for the centroid i with minimal Euclidean distance s, process defined as: $s(i) = 1$, if $i = \arg \min \left\| u_n - d_j \right\|^2$ s.t. $\left\| s \right\|_0 = 1$. A normalized version of the set of K centroid occurrences constitutes the histogram representation for the action in such particular interval of time.

A clear advantage of BoSRW is that actions can be described from partial sequences and is also robust to occlusion problems during capture. Compared with circular distribution points at mid-level, this analysis is not directly related to the neighboring trajectories over regions, allowing scale invariance in the description of motion patterns. Redundant information enrich the description of actions and add favorable complexity to the model. This representation of motion clusters result in a compact scheme for action recognition and constitutes the input for classifying human action sequences.

2.4. Supervised action classification and recognition

The computed spatio-temporal descriptor at the larger region ocurrence scale, coded from fine to coarse active point representation, is used as the input of machine learning strategies to obtain an automatic action recognition. To achieve a proper trade-off between prediction time and accuracy, in this work were evaluated different classification methods based on kernel functions and random forest strategies. On the one hand, a multi-class Support Vector Machine (SVM) was herein implemented by using a One against one SVM multiclass classification and evaluated using the linear and the Radial Basis Function (RBF) kernels [4]. The SVM with linear kernels allows to deal with fast action recognition, one of the main goal of the proposed work, while the RBF lead with non-linear class separations.

On the other hand, the random forest (RaF) classification consists on a set of decision tree classifiers grouped in an Ensemble Learning strategy, allowing to overcome sensibility problems of decision by taking mode decisions over the set of computed trees. In our specific approach, a set of independent DT algorithms are trained over different parts of global occurrence histograms to reduce the variability in the prediction. The final prediction may be carried out by averaging the predictions of individual trees f_b, or taking the majority vote as expressed $\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b$ ^ $\hat{y} = \arg\max \{f_1, \ldots, f_B\}$. Once the trained models are adjusted from spatio-temporal descriptor, the proposed strategy result efficient in time to perform online prediction over sequences of actions.

2.5. Data

The proposed approach was widely evaluated by using three different public and academic datasets. These datasets were selected to test hypothesis of the relationship between actions and the modelling of salient spatio-temporal points. Taking such fact into account, the selected datasets captured only one person into each video and the actions are mainly described by its dynamic. A description of the datasets is presented as follows:

- KTH [15]: This human action dataset contains six types of human actions: walking (Walk), jogging (Jog), running (Run), boxing (Box), hand waving (HW) and hand clapping (HC) which are performed by 25 subjects in indoor and open scenarios with some scale variation in several sequences. The subjects in this scenario exhibit different clothes. Following the suggestion of KTH authors, the total of 2391 video sequences were

split in three different groups to carry out the evaluation: training (760 sequences), test (863 sequences) and validation (768 sequences).

- Weizmann [5]: this dataset is composed by 90 video sequences that comprehend 10 actions such as walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack and skip. Validation was carried out using a leave-one-out strategy, taking into account the standard validation reported in the state of the art.

- UT-Interaction [14]: this surveillance dataset is captured in open scenarios, with some variations in terms of appearance and dynamic performance of actors executing actions. The dataset is made of two sets of 60 sequences, that contain 6 actions namely shake-hands, point, hug, push, kick and punch. Validation is executed with a 10-fold leave-one-out strategy, as suggested by the dataset's authors.

3. RESULTS

The proposed descriptor was evaluated in two different tasks: activity classification using complete sequences and partial sequence recognition from frame-level action representations. In both experiments a very compact descriptor achieved a trade-off between accuracy and fast activity prediction, being ideal for real time applications. Both evaluation tasks are described in the next subsections:

3.1. Activity classification

A first evaluation of the proposed approach was carried out to obtain the best configuration in terms of number of concentric circles (γ) and their respective radius (r), on each evaluated dataset. The number of circles over our circular grid was set as $\gamma=\{4, 5, 6, 7, 8\}$, increasing the size to reach of the neighborhood around the averaged center. Also, the radius of each circle was set as $r=\{6,8,10\}$ to admit more trajectories inside each region. These results were obtained w.r.t a global descriptor of 400 words and using the SVM with a RBF kernel, as strategy for classification.

Table 1. Preliminary experiment for KTH (top) and Weizmann (bottom) datasets: accuracy (%) for combinations of γ circles and radius r (px).

γ \ r	6	8	10
4	88.52	90.26	90.26
5	90.73	89.68	89.33
6	90.61	89.91	88.41
7	88.87	87.83	87.60
8	**91.20**	88.64	88.18

γ \ r	6	8	10
4	76.66	76.66	73.33
5	73.33	72.22	**78.88**
6	74.44	72.22	72.22
7	75.55	73.33	71.11
8	78.88	71.11	71.11

In table 1-top and 1-bottom is reported the performance of the proposed approach using different (γ, r) configurations on KTH and Weizmann datasets, respectively. For KTH the maximum score of 91.2% was achieved with the tuple (γ=8, r=6), standing out small radius steps because of the spatial frame resolution in this dataset. Hence a larger number of γ circles was necessary to codify actions in terms of trajectory densities around each point, mainly because the dynamic closeness of some activities, such as: jogging and running. As reported in table 1, the proposed approach achieves more than 87% of accuracy for almost all configurations, demonstrating its robustness to represent actions given the different (γ, r) configurations.

Regarding Weizmann dataset, the best score (78.88%) was achieved with the tuple (γ=5, r=10). This dataset groups most action classes and therefore the classification is more challenging, w.r.t to KTH dataset. Larger radius sizes have better performance, since the spatial frame resolution is much more larger than for KTH. Also, this frame resolution allows a better density trajectory description around each point, and therefore fewer γ circles are sufficient to describe the actions. As expected, for Weizmann dataset are also reported stable results for different (γ, r) configurations, obtaining more than 70% of accuracy to describe the 10 different actions.

Table 2. Preliminary experiment for UT-Interaction dataset (set1 - top, set2 - bottom): accuracy (%) for combinations of γ circles and radius r (px).

γ \ r	6	8	10
4	65	65	66.66
5	65	68.33	70
6	65	70	68.33
7	66.66	70	70
8	68.33	**71.66**	66.66

γ \ r	6	8	10
4	**65**	56.66	61.66
5	60	53.33	60
6	56.66	65	56.66
7	61.66	58.33	53.33
8	61.66	60	53.33

The proposed approach was also evaluated on UT-Interaction dataset that represent activities in almost real conditions. This dataset is split on: set 1, which was captured with a relatively static camera,

and set 2, where videos were recorded with some camera jitters. In tables 2-top and 2-bottom are reported the performance of the proposed approach for set 1 and set 2, respectively. As expected, the best accuracy (71.66%) is obtained on set 1 with a tuple configuration of (γ=8, r=8), while for set 2 with tuple (γ=4, r=6) was achieved a 65%. The set 2 of UT-Interaction has better performance on more reduced occurrence space, since most of the trajectories could be related with camera motions, introducing artifacts to the analysis of activities. In contrast, the set 1 achieves a better performance doing a wide analysis from several circles with a radius of 8 px. Interestingly enough and despite of the challenge of the UT-Interaction dataset, almost whole (γ, r) configuration achieves similar scores showing a stable performance of the proposed descriptor.

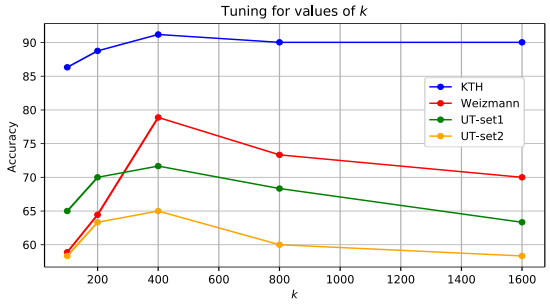

Fig. 5. Accuracy for different k-values given best configurations over KTH, Weizmann and UT-Interaction datasets.

In a second experiment, the dictionary size, that code main words for all actions on training video sequences, was adjusted for different k centroids, which allowed to find an adequate balance between dimensionality of descriptor and accuracy. Using a SVM with a RBF kernel, the size of the dictionary was increased from histograms with sizes of 100 to 1600. For all evaluated datasets the best overall accuracy was obtained for k=400 visual words, as shown in figure 5. It is worth noting that much of the state-of-the art strategies achieve stable results for descriptors with thousands of values, while the proposed approach achieve the best performance by only using 400 scalar values to represent a complete video sequence. Such compact descriptor result ideal as a first stage of identification, allowing also to highlight the importance of active spatial points to represent actions developed by the subject. Additionally, the descriptor shows a good trade-off between dimension size and accuracy.

In a third experiment, several classification strategies were evaluated over the best (γ, r) configurations obtained in previous experiments. The classification strategies herein selected allow to deal with non linear spaces, with appropriate balance regarding computational time. The classification strategies were: Random Forest (RaF), SVM (with linear kernel) and SVM (with RBF kernel). The whole

strategies were evaluated under a grid search parameter framework to obtain best configurations w.r.t to the proposed action descriptor and the respective dataset evaluated.

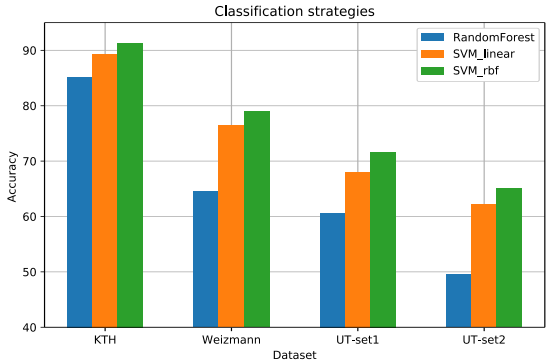

Fig. 6. Classification strategies for KTH, Weizmann and UT-Interaction datasets. Best overall accuracies are shown.

In three evaluated datasets the SVM with RBF kernel achieved the best performance in terms of accuracy. Such fact could be associated to the properties of kernel to separate actions through radial basis functions w.r.t support vectors of reference rather than linear boundaries. It should be also noted that linear kernel respond to a proper performance of accuracy but with the main advantage of low computational complexity, which could be a perfect tool to obtain fast prediction on real-time applications. Despite the lower performance of random forest classifier, the obtained results on KTH and Weizmann dataset are competitive and useful on low-level architectures. Also this classification strategy is interesting as a pre-processing step to discover most relevant features on the descriptor. Finally, prediction of human actions constitute a task that often requires reduced execution times. Our descriptor takes only 0.036 seconds in average to correctly recognize human actions on standard video sequences.

3.2. Partial sequence Recognition

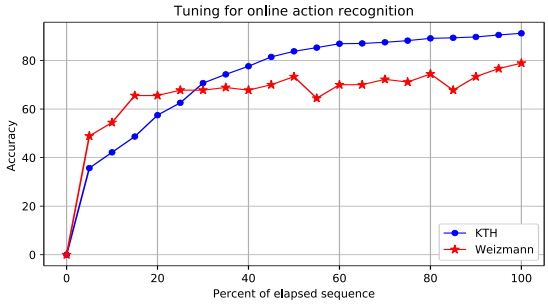

Fig. 7. (Blue line) Performance simulation for an online application over KTH dataset. Our method achieves promising results with just 25% of the total number of frames. (Red line) Performance simulation for an online application over Weizmann dataset. Our method achieves promising results with just 15% of the total number of frames.

A second evaluation of the proposed approach was carried out on partial sequences computed at frame level representations. Such experiments deal with the capability of the strategy regarding online and partial recognition of action, setting face to incomplete

information of the dynamic. To obtain a statistical quantification of partial recognition performance, the proposed approach was tested using different temporal percentages of the video sequences, and an averaged result was registered for all datasets.

In figure 7, blue line illustrates the recognition performance of the proposed approach for the KTH dataset. A proper action prediction is achieved by using almost half of the video sequences, reporting more than 70% of accuracy. Also, a natural increasing of prediction is obtained while more frames are used into the prediction. Such fact is associated to the periodic nature of recorded actions.

Additionally, the red line in figure 7 represents the recognition achieved by the proposed approach for Weizmann dataset. A fast recognition is herein achieved by using only the 15% of video sequences. After that, the proposed approach remains relatively stable, showing some mild increasing at the end of the sequences. Some little accuracy variations could be associated to action duration of different activities, and the relative non-regular cropping of video sequences.

Finally, in figure 8 is illustrated the online recognition performance for UT-interaction in both set of videos. Despite of difference of both sets, w.r.t, to the camera motion and other actions on the background of the scenes, the proposed approach achieves a very similar performance on the task of recognition.

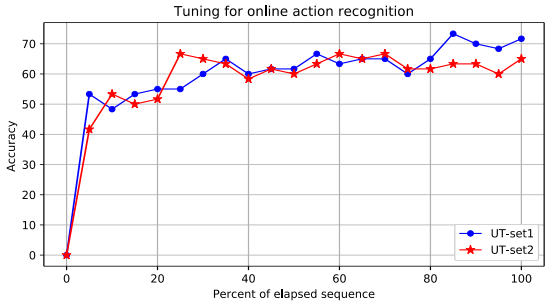

Fig. 8. Performance simulation for an online application over UT-Interaction dataset.

Interestingly enough, the proposed approach achieves the best score (with ~ 67%) on set 2 on an intermediate representation, by using only the 25% of the video sequences. Also for set 1, the best score (~ 73%) is achieved by using the 85% of the video sequence. These partial results are even better that the obtained results on classification evaluation, previously described. For both cases, the prediction, as in other datasets, remain relatively stable for almost all entire sequences.

4. DISCUSSION

The proposed approach introduced a reduced size descriptor that recognize

actions through a local and global occurrence analysis of active trajectories. Such recognition exploits spatio-temporal information that lies within neighboring trajectories. A proximity criteria defines which trajectories will belong inside a neighborhood that is bounded by concentric circles and angles. Experimental results evidence a competitive performance on classification and recognition task under academic and well know action datasets. A main advantage of the proposed approach is the compact frame-level representation of the action, which result useful on low level computational architectures and for applications that require online predictions on video streamings.

In literature, much of the proposed action recognition strategies are dedicated to classify action on well cropped sequences, which also namely require high dimensional descriptors to represent actions. A seminal word coding action representation was proposed by Laptev et. al. [10], that codes actions from occurrences of appearance patches computed along the video. This approach requires video descriptors with size of 4000 scalar values, and fails in much of the cases because of the dependency of appearance of the actions. Currently, as an alternative to appearance based descriptors, the work proposed by Wang et. al. [20] compute cuboids descriptors around motion trajectories along video sequences. Such cuboids codes gradient appearance and motion information of the actions. Significant results are achieved using this strategy, over challenging dataset but requiring high dimensional descriptors to represent the activities. In such case, the best configuration is achieved with cuboids of size 32x32x16, and global occurrence histograms of size 4000. In contrast, the proposed approach operates at two level of occurrences, obtaining very compact descriptors to represent actions. In terms of occurrences around trajectories, the size of descriptor is 9x8x16, while the global occurrence is achieved with compact histograms of 400 scalar values. Additionally, the strategy is able to recover a global representation from partial representations.

Currently, sophisticated deep learning strategies are proposed in the state-of-the-art to deal with action challenges, such as camera motion, scene representation and even w.r.t to the dynamic representation of the actions [2][12]. Such approaches nevertheless require a huge number of samples to achieve coherent results, and also require complex setups of training to achieve proper accuracy results. Because much of these approaches are based on convolutional representations, a fixed interval of the action is required. Other approaches based on LSTM architectures

have been proposed to deal with temporal and dynamic correlation of actions. For instance, Veeriah et al. [18] introduced a LSTM-based proposal that uses derivatives of gait states as a threshold criteria for a temporal segmentation of actions. This approach achieves a 92.12% accuracy over KTH dataset, using 450 scalar values obtained with PCA from a 56.000 sized feature vector. Regarding the proposed approach (accuracy of 91.2%), there is not statistical significance on the obtained results over the KTH dataset. Also, the LSTM-based approach achieve a compact description but requiring an additional processing of dimensionality reduction from the PCA.

The proposed method can be adapted to different classifying strategies and is not dependent on high performance hardware specifications, which facilitates online action recognition and a low-cost implementation. Taking into account the current reported results, our strategy can also be adapted to recurrent neural networks architectures in order to exploit the richness of sequential information regarding human action. Moreover, our method can be used as an input for finding correlation over different motion trajectories and the joint variability of motion kinematics in order to recognize actions.

5. CONCLUSSION

In this work was presented a compact descriptor to recognize actions in video sequences using a local strategy that statistically model active points distribution, integrating local and mid level analysis. The atomic point distributions are described by only using 72 features per active frame. At mid-level representation it was only necessary a global representation of 400 visual words. The dimension of descriptors at local and mid level are useful for real-time computations and for applications with limited hardware resources. The proposed approach also demonstrated competitive results in terms of accuracy for the KTH dataset. Future works include the analysis of additional features to represent active points, as well as additional statistical distributions. A more deeply evaluation over different academic public datasets will be also carried out.

FUNDING

This work was partially funded by the Universidad Industrial de Santander. The authors acknowledge the Vicerrectoría de Investigación y Extensión (VIE) of the Universidad Industrial de Santander for supporting this research registered by the project: Reconocimiento continuo de expresiones cortas del lenguaje de señas, with SIVIE code 2430.

REFERENCES

1. S. Al-Ali and M. Milanova and H. Al-Rizzo and V.L. Fox, "Human action recognition: contour-based and Silhouette-based approaches", In Computer Vision in Control Systems-2, Springer, pp. 11-47, (2015).
2. M. Baccouche and F. Mamalet and C. Wolf and C. Garcia and A. Baskurt, "Sequential deep learning for human action recognition", International Workshop on Human Behavior Understanding, Springer, pp. 29-39, (2011).
3. A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates", In IEEE Transactions on pattern analysis and machine intelligence, IEEE, Vol. 23, pp. 257-267, (2001).
4. C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines", In ACM transactions on intelligent systems and technology (TIST), ACM, Vol. 2, (2011).
5. L. Gorelick and M. Blank and E. Shechtman and M. Irani and R. Basri, "Actions as space-time shapes", In IEEE transactions on pattern analysis and machine intelligence, IEEE, Vol. 29, pp. 2247-2253, (2007).
6. G. Johansson, "Visual perception of biological motion and a model for its analysis", In Perception & psychophysics, Springer, Vol. 14, pp. 201-211, (1973).
7. I. Junejo and K.N. Junejo and Z. Al Aghbari, "Silhouette-based human action recognition using SAX-Shapes", In The Visual Computer, Springer, Vol. 30, pp. 259-269, (2014).
8. I. Laptev, "On space-time interest points", In International journal of computer vision, Springer, Vol. 64, pp. 107-123, (2005).
9. I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition", In Lecture notes in computer science, Springer, Vol. 3667, pp. 91-103, (2006).
10. I. Laptev and M. Marszalek and C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies", In Computer Vision and Pattern Recognition, 2008. CVPR 2008, IEEE, pp. 1-8, (2008).
11. R. Poppe, "A survey on vision-based human action recognition", In Image and vision computing, Elsevier, Vol. 28, pp. 976-990, (2010).
12. H. Rahmani and A. Mian and M. Shah, "Learning a deep model for human action recognition from novel viewpoints", In IEEE transactions on pattern analysis and machine intelligence, IEEE, Vol. 40, pp. 667-681, (2018).

13. M.D. Rodriguez and J. Ahmed and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition", In Computer Vision and Pattern Recognition, 2008. CVPR 2008, IEEE, pp. 1-8, (2008).
14. M.S. Ryoo and J.K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities", In International Conference on Computer Vision, IEEE, pp. 1593-1600, (2009).
15. C. Schuldt and I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", In International Conference on Pattern Recognition, IEEE, Vol. 3, pp. 32-36, (2004).
16. T. Subetha and S. Chitrakala, "A Survey on human activity recognition from videos", In Information Communication and Embedded Systems (ICICES), IEEE, pp. 1-7, (2016).
17. M. Takahashi and M. Naemura and M. Fujii and S. Satoh, "Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories", In Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, pp. 9-16, (2011).
18. V. Veeriah and N. Zhuang and G.J. Qi, "Differential recurrent neural networks for action recognition", In Proceedings of the IEEE international conference on computer vision, IEEE, pp. 4041-4049, (2015).
19. H. Wang and A. Kläser and C. Schmid and C.L. Liu, "Dense trajectories and motion boundary descriptors for action recognition", In International journal of computer vision, Springer US, Vol. 103, pp. 60-79, (2013).
20. H. Wang and C. Schmid, "Action recognition with improved trajectories", In Proceedings of the IEEE international conference on computer vision, IEEE, pp. 3551-3558, (2013).
21. Y. Wu and T.S. Huang, "Vision-based gesture recognition: A review", In Gesture Workshop, Springer, Vol. 1739, pp. 103-115, (1999).
22. G. Zhu and L. Zhang and P. Shen and J. Song, "An online continuous human action recognition algorithm based on the kinect sensor", In Sensors, Multidisciplinary Digital Publishing Institute, Vol. 16, pp. 161, (2016).

Gustavo Garzón Villamizar. Junior researcher at Biomedical Imaging, Vision and Learning Laboratory (BIVL2ab) of Universidad Industrial de Santander, Bucaramanga, Colombia. Graduated from Systems Engineering program in 2016. His research interest include action recognition, motion analysis and deep learning.

Fabio Martínez Carrillo. Is currently a full-time professor at Universidad Industrial de Santander on the Computer Engineering and Informatics School. He is part of the Biomedical Imaging, Vision and Learning Laboratory. In past, he did a postdoc working together with U2IS (ENSTA-ParisTech) - LIMSI-CNRS (Université Paris-Sud) of Université de Paris-Saclay. He has also worked in CIM@LAB laboratory and BioIngenium Research Group, Bogotá-Colombia, in Laboratoire du Traitement du Signal et de l'Image, Université de Rennes 1. Rennes-France and Laboratoire d'Electronique et Informatique. ENSTA-ParisTech. Paris-France. His major interest are on video processing, models of motion to action recognition, machine learning, applications related with computer vision, and medical imaging processing.