International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 1
ISSN: 2088-5334

Local Trajectory Occurrence Patterns for Partial Action and Gesture Recognition

Gustavo Garzón[a,*], Fabio Martínez[a]

[a]*Biomedical Imaging, Vision and Learning Laboratory (BIVL2ab). Universidad Industrial de Santander, Bucaramanga, Colombia*
Corresponding author: [*]gustavo.garzon@saber.uis.edu.co

***Abstract*— Action and gesture recognition is essential in computer vision because of their multiple and potential applications. Nowadays, in the literature, dramatic advances have been reported regarding recognizing gestures and actions under uncontrolled scenarios with significant appearance and motion variations. Nevertheless, much of these approaches still require manual segmentation of temporal action boundaries and complete processing of whole sequences to obtain a prediction. This work introduces a novel motion description that can recognize actions and gestures over partial sequences. The approach starts by representing video sequences as a set of key-point trajectories. Such trajectories are then hierarchically represented from a local and regional perspective, following a statistical counting process. Firstly, each trajectory is defined as a binary occurrence pattern that allows for standing out critical motions by neighborhood densities from a local perspective. Such occurrence patterns are involved in a regional bag-of-words representation of actions. Both representations could be obtained for any interval inside the video, achieving a partial recognition of motion, and regional representation is mapped to a support vector machine to obtain a prediction. The proposed approach was evaluated on academic action recognition datasets and a large gesture dataset used for sign recognition. Regarding partial video sequence recognition, the proposed approach achieves an accuracy rate of 63% using only 20% of frames. The proposed strategy achieved a very compact description, with only 400 scalar values, which ideal for online applications.**

***Keywords*— Action recognition; binary motion patterns; occurrence patterns; motion trajectories.**

Manuscript received 15 Jul. 2019; revised 30 Aug. 2020; accepted 10 Dec. 2020. Date of publication 28 Feb. 2021.

I. INTRODUCTION

Recognizing human actions is a fundamental task in many areas and applications, such as surveillance and crowd control [1], automatic annotation of human actions in videos [2] and video indexing [3], analysis of sports videos [4], HCI applications [5] and gesture-based video games interaction [6], among other examples [7]. Nevertheless, such applications hardly offer ideal conditions concerning 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 background movement. Substantial variations of the object of interest concerning the geometry, appearance, and motion patterns can also difficult the task of identification and recognition.

Action recognition has been approached with strategies such as representation with human silhouettes, Spatio-temporal patches, motion primitives, and exhaustive learning strategies that aim to comprehend motion patterns using significant amounts of data. These approaches have reached significant results regarding predicting and modeling actions, even in complex scenarios and with abrupt changes of dynamic modeling of activities. Nevertheless, these approaches are dependent on a temporal segmentation of activities, which strongly restrict their use on streaming activity prediction and online applications [8].

Classical approaches model actions from geometric primitives are computed from temporal silhouettes [9] to local representations that deal with occlusion problems [10]. These approaches are generally limited to describing global or appearance-based changes along time, losing dynamic representation that could be determinant to differentiate among comparable actions. A set of kinematic primitives have been explored to represent activities to overcome such limitations, which include optical flow descriptors [11], [12], tracking approaches, and current strategies based on long motion trajectories [13]–[15]. These trajectories and motion-based approaches include deep local representations that allow facing with complex real-life action datasets. However,

such representations require a complete motion sequence characterization to explore the coding activities.

Currently, exhaustive, deeply, and in some cases, recurrent learning strategies have emerged to represent activities in videos robustly. These approaches report state-of-the-art results regarding the accuracy of predicting and recognizing actions in very complex and large datasets [16]. Nevertheless, these approaches still require extensive training data and demand many hyper-parameters to represent the actions with a proper performance [17]. A significant limitation of these approaches is the dependency of temporal boundary priors that delimit actions to learn and predict activities. In such cases, deep and convolutional architectures require complete video-sequences to learn spatial and temporal patterns that are associated with actions of interest [18]. Some recent approaches have used recurrent structures to exploit temporal relationships but still require large datasets to detect and identify actions [8] virtually. Also, one-shot strategies have emerged from such deep architectures allowing to predict and anticipate action [19]. This recent approach is, however, limited to very controlled and motion structured datasets.

This work introduces a robust and compact action recognition strategy with the capability to predict high-level instances at frame level or partial video-sequences. The approach is based on an occurrence model of key-motion points that are tracked as long trajectories along with the video. Such trajectory points serve as support to locally quantify critical motion density around each trajectory. A binary Spatio-temporal point process is then defined as an occurrence measure of key-points around each trajectory and following a circular grid to perform the process. According to a local occurrence threshold, each cell inside the circular grid is discretized on a binary base. Such bit-vector representation constitutes a local motion descriptor that allows forming a dictionary of bit-vectors to code each of the video sequences. Then, for any video sequence interval, the computed bit-occurrence-vectors are mapped to the previously trained dictionary to obtain a frame occurrence descriptor representing the action or gesture present on a particular sequence. This frame-occurrence-descriptor is mapped to a previously trained machine learning algorithm to obtain an automatic prediction. The strategy is very compact and usable on online recognition applications.

On the one hand, the local bit-vector occurrence has an average size of 51 scalar values, while the frame-occurrence-descriptor has 400 values. For online action recognition, the proposed approach achieved 63% and 64% accuracy using only the 20% and 15% of the elapsed video for KTH and Weizmann's public datasets, respectively. Regarding gesture recognition, the proposed approach achieved an average score of 89%, using a total of 64 signs, recorded in a total of 3200 sequences. The proposed approach achieved an accuracy of 66% for partial video-sequence recognition using 65% of the elapsed video in such gesture dataset. A preliminary version of this article has appeared in Garzón et al. [20]. This paper is organized as follows: Section II consists of an in-depth review of a set of strategies that focused on local binary patterns to describe several images and video primitives in action classification and recognition tasks. Then, Section III introduces the proposed approach that model the occurrence of key-motion points as LBP patterns to code developed actions, with the capability to code information at frame-level, which results in useful for online applications. Then, in Section IV, the proposed approach is widely evaluated using several actions and gesture recognition public datasets. Finally, some final conclusions are developed in Section V.

II. Materials and Method

It has been widely demonstrated that local features achieve a more robust characterization of non-trivial objects during years, like humans developing complex actions. Among these descriptors, the Local Binary Patterns (LBP) have demonstrated relevant results on many different applications, ranged from texture classification, object detection, and video analysis [21], [22]. The principle of LBP is to measure local differences around central interest points and to code such differences in a bit-string vector. This binary vector representation, generated from a bit-vector quantification, allows robustly describing each key points information as a signature within a bounded space. For doing so, it was defined as a metric to compare neighborhood points, where positive differences concerning the center points are marked as one, and otherwise, the bit (of differences) is fixed as zero.

Specifically, for action recognition strategies, the LBP descriptors represent complex action dynamics by coding texture difference patterns and kinematic primitives computed from optical flow fields. A modification of this descriptor, the local ternary patterns (LTP), was proposed to compute coded human shapes over orthogonal planes in video sequences. This approach allowed a temporal analysis of shapes representing actions but resulted in the sensible to occlusion and limited to static camera captures [23]. The local trinary patterns, proposed by Yeffet et al. [24], are locally built Spatio-temporal texture descriptors that robustly represented complex sequences, even over challenging light conditions. In such a strategy, a high-level representation is achieved by occurrence histograms of these trinary patterns. Nevertheless, such an approach required complete video sequences to achieve a proper representation of the recorded phenomenon.

Complementary, Nguyen et al. [25] integrated LBP and LTP descriptors, allowing to follow patches along time. This strategy codified velocity and texture descriptors from local positions but was limited to static backgrounds. In this last case, with challenging scenes, the descriptor could be focused only on background patterns. These LBP strategies have been extended and integrated on RGBD sequences regarding depth information, as Depth Motion Maps (DMMs). The LBP has also been integrated on CNN's deep architectures for texture recognition, achieving complementary deep representation and enhancing image classification results [26]. Besides, the deep learning strategy, named TEX-Nets, was proposed to approach the dense expression of texture patterns by starting with a projection of LBP codes into a 3D metric space [27].

This work introduces a novel approach that spatially models motion key-points by considering local occurrence measures. As visual systems that codify spatial distribution of points to represent and identify complex actions, the proposed approach stands out with a dense local representation and coherent global distribution at each frame. These points are taken from the decomposition of motion trajectories that followed the main object of interest into a video sequence. Then, such points are described according to the occurrence

motion trajectory in the neighborhood. The shape distribution of such issues is coded into a circular grid and binarized according to a particular threshold. An additional modification was herein presented that considered each spatial distribution of key-points by weighting the occurrence with the sum of norm velocities of points inside the circular grid. Such local occurrence descriptors are projected to a dictionary to obtain a partial mid-level representation of actions at each frame. Then, at any video sequences interval, the resulting descriptor is projected to a machine learning strategy to achieve an automatic action recognition. The pipeline of the proposed approach is represented in Figure 1. A preliminary version of this work appeared in Garzón et al. [20].

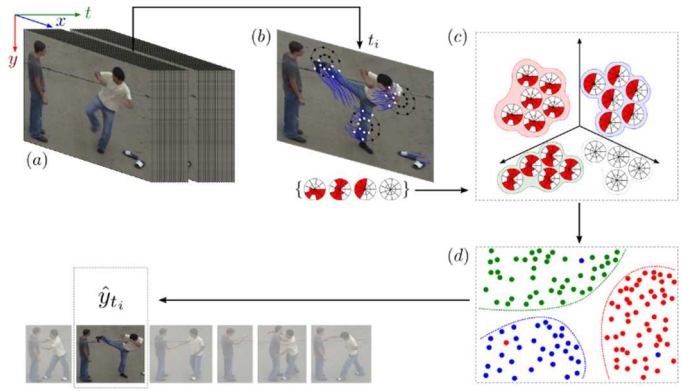

Fig. 1 General outline of the proposed approach: (a) For any frame t_i motion trajectories are obtained. (b) In order to perform an LBP-based analysis for motion vector codification, a concentric circular grid is defined. (c) A regional dictionary-based representation is created from a vector of motion density, and (d) a label $\hat{y}(t_i)$ for each frame t_i of the sequence is assigned to obtain a prediction.

A. Motion Trajectories Representation

Dense motion trajectories have become a reference point on video analysis to locally represent and stand out points with motion coherence along a particular period of a video sequence [15]. These primitives are a set of motion trajectories that densely follows velocity patterns computed from a Farneback dense optical flow field $\omega_t = (u_t, v_t)$. Hence, it is defined as a regular grid in a particular frame, from which is taken velocity pixel patterns. These selected velocity vectors point out to the next position that will be tracked. Then, each motion trajectory is computed by concatenating these spatial-points $(p_t, p_{t+1}, p_{t+2}, \ldots, p_{t+N})$ according to the corresponding velocity vectors. These motion trajectories are bounded in a fixed interval of N frames to prevent corrupt motion incoherence along the codified paths. These problems could occur because video sequences are captured under changing illumination conditions, with additional partial occlusion or abrupt changes of perspective.

In this work, the set of motion trajectories P, computed from a video, constituted a primary kinematic action representation. These trajectories are additionally filtered out to remove paths without relevant dynamic information. For instance, the static trajectories (small (x, y) displacement) or trajectories with sudden motion patterns between consecutive frames are removed from the representation. The resulting filtered paths then mainly represent the dynamic of the object of interest (see Figure 2), whose density allows to highlight local and regional signatures of the actions statistically. In this work, we consider that motion trajectories have sufficient information about the kinematics of activities, and therefore they are used exclusively to code descriptors. Additional appearance patterns could restrict the flexibility of representation on particular actions and gestures herein evaluated.

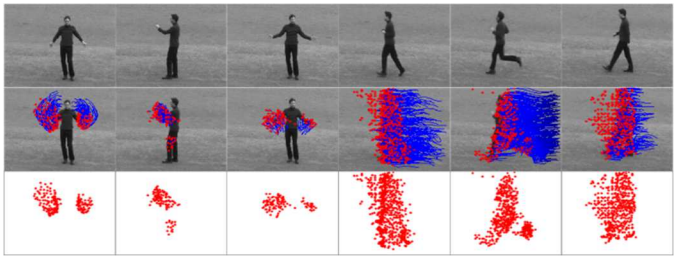

Fig. 2 Top row: a collection of frames related to six different actions from KTH dataset. Middle row: motion trajectories associated with six different actions. Bottom row: red pixels were demonstrating active trajectories over a high-contrast background frame.

B. Kinematic Occurrence Binary Patterns (ToBPs)

A set of significant local kinematic patterns are herein coded around each point, at the frame level, of motion trajectories to represent video sequences activities. Each motion point is then described by the density of motion information around its neighborhood and coded into a circular grid. This occurrence coding is technically defined as a spatial point process composed of spatial point trajectories that fall inside a bounded local region. Hence, each active trajectory, in a particular frame t, is modeled as a random variable $s = \{s_1(\tau), s_2(\tau), \ldots, s_n(\tau)\}$, whose density defined a particular signature of the point on the study.

For doing so, the trajectory counting process is implemented by fixing a set of circles, split up at several angles. This configuration allows increasing resolution of occurrence process into the defined circular region. Specifically, the sub-regions r_i bounded by γ concentric circles and α angular divisions stored occurrence patterns of spatial kinematic patterns. These patterns could be fixed by merely coding the number of point trajectories but eventually could store such points' velocity norm to recover high kinematic levels. A more detailed description of two alternatives for representing actions is described as follows:

1) Occurrence Motion Patterns (ToBPs): A first approximation of local occurrence-based representation, around each spatial point trajectory, was herein carried out by considering the spatial distribution of spatial motion points. Spatial distribution of points is measured as occurrences at each segment of a grid circle, following an operator like LBP (ToBPs). Additionally, to form binary information, a Minimal Number of Trajectories (MNT) threshold τ is defined to fix each grid segment with a value one or zero, according to the occurrence pattern. Then, the set of points $T \approx t(m_\tau(n_1), m_\tau(n_2), \ldots, m_\tau(n_k))$ where a function m_τ compares each value n_i with τ to form the bit-vector string. The coding of this feature vector is defined as:

$$u := \sum_{1 \le i \le \gamma \cdot \alpha} 2^{i-1} m_\tau(n_i) \qquad (1)$$

2) Speed-based Motion Patterns (SBPs): A second scheme calculates the amount of motion (speed) inside each bounded subregion r_i. This amount of motion is obtained by

accumulating the total displacement of each neighboring trajectory inside each subregion r_i. That is, for each trajectory τ_{t_k} occurring inside r_i, total displacement is defined as $d(\tau_{t_k})$ and speed inside each bounded region r_i as:

$$S(r_i) = \sum_{\tau_{t_k} \in r_i} d(\tau_{t_k}) \quad (2)$$

Such accumulation is compared with a given threshold using a function that evaluates if there is a minimal quantity of motion (MQM) and then, a feature vector will carry out the binary-codified values representing each atomic signature.

Then, any temporal interval Δt representation is herein achieved by considering a boundary C composed of a set of l-dimensional ToBPs points $\{u_1, u_2, \ldots, u_z\}$, code from any of two previously defined configurations for each motion point. This C representation locally code actions from occurrence kinematic patterns coded in ToBPs. The bit-string representation achieved a compact and robust model that allows a fast computation of action variations.

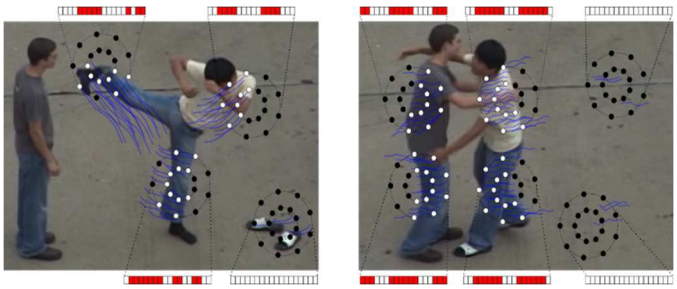

Fig. 3 Explained LBP-based scheme: delimited subregions r_i associated with motion density n_i are translated to a value of 1 (when $n_i \geq \tau$) or 0 alternatively. If the amount of trajectories is not acceptable, subregions are removed from the scheme.

C. Bag of Local Occurrence Binary Patterns

Bag-of-visual-words is a natural strategy to codify local patches, images, and videos and obtain mid-level representations. This strategy allows representing high-level concepts by computing occurrence histograms of local key descriptors regarding a previously learned dictionary. This mid-level representation scheme was herein adopted by using the set of motion-words as the local representation of ToBPs patterns, captured over a specific boundary interval C. From such ToBPs is computed a dictionary of more representative motion words, and then a mid-level occurrence histogram is computed for any interval of frames into the video sequence.

Firstly, It was recovered a ToBPs dictionary with more relevant motion patterns from a training set. For doing so, a K-means algorithm was herein implemented to compute principal K-centroids over the set of training ToBPs patterns, which correspond to the mean samples of corresponding k-clusters. Mainly, the set of k centroids $C = [c_1, c_2, \ldots, c_k] \in \Re^{l \times k}$ are extracted by an iterative function that tried to group each of the points in the respective k group, which represent the minimal distance, to the centroid $C(k)$, following the next objective function:

$$C(k) = min_{c_k} = \sum_{m=1}^{M} \quad \sum_{k=1}^{K} \|u(\tau)_m - c_k\|_2^2 \quad (3)$$

This dictionary coded the main ToBPs motion patterns across all activities and the whole set of training videos. Under this assumption, activities are defined as maximal occurrences of a particular set of centroids, flexible enough to obtain other patterns into the representation. Hence, for a particular frame (or set of frames Δt) it is computed the ToBPs motion patterns and is projected to the learned ToBPs dictionary. Each of the projected points is measured w.r.t. each centroid, to define the centroid's contribution in the representation concerning Euclidean distance s. The centroid i with minimal distance is counted in the respective bin of an occurrence histogram. Once all points are projected and measured, the resulting histogram is normalized, resulting in the current action representation for this interval of time.

This mid-level representation is robust to noisy scenarios, problems with occlusion and complex backgrounds containing a remaining dynamic of other activities. Besides, once the dictionary is trained, the occurrence histogram representation can operate at any interval of time, ranging from the simple frame-level representation to the complete video sequence representation.

D. Action Classification and Recognition of ToBPs Occurrence Histograms

The resulting ToBPs occurrence histograms are herein used to predict a particular action label, a relationship that it is modelled from a high-level machine learning algorithm. In such a case, we use a classical scheme of supervised machine learning strategies, being the occurrence histograms x_i the features and the corresponding action y_i the label to be predicted. In this case, the set of training video sequences, are also herein used to compute a set of (x_i, y_i) occurrence histograms that, together with labels allowed to fix boundaries into the machine learning strategies. We aimed to attain a favorable balance between accuracy and prediction time, and following such lines, the Random Forest classification (RF), and the Support Vector Machine (SVM) were considered to perform this analysis. Particularly, Random Forest classification (RF) reaches stable results outperforming much of the other machine learning alternatives. This RF strategy tried to model action representation space as a tree-based partition but learned from a set of decision tree classifiers (DT). Because the Gini-algorithm that optimize the branches of these trees can change abruptly from one representation to the other, the set of trees mitigate such fact by considering the confidence of the results as a voting approach of a group of DT classifiers. Also, the multi-class adaptation of Support Vector Machines (SVM) was evaluated using two different kernel configurations: the linear and Radial Basis Function (RBF) [28]. The linear kernel allows a fast performance on prediction task but assuming non-linear dependencies among ToBPs descriptors. In contrast, the RBF deal with non-linear representations, allowing a more accurate boundary definition theoretically but requiring a significant time on prediction to define the region of classes.

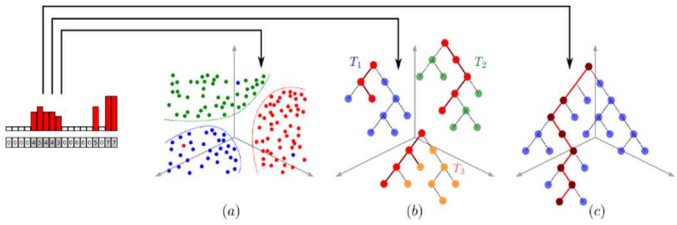

Fig. 4 Suggested classification strategies, namely: (a) Support Vector Machines, (b) Random Forest and (c) Decision Trees.

E. Data

An exhaustive evaluation was performed over four academic datasets of labeled video sequences to validate the proposed strategy. Three of these datasets are compiled for action recognition in general, while the last dataset is specialized in the task of gesture recognition. The used datasets are described as follows:

1) Twenty-five different subjects execute the KTH dataset [29], featuring six human actions. The experimental setup was considered as training (760 sequences), testing (863 sequences), and validation (768 sequences). The videos were recorded on open scenarios, with some variations in background and camera transformations. The resolution of each frame is 160x120 with a temporal resolution of 24 fps. On average, for these videos were captured 908 trajectories.

2) The Weizmann dataset [30] contains ten actions spread over 90 sequences with an evaluation using a leave-one-out strategy. These videos were captured with a common background, but the action dynamics are more challenging w.r.t to the previous dataset. Each of the videos has a spatial resolution of 180x144 with a temporal resolution of 50 fps. In average, for these videos were captured 457 trajectories.

3) The UT-Interaction dataset [31] includes surveillance sequences that were captured in uncontrolled scenarios and separated into two sets of 60 video sequences, containing 6 actions. A 10-fold leave-one-out method was performed, as suggested by the authors. This dataset has important variations in background and high variability on developed actions. One of the sets was recorded with camera jitters, and the background has other actions. A spatial resolution of 720x480 with 30 fps is featured.

4) The LS64 gesture recognition dataset was included in order to assess the robustness of the proposed method. The LSA64 dataset [32] features 64 gestures performed by ten subjects in 5 different scenarios. LSA64 features 3200 videos starting with a motionless segment. These videos were recorded with colored markers that better shape hands that develop the gestures. Nevertheless, the proposed approach only takes the dynamic information captured on dense motion trajectories. Handshape modeling is beyond the scope of this work. The video sequences have an original resolution of 1920x1080 with 60 fps. On average, for each video sequence was captured, 518 motion dense trajectories.

III. Results and Discussion

The entire components of the proposed strategy were herein widely evaluated on four public datasets. This evaluation's main objective was to analyze the performance of motion occurrence patterns in modeling dynamic actions. The proposed approach was evaluated in two different tasks: 1) for action classification, using the whole video sequence and 2) for action recognition using a frame-level representation and varying the temporal intervals of representation. Also, for each of these tasks, the two occurrence representations were evaluated, i.e., the ToBPs (trajectory occurrence binary patterns) and the SBPs (speed occurrence binary patterns). In the next subsections, the developed experiments are fully described.

A. Action Classification from Motion Occurrence-Based Descriptors

In this evaluation, a complete description of activities was captured by coding all possible ToBP and SBP descriptors, which are mapped to a dictionary to obtain a complete histogram occurrence descriptor of the actions. The evaluation is explained for each local descriptor, as follows:

1) Action ToBPs Classification: An evaluation of different parameters was carried out to adjust the proposed descriptor in other datasets to assess ToBPs motion descriptor's performance. Firstly, a spatial circular grid analysis was carried out to find the best configuration to describe the recorded actions. Tables 1 to 4 shows the performance obtained for the different action's classification datasets, namely KTH, Weizmann and UT (in both subsets), respectively. Initially, in Table 1, the KTH dataset's obtained results were reported by varying the tuple (γ,r) and measuring the performance concerning the accuracy in the classification task. As observed in Table 1, the best configuration was acquired with the tuple (γ=8,r=6), which suggests that it is required a major spatial domain to describe motion points. In contrast, such radial space only needed a six-pixel radial partition because of the sparse nature of trajectories to represent these activities.

TABLE I
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR KTH DATASET

γ	r (px)		
	6	**8**	**10**
4	87.40	89.11	89.11
5	89.57	88.54	88.20
6	89.46	88.77	87.28
7	87.74	86.71	86.48
8	**90.03**	87.51	87.06

Also, in Table 2 is reported the same analysis for Weizmann dataset. It should be noted that in this dataset, the best configuration was described in a tuple: (γ=5,r=10).

TABLE II
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR WEIZMANN DATASET

γ	r (px)		
	6	**8**	**10**
4	76.67	76.67	73.33
5	73.33	72.22	**78.88**
6	74.44	72.22	72.22
7	75.56	73.33	71.11
8	**78.88**	71.11	71.11

Contrary to KTH, this dataset requires an increased resolution of occurrence descriptor (increased radial splitting) and because the density of trajectories required only 5 concentric circles distributed spatially. Considering that this dataset's spatial resolution is slightly larger, the motion trajectories of actions are described more densely, and the result inside a circular grid is much more compact.

The same analysis was carried out on the surveillance UT-interaction dataset, which counts with two different subsets, shown in Table 3 and Table 4, respectively. In Table 3 is reported the different results obtained for different (γ,r) configurations. This dataset features activities with a

relatively static background, and therefore a major spatial circular grid configuration can help with local dynamic description. In this case, (γ=8, r=8) demonstrated the best configuration. In Table 4, it is reported the different (γ,r) configurations for sub-set two with a significant challenge on background representation. Subset two was recorded with camera jitters, and the background can contain other activities. In such a case, the best configuration was the tuple: (γ=4, r=6), showing that motion representation should be focused only on very near trajectories w.r.t. interest points. This fact is attributed to the major corrupted trajectories that correspond to the background.

A second analysis evaluated the best value for the Minimal Number of Trajectories (MNT) threshold τ, which allows a more stable action representation. This value is related to the transformation of circular grid occurrences into a bit-vector descriptor.

TABLE III
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR UT-INTERACTION (SET 1) DATASET

γ	r (px)		
	6	**8**	**10**
4	68.02	68.02	69.77
5	68.02	71.51	73.26
6	68.02	73.26	71.51
7	69.77	73.26	73.26
8	71.51	**75.00**	69.77

TABLE IV
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR UT-INTERACTION (SET 2) DATASET

γ	r (px)		
	6	**8**	**10**
4	**70.00**	61.03	66.41
5	64.62	57.44	64.62
6	61.03	**70.00**	61.03
7	66.41	62.82	57.44
8	66.41	64.62	57.44

The number of angles was fixed as α=9, while MNT values were ranged as τ={2,3,4,5,6,7,8}. In such case, with larger τ values, a major amount of trajectories n_i are admitted inside each region r_i .Figures 5 and 6 illustrate the proposed approach's performance regarding the MNT values in all different action classification datasets. Figure 5 reported the impact of τ parameter for the KTH (blue bars) and Weizmann (orange bars) datasets, respectively. For the KTH dataset, a competitive state-of-the-art result with τ values ranging from 2 to 8, and achieving an accuracy of 90%. The best configuration was achieved with τ=6 and a circular grid configuration of (γ=8, r=6). This dataset's periodic nature's recorded actions can justify this configuration that requires a spatial expansion and a lower τ to admit a rich trajectory representation. Figure 6 is reported the different accuracies obtained by changing the τ value in the Weizmann dataset. The best configuration was achieved with the same value of τ=6 and a spatial configuration of (γ=5, r=10). Using the same τ values for both datasets can result from a similar periodic dynamic of actions. Nevertheless, the Weizmann dataset has a major trajectory density and needs a more reduced space to assign importance to the trajectories.

Figure 6 is reported a similar analysis but for UT video sequences. These video-sequences are challenging, and the dynamic description corresponds to activities that involved several motions. The best configuration on UT-1 was also achieved for a τ value of 6. This parameter could be related to the static background of the three different datasets. In contrast, sequence UT-2, which report backgrounds with human interactions, require a more restricted threshold of trajectories. In such a case, the descriptor is built under reduced parameters in both: reduced space (γ=4,r=6) and τ=2.

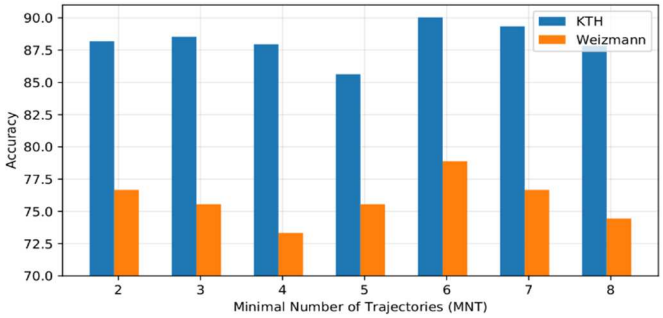

Fig. 5 The overall score associated with distinct values of MNT over the proposed circular grid for KTH and Weizmann datasets.

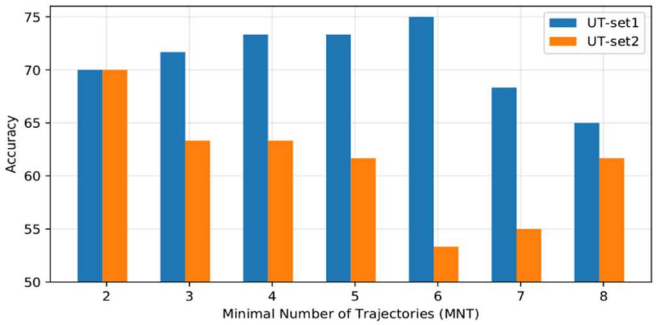

Fig. 6 The overall score associated with distinct values of MNT over the proposed circular grid for the UT-Interaction dataset.

Because the local proposed descriptors are dependent on trajectories, a length trajectory analysis was also carried out in this work. Length of trajectories was increased from 1 to 49 frames, allowing the major description of actions but resulting in noisy dynamic descriptions. As illustrated in Figure 7, For Weizmann, the dataset was achieved a local maximum with length on l=15 and l=37, being much more efficient than the l=15 for recognition applications. In this configuration, the resulting descriptor size could be estimated as (α×γ×15). A similar analysis was carried out in state-of-the-art for KTH, resulting in an ideal configuration of length l=15 [13]. Following this published work, the posterior analysis of KTH was fixed with this configuration.

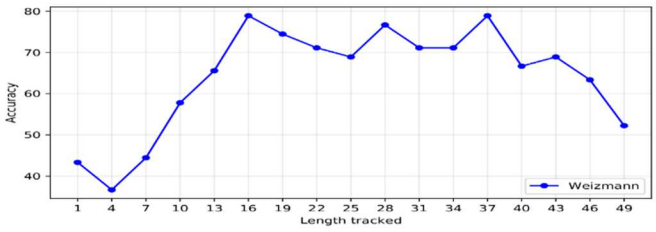

Fig. 7 The overall score associated with distinct lengths of tracking under the Weizmann dataset.

Figure 8 shows the performance of the proposed approach facing different length trajectory configurations under the UT-Interaction dataset. Regarding the UT-1 subset, the maximum accuracy value was achieved with an l=37 but unstable for online action recognition tasks. Best performance in both tasks was found with l=22 with a resultant descriptor size of (α×γ×22). For UT-2 subset, the best configuration was achieved on trajectories with length l=49 but with relevant increasing on descriptor size, and only an extra accuracy of 1%.

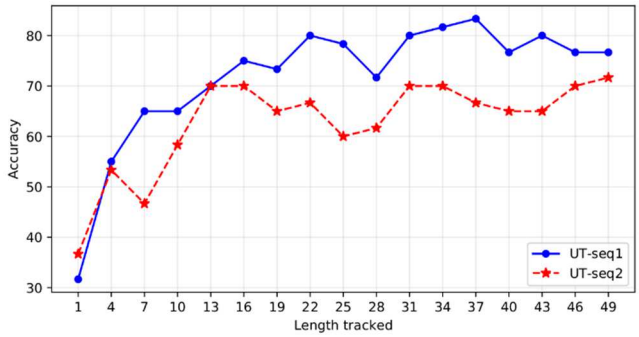

Fig. 8 The overall score associated with distinct lengths of tracking under UT-Interaction dataset.

Best configurations of proposed descriptors were used to evaluate different classifiers on action classification tasks. This approach evaluated the Random Forest (RF) and the Support Vector Machine (SVM) with two different kernel configurations. These classifiers have been successfully used on different tasks that required splitting complex training spaces with a reasonable computation time to obtain the learning models. An exhaustive grid search parameters for the classifiers were herein implemented to analyze the condition of the proposed occurrence descriptor's proper performance.

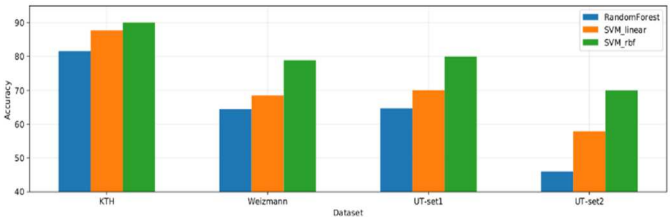

Fig. 9 The classification methods for KTH, Weizmann and UT-Interaction datasets. Major overall scores are depicted.

Best configuration was achieved with the SVM by using a radial basis kernel (RBF), which allows a non-linear partition of the feature space. As illustrated in Figure 9 in all datasets, this classifier outperforms the RF and linear SVM classifier. These results were obtained by training a dictionary of 400 centroids, and therefore the resulting global histogram occurrences are very compact to represent a complete video sequence. Nowadays, motion descriptors require histograms with thousands or millions of parameters to achieve a proper classification performance, even for datasets with relatively static backgrounds. These descriptors reported in the literature can deal with more challenging scenarios but are limited to online recognition.

2) Action SBPs classification: The second evaluation in the action classification task was performed by counting speed occurrence patterns counted in the circular grid, i.e., The SBPs patterns. Overall, this approximation looks more descriptive and for evaluated action classification datasets shows better accuracy scores. Additionally, a gesture recognition dataset evaluation was included for SBPs patterns, showing the robustness to represent different kinds of dynamics. This gesture recognition dataset, LSA64, was tested on ToBPs patterns but with negligible impact on the description. This fact could be explained because of the close representation of similar gestures that could be unrecognizable for this approach. As shown in the previous strategy, Support Vector Machines with an RBF kernel reported the best performance than other classification methods. Considering this fact, the same strategy is herein implemented to obtain a favorable tradeoff between accuracy and computation time. Hence all reported results for SBPs analysis were obtained using the SVM+RBF configuration. The parameter exploration and the analysis of obtained results for SBPs, are described as follows.

TABLE V
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR KTH DATASET

γ	r (px)		
	6	**8**	**10**
4	89.91	88.99	88.87
5	89.57	88.99	89.57
6	88.41	89.33	89.45
7	87.60	89.68	89.57
8	**90.73**	88.64	88.87

A first evaluation consists of exploring spatial circular grid resolution for SBPs local motion patterns. In Table 5 is reported the evaluation achieved for different circles and radii partitions. Like for ToBPs patterns, the best configuration was found in the tuple (γ=8, r=6). In this case, a complete spatial analysis is necessary to count related speed motion patterns and to obtain a significant statistical description on the circular grid descriptor.

In contrast, the resulting spatial analysis for Weizmann dataset obtained the best configuration (γ=4, r=8) (see Table 6). This fact could be associated with the spatial density of motion trajectories, that allows SBPs to focus only on near key-motion points. Because in this configuration, the kinematic occurrence is based on speed kinematic, a more restricted neighbourhood is sufficient to achieve a better score, concerning the configuration obtained for ToBPs (γ=5, r=10).

TABLE VI
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR WEIZMANN DATASET

γ	r (px)		
	6	**8**	**10**
4	75.55	**81.11**	76.66
5	76.66	74.44	77.77
6	75.55	77.77	74.44
7	76.66	76.66	74.44
8	77.77	74.44	75.55

The same spatial configuration analysis was carried out for both subsets of the UT-interaction dataset. Table 7 and Table 8 report the obtained results for UT in sequence 1 and UT in sequence 2, respectively. For the first subset (UT-1) the best configuration was achieved by the tuple (γ=7,r=8), which followed a similar pattern w.r.t ToBPs (γ=8,r=8)

configuration. In such a case, because of the relatively static background, a broader neighborhood was adequately explored. Regarding the configuration of UT-2, the best configuration was found with the tuple (γ=6,r=8), that admits a larger spatial exploration w.r.t. ToBps patterns. However, this configuration could admit a significant number of background trajectories and could lead to the classification's limited performance in this kind of dataset.

TABLE VII
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR UT-INTERACTION (SET 1) DATASET

γ	r (px)		
	6	**8**	**10**
4	63.33	63.33	68.33
5	68.33	73.33	70
6	68.33	68.33	65
7	68.33	**75**	73.33
8	63.33	73.33	68.33

TABLE VIII
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR UT-INTERACTION (SET 2) DATASET

γ	r (px)		
	6	**8**	**10**
4	51.66	55	55
5	56.66	60	56.66
6	58.33	**63.33**	55
7	56.66	50	55
8	60	58.33	56.66

SBPs patterns were also fully analyzed the LSA64 dataset to evaluate the capability to recognize the deaf-mute community (see Table 9).

TABLE IX
ACCURACY FOR DIFFERENT (γ,r) CONFIGURATIONS FOR LSA64 DATASET

γ	r (px)		
	14	**19**	**24**
4	88.28	**89.84**	88.43
5	89.53	89.68	88.75
6	88.12	87.03	87.18
7	88.43	88.59	88.12
8	89.68	88.90	89.53

In such a case, each of the gestures has a detailed dynamic description, but some gestures share very similar kinematic patterns. The best spatial configuration was achieved with a tuple (γ=4,r=19) that suggest a more extensive search space to code occurrence speed but requiring less concentric circles to obtain a proper motion descriptor. Nevertheless, similar scores are obtained for different configurations, which also justify the fact of motion density with local coherence

A second analysis carried out on SBPs motion descriptors was the minimal quantity of motion (MQM) that stand out the dynamic representation of actions, activities and gestures. The analogy with ToBPs is the MNT (τ value) that is required to byte-vectorize the descriptor. Figure 10 shows the performance of the proposed approach by changing the MQM parameter. An MQM value of 10 looks to be adequate for this SBPs configuration. It is also worth noting that each dataset exhibited a distinct MQM value due to the different kinematic features. Especially for LS64, there is an abrupt decay for larger MQM values, which is justified on the dynamic nature of these gestures requiring a major description in small intervals. For action classification datasets, larger values have long-lasting results but resulted ineffective for online recognition task, from which intermediate or partial descriptions are much more sensible during prediction.

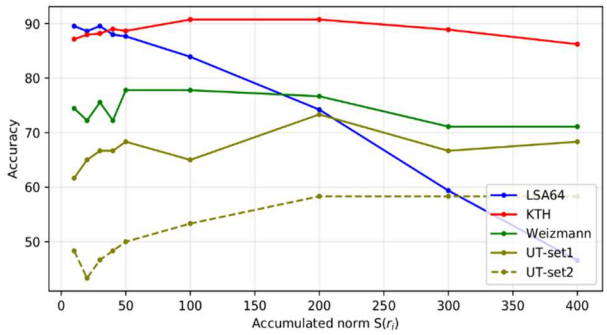

Fig. 10 Best configuration for accumulated norm (subregional speed) over: LSA64 dataset (blue) a score of 89.84% is obtained with MQM=10; KTH dataset (red) featured 90.7% with MQM=100; Weizmann dataset (green) exhibit 81.1% for an MQM=50 and UT set 1 (olive solid) and set 2 (olive dashed) with MQM=200 for 75% and 63.3% respectively.

Regarding the motion length, as analyzed on the ToBPs motion patterns, any length of tracking bigger than 20 frames will cause a significant increase in computation time of motion trajectories. Also, descriptor size is affected since the length of the track is one of the components, namely, the influence of l in (α×γ×l) number of elements. A fixed length of 16 was assumed, as suggested in the literature for different tasks.

In summary, except for UT-Interaction dataset, the SBPs strategy outperformed the scores of ToBPs, namely: 89.8% for LSA64 dataset (γ=4,r=19), 90.7% for KTH dataset (γ=8,r=6), and 81.1% for Weizmann dataset (γ=4,r=8). It is worth mentioning that LSA64 dataset showed a particularly unfavorable performance with ToBPs scheme, with scores around 1-3%. Such behavior is explained by the degree of detail that speed norm offers, in contrast to simple motion trajectories, given the kinematic nature of shorthand gesture motion. The proposed approach in any two configurations (ToBPs and SBPs) shows favorable and competitive accuracy results with a proper balance with computational time. This fact results in the implementation of real-time scenarios or complements the strategy with much more dense descriptions of the actions.

B. Action Recognition from Partial Sequences

Nowadays, many computer vision applications require an instant prediction about the high-level class that is happening. Despite that in the state-of-the-art, some very accurate approaches to perform action classification still miss an effective prediction on time. Following such facts, an online recognition evaluation was herein performed to explore the capabilities of the proposed approach to predict and update an action prediction at each frame t_i of the video sequence. In such a sense, the proposed approach in both configuration ToBPs and SBPs should predict actions, activities or gestures from partial representations or incomplete dynamics. The experiment was then designated by coding the motion descriptor using different incremental versions of the sequences and evaluating at each time the accuracy reported for all datasets.

An initial evaluation was performed by using local ToBPs motion configuration. Figure 11 reported the achieved results for KTH (blue line) and Weizmann (red line) datasets, using partial and online action recognition. For KTH, a competitive action prediction of 74% is achieved by using the 40% of the video sequences. Far better, for the Weizmann dataset, at only 20% of the total number of frames is achieved a significant prediction, w.r.t. using the entire sequence. The periodic nature of the actions can explain the proposed approach's capability to recover action over incomplete sequences, but with some mistakes corresponding to challenging capture and image transformations. Also, both datasets' actions result very interestingly that using only 10% of the video is available a proper prediction. These facts justify the use of this strategy in a scheme of real-time recognition.

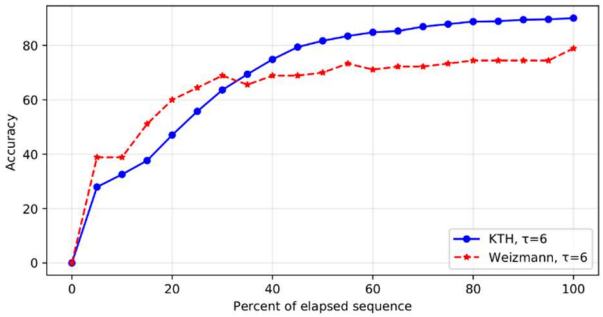

Fig. 11 Simulating an online application streaming over KTH dataset (blue): an acceptable score is obtained with just 30% of the sequences' total length. Weizmann dataset (red dashed): a sufficient accuracy is acquired with only 20% of the sequences' total length.

A similar analysis was carried out to UT-interaction, as illustrated in Figure 12. In this case, online recognition is reported independently for UT-set 1 (blue line) and UT-set 2 (red line). Despite that this dataset is dedicated to describing activities (composed of small motion actions), set 1 exposes promising results, requiring only 15% of the sequence to obtain a similar score to the average score achieved along the sequence. In this case, only the last part of the action, that defined the signature of the activity changed and incremented the proposed representation's performance. Additionally, different length trajectories l={15,22,37} were evaluated on this partial online recognition, showing that l=37 obtained a better performance for the overall classification but an inefficient representation for partial sequence recognition. As expected, the accuracy obtained for UT-2 is lower because of the complex background conditions.

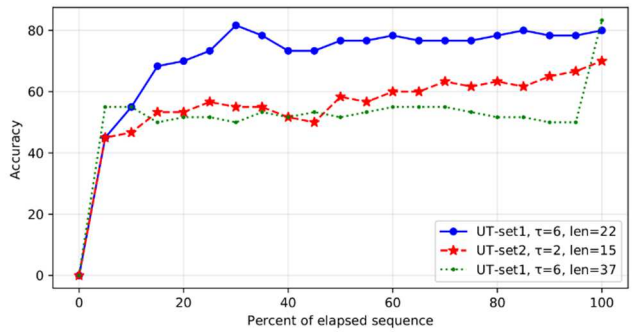

Fig. 12 Simulating an online application streaming over UT-Interaction dataset, acceptable results are acquired with only 15% of the total length of sequences for set 1.

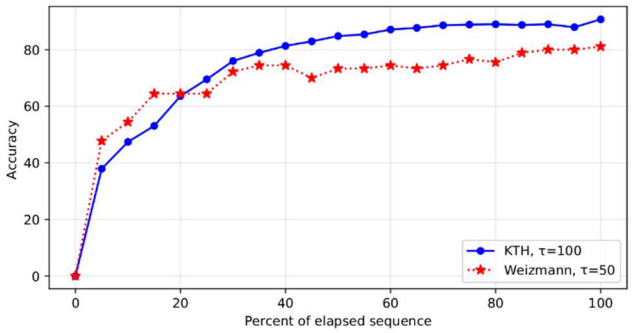

Fig. 13 Performance simulation for an online application over (blue) KTH dataset: our method achieves promising results with just 20% of the total number of frames. Weizmann dataset (red): promising results are obtained with just 15% of the total number of frames.

Regarding SBPs motion descriptors, an extensive set of experiments were carried out to obtain action and gesture recognition from partial cumulative sequences, simulating a streaming application. As for KTH dataset, a right action prediction was achieved with just 20% of the total sequence. Similarly, Weizmann dataset features an acceptable forecast with only 15% of the sequences, as shown in Figure 13. In such two cases, the SBP motion patterns result much more descriptive that ToBPs descriptors, with a more coherent motion description using a much more partial and sparse representation. This result could be justified by using the speed kinematic information as input to build bit-vector descriptors. In such a sense, activities like running and jogging could be better differentiated. Also, both datasets showed an incremental accuracy on the representation of the action, a fact associated with recorded activities' periodicity.

Figure 14 depicts the computed results for online action recognition for the UT-interaction dataset using the SBPs patterns. In this case, the activity prediction slightly decreased w.r.t. ToBPs, a fact associated with the initial exploration of parameters, from which the circular grid covers more trajectories. Much of the corrupted background trajectories can alter the local motion description of key-motion points in such a case. Additionally, the similar speed of actions featured on it makes the SBPs strategy less discriminant in action classification. Despite the comparison with ToBPs, the online recognition showed stable performance, being in general incremental accuracy after 15 frames. This start point of 15 frames is related to an original length of motion trajectories.

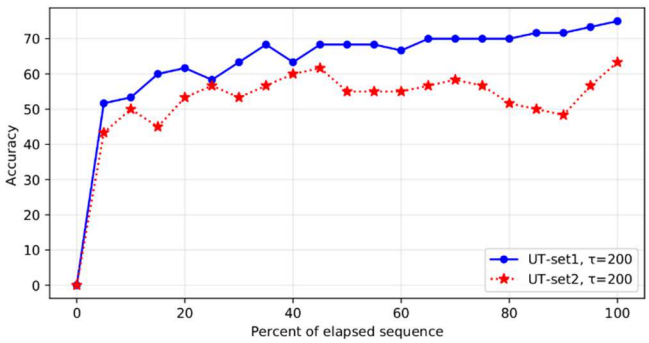

Fig. 14 Performance simulation for an online application over UT-Interaction dataset: sequence 1 (blue) achieved promising results, with only 15% of the total number of frames. As for sequence 2 (red), such score is obtained with 40% of the total number of frames.

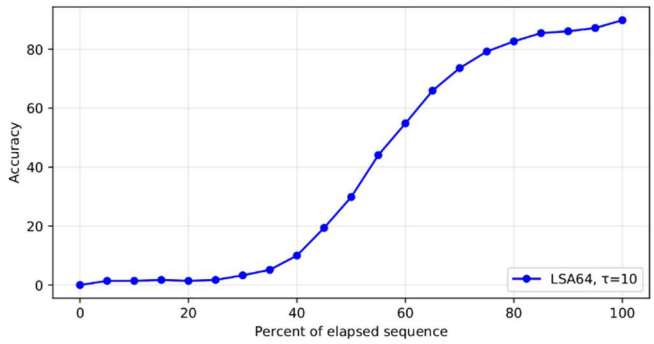

Fig. 15 Performance simulation for an online application over LSA64 dataset: our method achieves promising results with 60% of the total number of frames. The existence of motionless segments explains this at the beginning of each sequence.

Finally, online recognition performance was also evaluated on the LSA64 dataset, which could be used to get ahead word prediction and obtain more fluent translations. Figure 15 report the results associated with online recognition in this dataset. Notably, there is no gesture information in the first 40% of all sequences. For this reason, the first part of the plot shows accuracies close to zero. Once the gesture start (approximately after 40% of the elapsed sequence), there is a significant increase in accuracy that demonstrates the robustness of the proposed approach requiring less than the middle of the sequence to obtain a stable gesture prediction. Despite the close dynamic description among some recorded gestures, the SBPs motion patterns achieve an accurate description and differentiation using partial gesture codification.

IV. Conclusion

This strategy introduced a compact descriptor capable of predicting human actions and gestures in video streaming using a set of statistics obtained from neighboring motion trajectories. A fixed bounded structure allows quantifying occurrences of said trajectories, and then a binary representation is acquired. An extension of this scheme accumulated the speed of trajectories (SBPs) inside bounded subregions, resulting in a motion description in terms of trajectory length. Both schemes featured 400 scalar values in a mid-level representation and an averaged size of 9x6x16 for online frame-level presentation. Our method was extensively validated through three action recognition datasets and a hand gesture recognition dataset. An overall improvement is reported for the SBPs scheme, which is explained by the richer kinematic information behind the accumulated speed. Also, modifying the fixed circular grid's external boundary allowed to capture more details associated with trajectories on other grids, that is, a better perspective of the amount of motion around the fixed grid. The proposed method demonstrated competitive results over three action recognition datasets and a gesture recognition dataset, proving the robustness on different scenarios of human-related motion. Further research will be developed adopting different architectures, more challenging datasets, and additional kinematic primitives.

Acknowledgment

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] Mahmood, A., Al-Maadeed, S. "Action recognition in poor-quality spectator crowd videos using head distribution-based person segmentation" Machine Vision and Applications, 30(6), pp. 1083-1096 (2019).

[2] Sultani, W., Shah, M. "Automatic action annotation in weakly labeled videos". Computer Vision and Image Understanding, 161, pp. 77-86 (2017).

[3] Saravanan, D. "Efficient Video Indexing and Retrieval Using Hierarchical Clustering Technique". In Proceedings of the Second International Conference on Computational Intelligence and Informatics. pp. 1-8. Springer (2018).

[4] Kong, L., Huang, D., Qin, J., Wang, Y. "A joint framework for athlete tracking and action recognition in sports videos". IEEE Transactions on Circuits and Systems for Video Technology, 30(2), pp. 532-548 (2019).

[5] Narayana, P., Beveridge, R., Draper, B. A. "Gesture recognition: Focus on the hands". In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5235-5244 (2018).

[6] G. Zhu, L. Zhang, P. Shen and J. Song, "An online continuous human action recognition algorithm based on the kinect sensor" Sensors, Multidisciplinary Digital Publishing Institute., vol. 16, 2016.

[7] Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., Chen, D. S. "A comprehensive survey of vision-based human action recognition methods". Sensors, 19(5), p. 1005 (2019).

[8] V. Veeriah, N. Zhuang, and Qi, Guo-Jun, "Differential recurrent neural networks for action recognition" Proceedings of the IEEE international conference on computer vision, IEEE., pp. 4041-4049, 2015.

[9] S. Al-Ali, M. Milanova, H. Al-Rizzo and V. Fox, "Human action recognition: contour-based and Silhouette-based approaches" Computer Vision in Control Systems-2, Springer., pp. 11-47, 2015.

[10] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition" Lecture notes in computer science, Springer., vol. 3667, pp. 91-103, 2006.

[11] W. Moreno, G. Garzon and F. Martínez, "Frame-Level Covariance Descriptor for Action Recognition" Colombian Conference on Computing, Springer., pp. 276-290, 2018.

[12] J. Rodriguez and F. Martínez, "A Kinematic Gesture Representation Based on Shape Difference VLAD for Sign Language Recognition" International Conference on Computer Vision and Graphics, Springer., pp. 438-449, 2018.

[13] H. Wang, A. Klaser, C. Schmid and C. Liu "Action recognition by dense trajectories" Computer Vision and Pattern Recognition (CVPR), IEEE., pp. 3169-3176, 2011.

[14] H. Wang and C. Schmid "Action recognition with improved trajectories" Proceedings of the IEEE international conference on computer vision, IEEE., pp. 3551-3558, 2013.

[15] H. Wang, A. Klaser, C. Schmid and C. Liu "Dense trajectories and motion boundary descriptors for action recognition" International journal of computer vision, Springer US., vol. 103, pp. 60-79, 2013.

[16] F. Caba, V. Escorcia, B. Ghanem and J. Niebles "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE., vol. 103, pp. 961-970, 2015.

[17] H. Rahmani, A. Mian and M. Shah "Learning a deep model for human action recognition from novel viewpoints" IEEE transactions on pattern analysis and machine intelligence, IEEE., vol. 40, pp. 667-681, 2018.

[18] G. Varol, I. Laptev and C. Schmid "Long-term temporal convolutions for action recognition" IEEE transactions on pattern analysis and machine intelligence, IEEE., vol. 40, pp. 1510-1517, 2017.

[19] Q. Ke, M. Fritz and B. Schiele "Time-Conditioned Action Anticipation in One Shot" Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, IEEE., pp. 9925-9934, 2019.
[20] G. Garzon and F. Martínez "Online action recognition from trajectory occurrence binary patterns (ToBPs)" Proceedings of the International Conference on Advances in Emerging Trends and Technologies, Springer., 2019.
[21] T. Ojala, M. Pietikainen and D. Harwood "A comparative study of texture measures with classification based on featured distributions" Pattern recognition, Elsevier., vol. 29, pp. 51-59, 1996.
[22] T. Bouwmans, C. Silva, C. Marghes, M. Zitouni, S. Mohammed, H. Bhaskar and C. Frelicot "On the role and the importance of features for background modeling and foreground detection" Computer Science Review, Elsevier., vol. 28, pp. 26-91, 2018.
[23] L. Nanni, S. Brahnam and A. Lumini "Local ternary patterns from three orthogonal planes for human action classification" Expert Systems with Applications, Elsevier., vol. 38, pp. 5125-5128, 2011.
[24] L. Yeffet and L. Wolf "Local trinary patterns for human action recognition" 12th International Conference on Computer Vision, IEEE., pp. 492-497, 2009.
[25] T. Nguyen, A. Manzanera, N. Vu and M. Garrigues "Revisiting lbp-based texture models for human action recognition" Iberoamerican Congress on Pattern Recognition, Springer., pp. 286-293, 2013.
[26] R. Anwer, F. Khan, J. van de Weijer, M. Molinier and J. Laaksonen "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification" ISPRS journal of photogrammetry and remote sensing, Elsevier., vol. 138, pp. 74-85, 2018.
[27] R. Muhammad Anwer, F. Khan, J. van de Weijer and J. Laaksonen "Tex-nets: Binary patterns encoded convolutional neural networks for texture recognition" Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ACM., pp. 125-132, 2017.
[28] C. Chang and C. Lin "LIBSVM: a library for support vector machines" ACM transactions on intelligent systems and technology (TIST), ACM., vol. 2, p. 27, 2011.
[29] C. Schuldt, I. Laptev and B. Caputo "Recognizing human actions: a local SVM approach" Proceedings of the 17th International Conference on Pattern Recognition, IEEE., vol. 3, pp. 32-36, 2004.
[30] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri "Actions as space-time shapes" IEEE transactions on pattern analysis and machine intelligence, IEEE., vol. 29, pp. 2247-2253, 2007.
[31] M. Ryoo and J. Aggarwal "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities" 12th international conference on Computer vision, IEEE., pp. 1593-1600, 2009.
[32] F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini and A. Rosete "LSA64: an Argentinian sign language dataset" XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016), 2016.