A spatio-temporal multi-scale motion descriptor from a spatially-constrained decomposition for online action recognition

Fabio Martínez[1,2,3], Antoine Manzanera[2], Eduardo Romero[1]

[1]CIM@LAB, Universidad Nacional de Colombia, Bogotá, Colombia
[2]U2IS/Robotics-Vision, ENSTA-ParisTech, Université de Paris-Saclay, France
[3]Escuela de Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, Colombia
[*]edromero@unal.edu.co, famarcar@uis.edu.co, antoine.manzanera@ensta-paristech.fr

Abstract: *This work presents a spatio-temporal motion descriptor that is computed from a spatially-constrained decomposition and applied to online classification and recognition of human activities.* The method starts by computing a multi-scale dense optical flow that provides instantaneous velocity information for every pixel without explicit spatial regularization. Potential human actions are detected at each frame as spatially consistent moving regions and marked as Regions of Interest (RoIs). Each of these RoIs is then sequentially partitioned to obtain a spatial representation of small overlapped subregions with different sizes. Each of these region parts is characterized by a set of flow orientation histograms. A particular RoI is then described along the time by a set of recursively calculated statistics, that collect information from the temporal history of orientation histograms, to form the action descriptor. At any time, the whole descriptor can be extracted and labelled by a previously trained support vector machine. The method was evaluated using three different public datasets: (1) *The VISOR dataset was used for two purposes: first, for global classification of short sequences containing individual actions, a task for which the method reached an average accuracy of* 95% *(sequence rate)*. Also, this dataset was used for recognition of multiple actions in long sequences, achieving an average per-frame accuracy of 92.3%. (2) the KTH dataset was used for global classification of activities and (3) the UT-datasets were used for evaluating the recognition task, obtaining an average accuracy of 80% (frame rate).

1. Introduction

Action recognition is a very active research domain with a large variety of potential applications: human computer interaction, biometric, health care assistance or surveillance, among others. This task, aimed to automatically segment and identify human activities in video sequences, is particularly difficult because of the high variations in geometry, scale and appearance. Such variations are namely present under not controlled illumination or occlusion conditions. Moreover, characterization of human dynamic introduces additional challenges, specifically: (1) activities can share similar gestures or motion primitives, for instance, leaving an object or get into a car, (2) interactions with other humans or objects may occur, occluding the capture or producing very different dynamic patterns and (3) many of the proposed descriptors require a complete description of the activity along the video sequence to perform the recognition, this limitation could be critical in many real scenarios. *Emerging recognition applications include the detection of actions in challenging scenarios*, *as well as detection of group activities, search for salient events within*

particular time intervals and object detection from a single view. These new tasks demand flexible descriptors that achieve an appropriate trade-off between accuracy and computation time for real applications. For instance, some new applications require to capture local and salient regions with their associated semantics. Many other tasks require a competitive detection of salient regions but using efficient and soft descriptors that allow decision in real time. Comprehensive surveys of the different proposed approaches and applications can be found in [1, 2, 3, 4]

Under certain controlled conditions, the actions can be represented as a continuous progression of the body geometry, whereby temporal variations of the human shape are associated with specific activities [5]. *These approaches are however limited in outdoors or open scenarios where no illumination control is possible. The motion captured in such conditions is in general contaminated since the dynamic quantification depends on a proper computation of silhouettes. Furthermore, these approaches may fail by the under-segmentation or occlusion problems [4].*

Additional methods for action recognition are based on the computation of local features along the video sequence, that in general, overcome problems relative to the occlusion and some geometrical variations. Once the spatio-temporal features are computed, several statistics are applied to code the action description, for instance by computing bag-of-features to analyse the occurrence of patches [9, 7], or by using rule based methods, where the activity detection is represented as a maximum-weight connected sub-graph [6]. *These local features are however dependent on the object appearance, the recording conditions and a large number of patches. In such approaches, the set of patches used to represent the sequence of video ignore temporal correlations. However, some approaches do compute space-time information from particular motion patterns, aiming to follow the objects in motion along the sequence [10]. These approaches characterize the space-time information by computing features such as HOF (Histograms of Optical Flow), MBH (Motion Boundary Histograms) and HOG (Histograms of Oriented Gradients), which together represent the activities under a bag-of-feature framework. In this case, the features computed form a dictionary which is used to compute the signature of the video. These space-time features nevertheless are computed for fixed intervals of time, a condition too restrictive for on-line applications containing a wide variety of actions. Moreover, variability can even be larger if one considers computed features mainly dependent on their particular appearance under variable illumination conditions in open scenarios where typically activities are developed.*

Motion description from dense or sparse optical flow primitives, has been also widely used to characterize and recognize individual and interactive activities. Such descriptors have been popular since they are relatively independent of the visual appearance and allow to capture complex patterns of human actions. For instance, in [11] human gestures were recognized by applying histograms of oriented optical Flow (HOOF), made invariant to vertical symmetry. This approach however misses local details that might define a particular activity, for instance the motion relationship between the limbs. Likewise, Riemenschneider *et. al* [8] compute different dynamic relationships from an optical flow combined with a bag-of-features model that aims to determine the occurrence of motion patterns. These methods in general highlight the most frequent dynamic patterns but lose temporal and spatial probabilistic feature distributions. Also, block-based histograms of optical flow have been proposed to partially preserve the spatial distribution of motion patterns, which in turn can be combined with local contour orientations [12]. This method can distinguish simple periodic actions, but the motion characterization may be dependent on the spatial grid configuration. Other approaches like [13] and [14] use polar space representations to code regionally the optical flow and to characterize activities. Besides, in the work proposed by Michalis *et. al* [18] a set of time series features are computed from the optical flow to represent the activ-

ities. These series are clustered using Gaussian mixture modeling (GMM) and then a canonical time warping allows the comparison between the training and the test samples. A main limitation of these methods is that they compute their descriptors on entire sequences, thus do not explicitly provide on-line recognition capabilities. Likewise, the local features coded in their descriptors are in most cases appearance dependent.

Currently, machine learning approaches based on convolutional deep models have been applied to construct features used in action classification, achieving high accuracy from large training sets [15], [16], [17]. Nevertheless, these methods are computationally expensive and present an intrinsic time-delay that together constitute a critical limitation in real action recognition applications. Additionally, these methods are dependent on a bias parameter that has to be learned along with the spatio-temporal kernel weight parameters, a task for which the whole video sequence is required. *For instance, in [37], different temporal scales were proposed to recognize activities in single perspective videos, as input to a CNN architecture. Such work reports high recognition rates but at a high computational cost because of the resultant high dimensional descriptors that must be learned from video sequences.*

The main contribution of this work is a spatio-temporal multi-scale descriptor composed of a set of recursive statistics that collect the RoI temporal information in orientation histograms. Each of these RoIs is spatially divided into several overlapping sub-regions with different sizes and each of these subregions is in turn temporally characterized by computing a flow orientation histogram, weighted by the norm of the velocity. A complete dynamic description is then achieved by recursively computing statistics of the histograms spanning different temporal intervals. Such descriptor is capable of determining locations of human actions with potential interest. The resultant descriptor is used as input to a trained SVM classifier. Evaluation is performed using two video-surveillance based human action recognition datasets, made of real actions recorded with static (or quasi static) cameras. The performance of the proposed approach proved to be competitive with respect to the state-of-the-art. This paper is organized as follows: Section 2 introduces the proposed method, section 3 presents results and the evaluation of the method, and finally section 4 presents a discussion and concludes with possible future works.

2. The Proposed Approach

Visual systems are naturally entailed with the ability of optimally detecting, recognizing and interpreting visual information, in many cluttered scenarios, using practically the same evolved mechanism [19]. In general, the visual system explores an overfragmented environment and constructs a valid world representation by recognizing a relevant motion when there exists a sort of temporal coherence during a time interval. Hence, a major challenge at analysing any human action is then the optimal duration, during which such analysis should be carried out. This interval is obviously dependent on the action complexity, for instance a walking action may be characterized during very short periods while composed actions, such as getting into a car, may require longer times. This work introduces a novel strategy that integrates several temporal scales of an overlapped representation of a Region of Interest, which in due turn is determined from the optical flow field. The method starts by computing a dense optical flow, using a local jet feature approach, from which a spatial coherence during a certain time helps to discover potential RoIs. Each RoI is then divided to obtain an overlapped representation that allows to integrate local and global dynamic information. Any of these subregions is basically characterized by a set of statistics, computed from the history of the orientation histograms, a complex dynamic structure that, at each time, stores the

information flow between consecutive frames.

2.1. Dense optical flow estimation using local jet features

The computation of an apparent velocity flow field has been successfully applied to recover motion patterns from object orientation [2],[3]. The herein proposed strategy characterizes the motion by computing a velocity flow field that is used, first, to localise potential actions in the scene, and second to describe such actions using space $\times$ time statistics of velocities. Any dense or semi-dense optical flow algorithm may be used for such task. In this work we used the nearest neighbour search in the local jet feature space (see [20] for more details). It consists in projecting every pixel to the feature space made of the spatial derivatives of different orders, computed at several scales (the local jet). For each frame t and every pixel $\mathbf{x}$, the apparent velocity vector $\mathbf{V}_t(\mathbf{x})$ is estimated by searching the most similar feature in the local jet space at the frame $t-1$. *In the literature, most dense optical flow algorithms obtain smooth dense velocity fields by introducing different orders of spatial restrictions that limit global motion computations. In contrast, the method herein used provides a dense optical flow field without explicit spatial regularization and an implicit multi-scale estimation by using regional spatial characteristics. In this sense, the proposes approach provides a better motion estimation at a global level.* In our experiments, we used 5 scales, with $\sigma_{n+1} = 2\sigma_n$, and three first order derivatives, resulting in a pixel-level descriptor vector of dimension 15 (Figure 1, *first column).*

2.2. Localisation of potential actions and overlapped RoIs representation

The second step consists in localising the potential actions by extracting rectangular Regions of Interest (RoIs) containing a set of the closest pixels with significant motion. The pixels whose velocity norm is above a certain threshold τ_s are spatially aggregated using a morphological closing with a disk of radius R_c. Then, multiple moving sub-regions can pop out the scene $\{a_i\}_{i=0}^{n}$. In a particular frame, the regions of connected components with area smaller than A_m are discarded, and the rest are grouped if their distance D_m is less than τ_t. Otherwise, they will be considered independent regions of interest that will be processed as independent actions. For consecutive frames, such spatial regions are also indexed as the same region if the distance D_m is less than the threshold τ_t. Such temporal association of the RoI is fundamental for computing the time statistics (see Sec. 2.4). The distance $D_m = \|a_i - a_j\|$ is simply defined as an Euclidean metric between the centroid coordinates of any two connected components a_i, a_j, in both spatial and temporal axes. (see Sec.2.4)

In video sequences with multiple targets, multiple regions of interest are identified if the distance between these regions is larger than the pre-established distance threshold τ_t. If there are different actors in the scene and the between-object distance D_m is small, then pair-wise actions are grouped and the complete motion is considered as human interaction. For a typical sequence of human interaction, initially independent actions can be considered because of the established distance threshold τ_t, but with the time such RoIs can be grouped as a unique interaction. For each detected RoI, an independent motion descriptor is computed and then mapped to the support vector machine as described hereafter. The threshold for grouping the spatiotemporal RoIs was set to $\tau_t = \lceil 0.1 \cdot (2 \cdot (W + H)) \rceil$, while the grouping of subregions was set to $t_s = \lceil 0.1 A_m \rceil$. Here W and H correspond to frame width and height, while A_m corresponds to the area of region m.

Afterwards, each of the selected RoIs is partitioned as illustrated in Figure 1, obtaining a set of overlapped subregions with different sizes. The total number of subregions for n layers (splits) is:

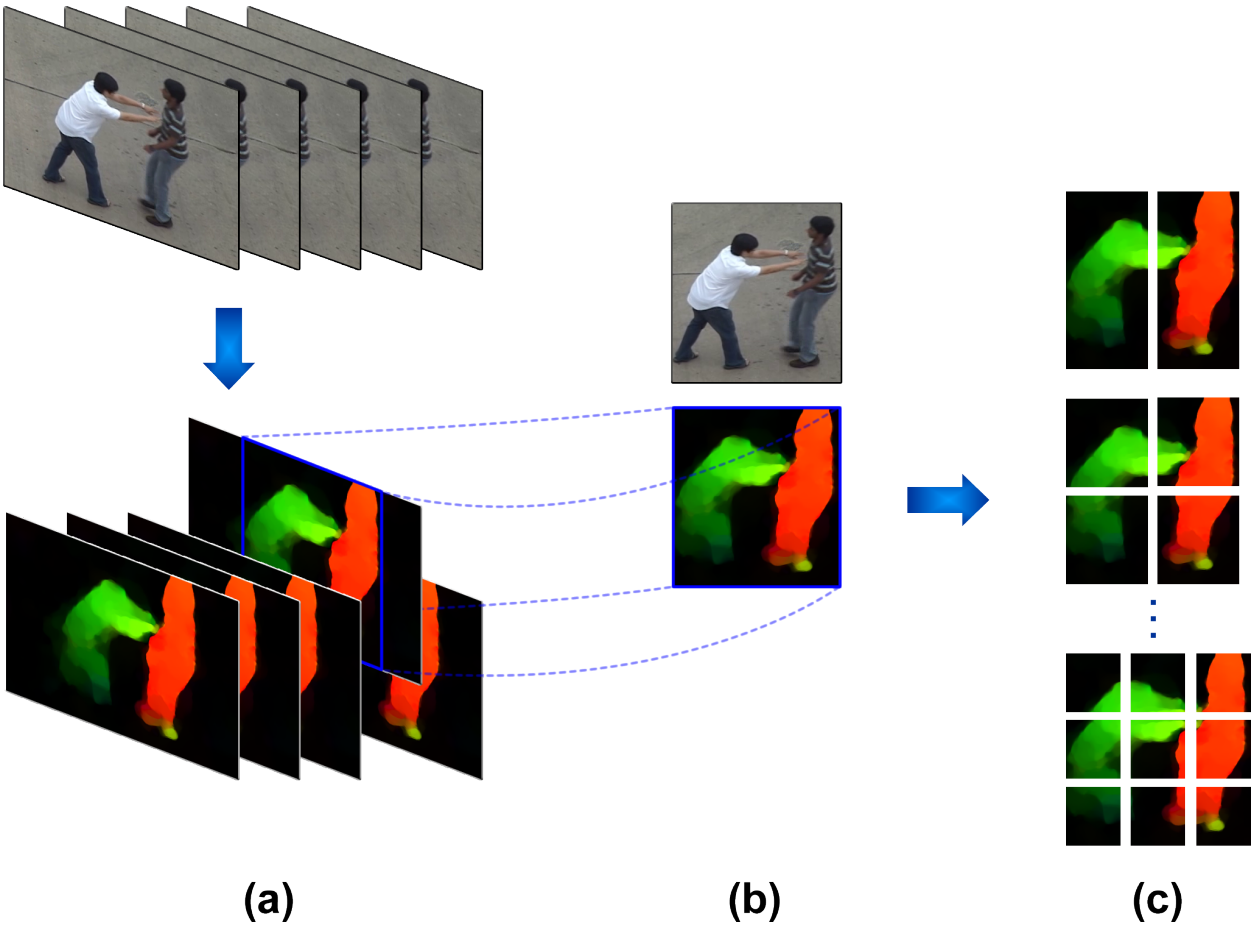


Fig. 1. *Motion RoI segmentation and the spatially-constrained decomposition that together form a spatio-temporal representation. In (a) is computed the dense optical flow based on a multi-scale local jet representation. Such flow field characterization allows to bound the potential activities as Region of Interest (b). Each of these RoIs is then sequentially partitioned, up to obtain a spatial overlapped subregion representation (c).*

$L_n = 2 + \sum_{L=2}^{n} L^2$*, where the first layer* L_1 *vertically splits the RoI into two parts, and then each layer is proportionally split into the same number of divisions for the two axes. For instance, the total number of subregions using four layers* (L_4) *is* 31*. This spatial RoI representation allows us to represent the RoI as a set of different layers which at each time captures finer regions, and therefore more localized dynamic patterns. Perceptually inspired [21], such representation captures global and local appearance motion patterns, i.e., high statistical dependence of these subregions. Thus, the salient dynamic emerges at the different scales, making the computed statistics store the dependency of the different patterns at the several scales and times. Thus, the spatially located salient dynamics is highlighted by the different scales, which produce a redundant representation of the motion patterns. In the proposed approach there are not explicit metrics to capture the correlation among the regions captured at different layers. However, since the RoI is represented using different layers, the regions with motion will be predominant in the distribution of the motion descriptor.*

2.3. Histogram of Velocity Orientations

For each of these RoI subregions, a per-frame temporal descriptor is built up using the distribution of the instantaneous motion orientations. For a non-null flow vector $\mathbf{V}$ let $\Omega(\mathbf{V})$ be its orientation, quantised to N bins. As in the Histogram of Oriented Gradients, HOG [22], a per-frame motion orientation histogram is computed as the occurrence of flow vectors with similar orientations, weighted by their norms. A dominant direction may then be the result of many vectors or few vectors with large norms; such histogram reads as:

$$H_t(\phi) = \frac{\sum_{\{\mathbf{x} \in RoI; \mathbf{V}_t(\mathbf{x}) \neq 0; \Omega(\mathbf{V}_t(\mathbf{x})) = \phi\}} \|\mathbf{V}_t(\mathbf{x})\|}{\sum_{\{\mathbf{x} \in RoI\}} \|\mathbf{V}_t(\mathbf{x})\|}$$

where $\phi \in \{\phi_0 \dots \phi_{N-1}\}$, N being the number of orientations, herein set to 64 (see Figure 2, first row).

2.4. Multi-scale motion descriptor

Visual systems are in general capable of recognizing activities by somehow integrating simple primitives during different intervals of time [19]. It is well known that most of the retina cells respond to transients and basic information of edges. Overall, this information is organized in terms of space and time by cells that are topographically connected to the visual field and that are triggered by variable time stimuli. This characterization allows, among others, to filter specific noise out, to analyse complex dynamics and to recognize objects during variable intervals of information. *In action recognition applications, human activities may be thought of as the succession of atomic motions / gestures that can be described by simpler dynamics. Actions involving periodic motions like "walking" or "boxing" can be described efficiently since all their atomic gestures can be characterized during limited temporal intervals. Activities like "get into a car" or "leave an object" are more complex and typically involve several periodic and aperiodic parts.* In such a case, the computation of global features over the flow trajectory may result insufficient and, in many times, with an important loss of relevant information, as for instance the transition between simple actions. Hence, a successful temporal descriptor should combine the analysis spanning different temporal windows.

The new descriptor herein introduced is designed to combine information from different time periods. For doing so, a set of relevant motion features are computed during variable time intervals (temporal scales) using recursive estimations as the cumulated statistics from the orientation histograms. *Firstly, the temporal mean and variance are estimated using the recursive exponential filters, where $H_t(\phi_j)$ is the j_{th} histogram bin, computed at time t and $\alpha \in [0, 1]$ is a decay parameter relative to the time depth. For each histogram bin, two additional statistics are recursively estimated: the temporal maximum and minimum, using forgetting morphological operators [23]. The computation of the recursive statistics is illustrated in Algorithm 1.*

The computed non-linear features complement the dynamic information estimated by the mean and standard deviation. The proposed descriptor combines a set of features recursively computed to cope with different complex and periodic human activities. Furthermore, the motion descriptor shows interesting properties: (1) little sensitivity to the impulse noise because of the forgetting term, corresponding to the exponentially decreasing weights attached to the past values; (2) periodic and composed motions are characterized by a set of global features computed at several time

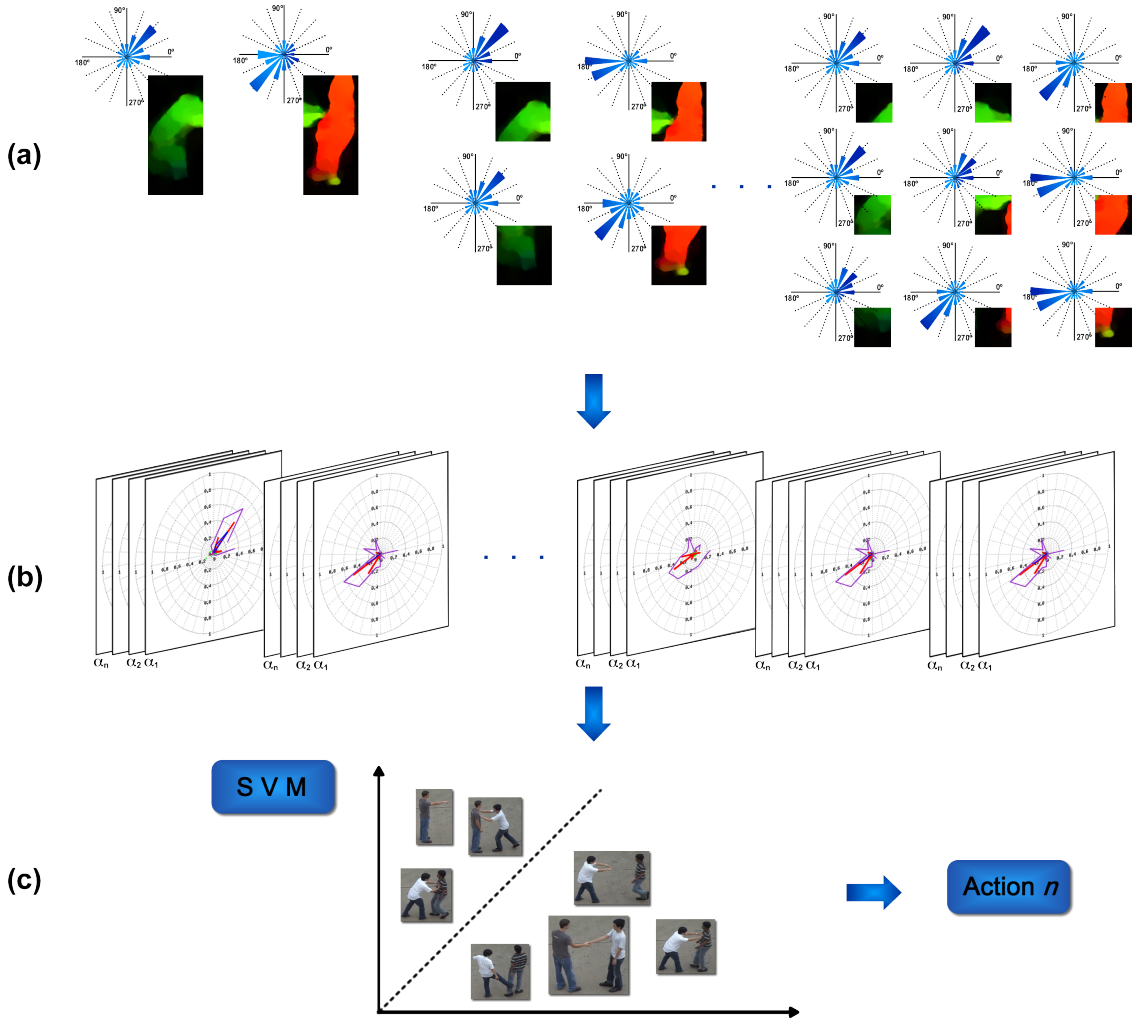


Fig. 2. *Computation of the multiscale motion description from a spatially-constrained decomposed RoI. In (a) the candidate action bounded in a RoI is sequentially partitioned and for each sub-region a motion orientation histogram is computed. In (b), several temporal recursive statistics are illustrated, they are computed per bin, using different α parameters that achieve variable time intervals for the analysis. This figure illustrates the principles of the different steps but does not necessarily display the features corresponding to the real data. Finally, in (c) each multiscale-motion descriptor is mapped over a previously trained support vector machine to predict the action.*

scales, including both recent and old dynamic information; (3) the recursive computation achieves an efficient use of the memory. Provided that some activities are commonly composed of different simpler actions, the statistics and the multiscale estimation facilitate discrimination of periodic and aperiodic motions.

2.5. Frame activity recognition

Recursively computed statistics estimate different characteristics for several time scales. This particular attribute of the proposed descriptor is a powerful tool for representing actions of interest in incomplete videos or streaming sequences. In practice, at each frame, the motion descriptor is

Algorithm 1 Recursive computation of statistics that temporally integrate the motion histograms computed at each sub-region.

Initialization:
for each bin ϕ_j of the orientation **do**
$\mu_0(\phi_j) = m_0(\phi_j) = M_0(\phi_j) = H_0(\phi_j)$
$v_0(\phi_j) = 0$
end for

for each time t **do**
for each sub-region of interest **do**
calculate Histogram $H_t(\phi)$
for each bin j of the $H_t(\phi)$ Histogram **do**
1. Recursive mean:
$\mu_t(\phi_j) = \mu_{t-1}(\phi_j) + \alpha(H_t(\phi_j) - \mu_{t-1}(\phi_j))$
2. Recursive variance:
$v_t(\phi_j) = v_{t-1}(\phi_j) + \alpha\left((H_t(\phi_j) - \mu_t(\phi_j))^2 - v_{t-1}(\phi_j)\right)$
3. Forgetting Max:
$M_t(\phi_j) = \alpha H_t(\phi_j) + (1-\alpha)\max(H_t(\phi_j), M_{t-1}(\phi_j))$
4. Forgetting Min:
$m_t(\phi_j) = \alpha H_t(\phi_j) + (1-\alpha)\min(H_t(\phi_j), m_{t-1}(\phi_j))$
end for
end for
end for

composed of the set of scalar statistics computed at different time scales and regions as $SR_d = \{\mu_{\alpha_i}, \sigma_{\alpha_i}, M_{\alpha_i}, m_{\alpha_i}\}_{i=0}^{n}$. *These statistics are constantly updated, storing information from all the time scales considered, using the* α_i *parameters. The vector of updated statistics feeds a previously trained support vector machine and a corresponding activity label is returned for the particular region of interest. This methodology is illustrated in Figure 3.*

For training purposes, several samples of incomplete videos were obtained from the original training set. For doing so, each of the videos was split several times, setting the video length to the alpha parameters considered in the experiments. This set was used to train the SVM that returns a label action for each time.

2.6. SVM Classification and Recognition

Finally, for each potential action, a descriptor of dimension d is produced, such that:

$$d = N_{div} \times N_{\phi} \times N_{stat} \times N_{\alpha}$$

with N_{div} the total number of subregions of the overlapped representation, N_{ϕ} the number of orientations for the velocity, N_{stat} the number of time statistics computed for each bin, and N_{α} the number of time scales.

The recognition of each potential activity is carried out by a Support Vector Machine (SVM) classifier since this constitutes a proper balance between accuracy and low computational cost. SVMs have been successfully applied to many pattern recognition problems, given their robustness, generalization aptness and low computational cost. Particularly, for action recognition, several

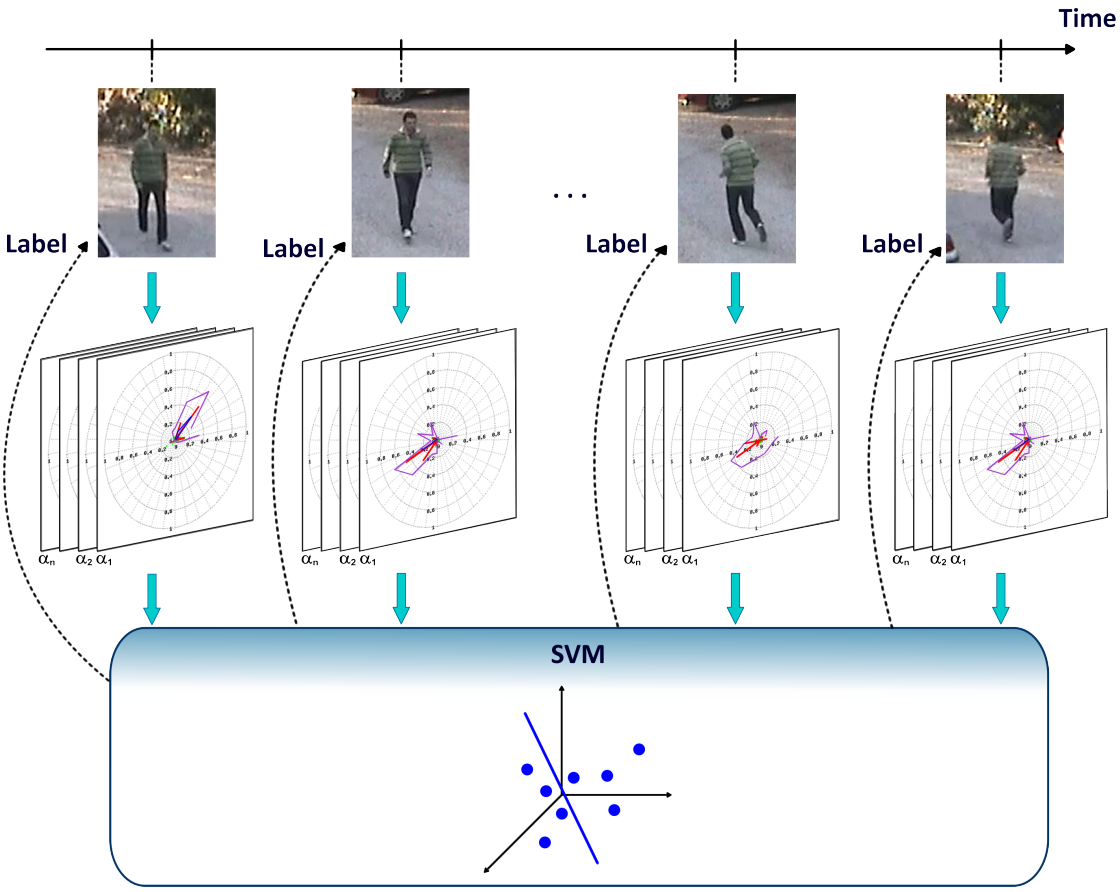


Fig. 3. *Pipeline of the frame-level action classification. The spatio-temporal histograms are quantified for each detected RoI and for each of the defined subregions from the overlapped representation. The motion descriptor is updated by the several statistics representing different temporal windows of analysis. The updated descriptor feeds a previously trained SVM and a label is returned for a particular RoI.*

approaches have been reported to use SVM classifiers [1], [35], [36]. The present approach was implemented using a One against one SVM multiclass classification with a Radial Basis Function (RBF) kernel [24]. Here, the classes represent the actions and optimal hyperplanes separate them by a classical max-margin formulation. For k motion classes, a majority voting strategy is applied on the outputs of the $\frac{k(k-1)}{2}$ binary classifiers. A (γ, C)-parameter sensitivity analysis was performed with a grid-search using a cross-validation scheme and selecting the parameters with the largest number of true positives.

3. Evaluation and Results

Experimentation was carried out with three public datasets, commonly used for assessing human action recognition tasks: (1) the ViSOR dataset (Video Surveillance Online Repository), captured by a real world surveillance system [26, 25], (2) UT-Interaction dataset (High-level Human Interaction Recognition Challenge) which is dedicated to complex human activities in real world scenarios [27], and (3) the classical KTH dataset which contains several activities recorded from

different viewpoints [35]. The ViSOR dataset is composed of videos showing 5 different human activities: walking, running, getting into a car, leaving an object and people shaking hands. These videos were captured by a stationary camera and contain a different number of actors and activities, (examples of actions are shown in Figure 4, first row).

The challenge is related with recognizing individual activities performed by several actors with different appearance, when the scene background differs and the motion direction may vary during the video sequence. Experiments with the ViSOR dataset consisted in classifying the different actions in 150 *videos with individual human activities. For testing, the dataset was partitioned into two complementary subsets, performing the analysis on the training set (*$60\,\%$*), and validating such analysis on the testing set (*$40\,\%$*). Each round of the validation scheme we applied randomly selects those two subsets of the original ViSOR set. A total of four rounds were performed, and the reported accuracy was the average over the rounds.*

The best reported performance was herein obtained by quantifying the motion into 32 *bins. The actions in ViSOR sequences have a relative periodic pattern and the dynamic description can be more easily recovered, therefore a 32 bins histogram allowed to capture these motions. In consequence, this reduced resolution allows us to built a compact descriptor that can be operated in online applications. Every RoI was built using three layers: then it was split into 2, 4, and then 9 divisions, representing a total of 15 histogram supports.*

The multi-temporal motion descriptor was computed using different decay parameters $\alpha_i = 2^{-i}$*, with* $i \in \{4, 5, 6\}$ *for* 3 *scales (i.e. time depths varying between 16 and 64 frames) and* $i \in \{5, 6, 7, 8, 9\}$ *for* 5 *scales (i.e. time depths varying between 32 and 512 frames). In this dataset,* 3 *scales obtained good results for the classification task, basically because most activities span a time period between 16 and 64 frames.*



Fig. 4. *The first row illustrates different examples of the human activities recorded in the ViSOR dataset. The second row shows different examples of activities in UT-interaction dataset*

The herein proposed strategy has succeeded in recognizing complex human actions in real scenarios. Such spatio-temporal descriptor has been able of popping out the most salient regions in terms of dynamic information by tracking the statistical dependency among the different partitions of the potential RoIs. This strategy was crucial when discriminating composed actions such as getting into a car or leave an object, since the decomposition into motion primitives captured the differences between similar actions.

Category	gc	lo	w	r	h
get car	100	0	0	0	0
leave Object	0	100	0	0	0
walk	0	0	83.3	16.7	0
run	0	0	14.29	85.71	0
hand shake	0	0	0	0	100

Table 1 Confusion Matrix obtained with the ViSOR dataset, using the proposed motion descriptor with 3 temporal scales. Results are in %. In average, the proposed approach achieves an accuracy of 93.3% . Some misclassified actions are walking and running

Table 1 and 2 show the confusion matrices obtained with the ViSOR dataset using a motion descriptor parametrized with 3 and 5 temporal scales, respectively. In general, these results demonstrate a good performance of the proposed descriptor in real-surveillance applications, obtaining an average accuracy of 93.3 and 96.7. *As expected, the motion descriptor may confuse actions like "walking" and "running", basically because their representation in terms of dynamic primitives is similar. Particularly, the activity labeled as "get into a car" is composed of a first part with a classical walking pattern which is followed by a bending down action to open the car. The proposed descriptor fails because most of the activity is devoted to the "walking" action. These failures were reported in Table 2.*

Category	gc	lo	w	r	h
get car	85.76	0	0	14.24	0
leave Object	0	100	0	0	0
walk	0	0	100	0	0
run	0	0	0	100	0
hand shake	0	0	0	0	100

Table 2 Confusion Matrix obtained with the ViSOR dataset using the proposed motion descriptor with 5 temporal scales. Results are in %. In average the proposed approach achieved an accuracy of 96.67%

Thanks to the recursive nature of the descriptor, action prediction can be made at any time of the sequence, which makes this approach adapted to online detection. For doing so, the motion is computed at each frame and mapped to a previously trained SVM model. We evaluated the accuracy of our approach in an online action recognition task for 5 long videos (each ~ 400 frames long) of the ViSOR dataset. In this evaluation, the online prediction performed by the herein proposed approach achieved an average per-frame accuracy of 92.3%.

Figure 5 illustrates the performance of the proposed method at the frame level for different ViSOR video sequences. The proposed approach generally achieves good recognition rates after few frames thanks to the recursive nature of the descriptor and its multiple time scales. When an action starts, the motion descriptor oscillates among different activities with close dynamics, until the history of the action reaches a sufficient number of frames. Specifically, a prediction delay is observed between frames 200 and 220 since the online descriptor, previously stabilized to the "walking" action, requires to recursively "forget" this motion pattern. On the left of the figure, the motion descriptor oscillates in the first frames by the similarity between "walking" and "get into a car" during the first part of the action. Some fluctuations are also observed in the middle of the sequence since actors remain sometimes static and therefore the optical flow is insufficient to

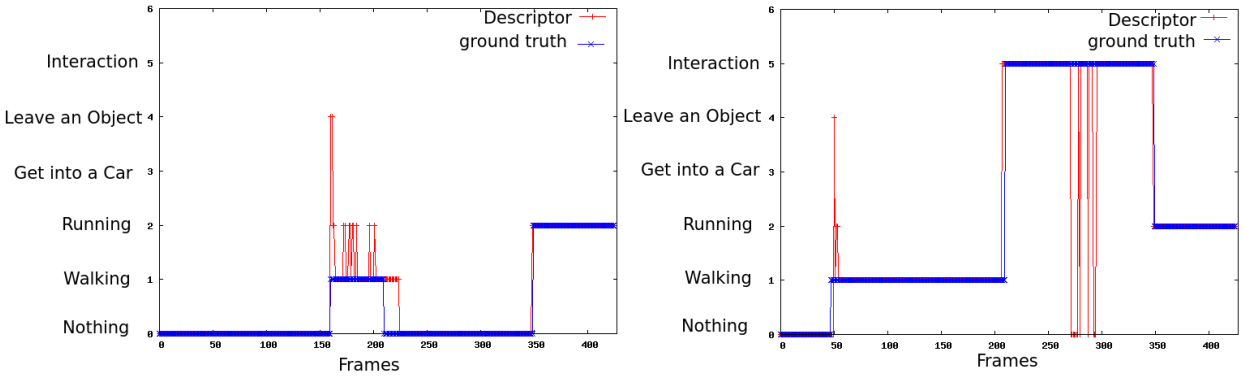


Fig. 5. *On-line action recognition for different videos recording several human motion activities. The frame-level recognition is carried out by mapping the motion descriptor to the SVM model at each time. The action label with minimal distance to the hyperplanes is assigned to the corresponding RoI.*

properly update the descriptor.

Additionally, the proposed motion descriptor was evaluated in the classical KTH dataset. This dataset contains six human action classes: walking, jogging, running, boxing, waving *and* clapping. *Each action is performed by* 25 *subjects in four different scenarios with different scales, clothes and scene variations. This dataset contains a total of* $2\,391$ *video sequences. The proposed approach was evaluated following the original experimental setup described in [35]. Since KTH contains individual actions recorded over relatively static scenarios, the same parameter configuration as ViSOR was applied. Every RoI was built using three layers as:* $\{2, 4, 9\}$*, representing a total of* 15 *histogram supports of* 32 *bins. In this dataset, the motion descriptor was computed using* 3 *temporal scales, using decay parameters* 2^{-i} *with* $i \in \{4, 5, 6\}$*. The proposed motion descriptor was also evaluated using four layers and five temporal scales but there were no significant improvements in the classification task.*

Table 3 shows the confusion matrix obtained for the KTH dataset using the proposed motion descriptor for the complete video sequence.The action label of each video was assigned by selecting the label most frequently predicted after computing the motion descriptor at each frame. The motion descriptor, as expected, confuses in some cases "jogging" and "running" activities, because of the similarities of such actions. The "clapping" action also generates misclassification because the periodic motion of limbs can be confused with other periodic actions. However these results show robustness with respect to the viewpoint, that can be further enhanced in future extensions of the proposed action descriptor.

The proposed spatio-temporal motion descriptor was also evaluated using the UT-interaction dataset. This dataset contains six different human interactions: shake-hands, point, hug, push, kick and punch (example of these actions are illustrated in Figure 4, second row)[27]. The actions included in this dataset are much more complex, i.e., they contain more interactions between different people, with a higher variability in the human appearance and motion patterns. A total of 120 *videos of this dataset were used for assessment. Each video has a spatial resolution of* 720×480 *and a frame rate of* $30\ fps$*. A ten-fold leave-one-out cross-validation was performed, as described in [27].*

Given the complexity of interactions and the similar dynamic relationships among the different

	box	hc	hw	jog	run	walk
box	91	4.8	0.0	0.0	0.0	4.2
hc	0.0	89	0.0	0.0	11	0.0
hw	3.0	5.0	92	0.0	0.0	0.0
jogg	0.0	0.0	0.0	89	4.0	7.0
run	0.0	0.0	0.0	12	88	0.0
walk	0.0	0.0	10	0.0	0.0	90

Table 3 Confusion Matrix obtained with the KTH dataset.Results are in $\%$. The proposed approach achieves an average score close to 90% for the multi-class recognition

recorded activities, the motion directions were quantized into 64 *bins for each of the different RoI sub-regions. Such histogram resolution allows us to improve discrimination between actions like Punching and Pushing while maintaining a low computational cost.*

A denser histogram representation was set for this dataset because of the dynamic complexity of the composed activities. *Two different subregion RoI configurations were herein considered by using three* $(2, 4, 9)$ *and four* $(2, 4, 9, 16)$ *layers. The best configuration of the motion descriptor for the UT-interaction dataset was obtained when the RoI was split using four layers, corresponding to a total of* 31 *subregions. In average, for the two UT-interaction datasets, the average accuracy decreased* 2% *when the proposed motion descriptor was run using only* 3 *layers for the RoI representation. The proposed motion descriptor was also tested using two different configurations for the temporal scales as follows:* $\{\alpha_i = 2^{-i}; i \in \{4.5, 6\}\}$ *and* $\{\alpha_i = 2^{-i}; i \in \{5, 6, 7, 8, 9\}\}$. *Three temporal scales proved to be sufficient for representing the UT-interaction actions and four scales showed no significant improvement. Although the actions recorded in such dataset are much more complex, the local movement that characterizes such actions occurs rapidly within a short time interval. From this observation, it is more useful to increase the histogram resolution rather than the temporal scales.*

Table 4 and 5 show the confusion matrices obtained when assessing with UT-interaction, using its two different datasets. In average, it was obtained an accuracy of 81.6 and 78.3 for dataset one and two, respectively. In summary, the proposed approach achieves a relevant dynamic characterization of the different human interaction activities. *However, such activities are often the result of combinations of complex motion patterns that may occur during a short time interval. Likewise, some of these interaction activities share local motion patterns that may lead to wrong predictions. For instance, interactions like "hand shaking", "pointing" or "pushing", share similar limb movements during certain temporal interval. Additionally, such scenarios highly increase the complexity of description, consider for instance a group of moving actors with a large variability in terms of appearance, interaction and background.*

Finally, Table 6 reports the comparison of the proposed motion descriptor with other state-of-the-art strategies. Some of these approaches achieve high accuracy rates in problems related with action recognition but they demand a complete processing of the video to compute the features that describe the sequence. For instance, the propagative voting approach [28] reports a computational complexity of $O(N_M)+O(WHT)$, *where* N_M *is the number of matches and* W, H, T *is the spatial(width* $\times$ *height) and temporal video resolutions. Such number of matches is computed by using random projection trees, a precise strategy that results computationally expensive and prohibitive for online applications.*

Other approaches combine different appearance and motion features that improve the perfor-

Category	hs	hg	ki	po	pun	pus
Hand Shaking	80	10	0	10	0	0
Hugging	10	80	0	0	0	10
Kicking	0	0	90	0	0	10
Pointing	10	0	0	90	0	0
Punching	0	10	10	0	80	0
Pushing	20	10	0	0	0	70

Table 4 Confusion matrix for UT-interaction dataset No-1.Results are in %

Category	hs	hg	ki	po	pun	pus
Hand Shaking	90	0	0	10	0	0
Hugging	0	80	10	0	10	0
Kicking	10	0	80	0	0	10
Pointing	10	0	0	80	10	0
Punching	0	0	20	0	80	0
Pushing	10	20	0	10	0	60

Table 5 Confusion matrix for UT-interaction dataset No-2. Results are in %.

mance in action classification tasks. Xiaofei et. al encode spatio-temporal points as enhanced BoW occurrence histograms after a preprossessing step (difference between consecutive frames). These histograms are combined with HoG to improve the action representation. However, the multiple steps of this approach result in a computationally expensive algorithm with particular parameterization for each of these steps, a limitation for specific scenarios and spatial configurations (appearance dependency). Particularly, this approach starts by computing frame differences to detect interactions in the video and define regions of interest. Then, it calculates 3d-SIFT points on the video. This computation requires the complete processing of fixed volumes to determine the keypoint candidates based on scale-time-space extrema detection. Such detection cannot be performed incrementally at each frame, and is applied on fixed time intervals of the video sequence. Then, 3d-SIFT descriptors (coded in 256 scalar values) are calculated on the detected points into a mid-level representation as visual words to compute a codebook of the actions. The codebook is based on a Multi-View Space Hidden Markov Models [34] that allows to learn time coherent fixed 3d-SIFT volumes. Although such mid-level strategy is computationally more expensive than classical Bag-of-Words, it allows to represent the actions for each temporal segment of volume computed for the 3d-SIFT. An additional step of video sequence characterization in this strategy consists in detecting regions of interest and split them into 16 *subregions. Each of these subregions is then described by HOG histograms. Once both representations are computed, independent nearest neighbor searches are performed to obtain a likelihood measure of each feature w.r.t the closest trained action. Finally the histograms (3d-SIFT and HOG occurrence) are independently normalized and weighted according to the estimated likelihood. The two histograms are concatenated and mapped to a new nearest neighbor to obtain classification of the actions. In [31], a bipartite graph key pose doublets identifies interactions from a large multidimensional pose descriptor which requires expensive algorithms to match the graph patterns with the new sequences. The approaches presented in [30] and [32] describe partial activities using histogram representations of some primitives computed at each frame but limited in term of accuracy. In contrast, the proposed*

Approaches	Accuracy UT-dataset 1	Accuracy UT-dataset 2
Propagative voting [28]	93	91
Proposed approach	81.6	78.3
Daysy [9]	71	51
SIFT 3D [29]	63	55
Slimani 2014 [30]	41	
Ryoo 2011 [32]	71.7	
Mukherjee [31]	79.17	
Xiaofei [33]	83.33	

Table 6 Average accuracy for different reported state-of-the-art strategies. Although the propagation voting achieves better results in terms of accuracy, the match of features using random projection trees is computationally expensive. The Xiaoafei *et. al.* work integrates BoW occurrence histogram with HoG, representing again a high computational time to obtain an action representation. In contrast, the proposed approach produces a compact descriptor that takes into account different time interval depths by using the same source of primitives, i.e., a dense optical flow. Additionally, the recursive nature of the proposed approach makes this estimator is constantly updated so that partial sequences can be predicted.

approach encodes motion orientations as histograms from a global perspective, as the region of interest, but also spatially localized from the spatially constrained representations. Such orientation histograms are summarized in a compact descriptor that involves the analysis of different time depth scales. Hence, the proposed descriptor robustly represents the spatio-temporal patterns of the activity and achieves a per-frame recognition. This proposed approach is suitable for real time applications because of its recursive construction and the partial description of the video.

Indeed, the proposed motion descriptor is quite simple and can be easily customized for real-time applications. The size of the frame-level action descriptor is $d = N_{div} \times N_\phi \times N_{stat} \times N_\alpha$ *with* N_{div} *the total number of subregions of the overlapped representation,* N_ϕ *the number of orientation bins for the velocity,* N_{stat} *the number of temporal statistics computed for each bin, and* N_α *the number of time scales. The computational complexity of frame-level classification is* $C_c = O(d)$*. However, such figures are constant and the number of orientation bins* N_ϕ *is the dominant factor. Although we used a dense optical flow field in our experiments, the proposed motion descriptor can be used together with any semi-dense optical flow method. In this work, we used a particular optical flow implementation based on the multiscale local jet nearest neighbor search. The local jet representation consists on characterizing each pixel by a set of derivatives estimated at different scales. Then pixel matching in adjacent frames is carried out by nearest neighbor search in a structured kd-tree space of local jet features. The computational complexity of the herein implemented optical flow is then* $C_f = O(W \times H \times d) + O(N \log N)$*, where the first term refers to the computation of the local jet feature (*W *and* H *being the image dimensions, and* d *the number of derivatives), and the second term to the nearest neighbor search,* N *being the number of points to match (which is at most* $W \times H$*). Finally, the complete complexity for the local jet dense flow computation and the frame level classification is given by:* $C_d = C_c + C_f$*.*

4. Discussion and concluding remarks

This article has introduced a novel motion descriptor consisting in the recursive computation of velocity orientation primitives that cover different temporal intervals and that are calculated on an

overlapped partition of moving regions of interest. Inspired by visual systems, the proposed approach carries out a spatio-temporal analysis of the environment and determines salient information as those regions with coherent motion patterns. The approach, assessed in action recognition applications, demonstrated action prediction capability at any time, becoming then a good candidate to be used in real time applications, while being easy to implement and efficient in terms of accuracy and time.

Characterization of flow motion, as the orientation occurrence, has been largely exploited in action recognition tasks, mainly because of the relative independence of the visual appearance as well as the flexibility to describe very different motions [11], [12]. However, these descriptors are in general restricted to represent invariant motions and therefore limited to specific applications. Also, they are commonly combined with appearance information that improves the description in specific scenarios but loses the flexibility of the motion flow information. The proposed approach is in contrast flexible and can represent a wide variety of real circumstances, from atomic gestures to interaction activities by computing cumulated statistics from the flow orientation histograms during variable time intervals. The different K temporal action periods (multi-frequency temporal representation) can be applied by adapting the hyper-parameter α as $\{\alpha_1 \dots \alpha_k\}$, which enhances the action description for different motion models while preserving the robustness to represent actions in high scene variations. In fact, for a simple gesture recognition, the temporal descriptor requires few temporal scales, while for interaction and behavioural activities, the different scales have to cover the dynamic description.

In contrast to block-based representations [9, 13, 12], in which every block appearance representation of the video has the same relevance, the representation herein introduced increases the importance of several flow orientation patterns that remain at different subregions and during a particular time interval. *The proposed motion descriptor has the capability to predict the actions at any time of the sequence by updating a motion descriptor from a multi-scale recursive framework. The computation time of the motion descriptor is acceptable, taking in average* $10\,ms$ *to be updated at each frame with the current motion information. In this time testing was computed motion descriptor using 3 temporal scales, 32 bins per hitogram and a total of 15 subregions of the RoI. Additionally, the mapping of each motion descriptor to the SVM model at each time takes in average* $11\,ms$*. The experiments were carried out on a single core i3-3240 CPU @3.40 GHz. According to the types of scene and actions, the proposed motion descriptor can be designed as a trade-off between time/memory efficiency and classification accuracy.*

This paper presented a novel approach that recognizes multiple human actions and classifies human activities as simple or complex. The descriptor consists of a series of statistics computed for different multiscale orientations and adapted intervals of time. The algorithm can run on line*, thanks to its recursive nature and* fast*, even on a main stream architecture. The proposed descriptor achieved an average accuracy of* 95% *in the real surveillance dataset VISOR and* 80% *in the UT-interaction dataset. The proposed descriptor was mainly assessed in videos acquired with a stationary camera, yet it can be adapted to actions occuring in mobile scenarios, by adjusting the time intervals w.r.t to the camera motion, or by discarding the main optical flow clusters as representing the background. Future works include evaluation of the proposed descriptor in such scenarios. Likewise, this descriptor will be extended to recognition of interactive actions such as human group activities.*

5. References

[1] Aggarwal, J. K., Ryoo, M. S.: 'Human activity analysis: A review', ACM Computing Surveys (CSUR)., 2011, 43(3), pp. 16.

[2] Vishwakarma, S., Agrawal, A.: 'A survey on activity recognition and behavior understanding in video surveillance', The Visual Computer, 2013, 29(10), pp 983-1009

[3] Borges, P. V. K., Conci, N., Cavallaro, A.: 'Video-based human behavior understanding: A survey', IEEE transactions on circuits and systems for video technology, 2013, 23(11), pp 1993-2008

[4] Vrigkas, M., Nikou, C., Kakadiaris, I. A.: 'A review of human activity recognition methods', Frontiers in Robotics and AI, 2015 , 2, pp. 28

[5] Liu, A. A., Xu, N., Su, Y. T., Lin, H., Hao, T., Yang, Z. X.: 'Single/multi-view human action recognition via regularized multi-task learning', Neurocomputing, 2015, 151, pp. 544-553

[6] Chen, C. Y., Grauman, K.: 'Efficient activity detection with max-subgraph search' In Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1274-1281

[7] Samanta, S., Chanda, B.: 'Space-time facet model for human activity classification', IEEE Transactions on Multimedia, 2014, 16(6), pp. 1525-1535

[8] Riemenschneider, H., Donoser, M., Bischof, H.: 'Bag of Optical Flow Volumes for Image Sequence Recognition', In BMVC, 2009, pp. 1-11

[9] Cao, X., Zhang, H., Deng, C., Liu, Q., Liu, H.: 'Action recognition using 3D DAISY descriptor', Machine vision and applications, 2014, 25(1), pp. 159-171

[10] Wang, H., Oneata, D., Verbeek, J., Schmid, C.: 'A robust and efficient video representation for action recognition', International Journal of Computer Vision, 2016, 119(3), pp. 219-238.

[11] Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: 'Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions', In computer vision and pattern recognition, 2009. CVPR 2009. pp. 1932-1939

[12] Ikizler, N., Cinbis, R. G., Duygulu, P.: 'Human action recognition with line and flow histograms', In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pp. 1-4

[13] Tabia, H., Gouiffes, M., Lacassagne, L.: 'Motion histogram quantification for human action recognition', In Pattern Recognition (ICPR), 2012, pp. 2404-2407

[14] Zhang, Z., Hu, Y., Chan, S., Chia, L. T.: 'Motion context: A new representation for human action recognition', In European Conference on Computer Vision, 2008, pp. 817-829

[15] Ji, S., Xu, W., Yang, M., Yu, K.: '3D convolutional neural networks for human action recognition', IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1), pp. 221-231.

[16] Taylor, G. W., Fergus, R., LeCun, Y., Bregler, C.: 'Convolutional learning of spatio-temporal features', In European conference on computer vision, 2010, pp. 140-153

[17] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: 'Sequential deep learning for human action recognition', In International Workshop on Human Behavior Understanding, 2011, pp. 29-39

[18] Vrigkas, M., Karavasilis, V., Nikou, C., Kakadiaris, I. A.: 'Matching mixtures of curves for human action recognition', Computer Vision and Image Understanding, 2014, 119, pp. 27-40

[19] Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., Sinha, P.: 'Visual parsing after recovery from blindness', Psychological Science, 2009, 20(12), pp. 1484-1491.

[20] Manzanera, A.: 'Local jet feature space framework for image processing and representation', In Signal-Image Technology and Internet-Based Systems (SITIS), 2011, pp. 261-268

[21] van Hateren, J. H., Ruderman, D. L.: 'Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in p0rimary visual cortex', Proceedings of the Royal Society of London B: Biological Sciences, 1998, 265(1412), pp. 2315-2320.

[22] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection', In Computer Vision and Pattern Recognition, CVPR 2005, 1, pp. 886-893

[23] Richefeu, J., Manzanera, A.: 'A new hybrid differential filter for motion detection', In Computer Vision and Graphics, 2006, 32, pp. 727-732

[24] Chang, C. C., Lin, C. J.: 'LIBSVM: a library for support vector machines', ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3), pp. 27

[25] Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: 'Effective codebooks for human action categorization' In Computer Vision Workshops (ICCV Workshops), 2009, pp. 506-513

[26] Vezzani, R., Cucchiara, R. 'Video surveillance online repository (visor): an integrated framework', Multimedia Tools and Applications, 2010, 50(2), pp. 359-380.

[27] Ryoo M., Aggarwal J.: ' UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)', 2010

[28] Yu, G., Yuan, J., Liu, Z.: 'Propagative hough voting for human activity recognition', In European Conference on Computer Vision, 2012, pp. 693-706.

[29] Scovanner, P., Ali, S., Shah, M.: 'A 3-dimensional sift descriptor and its application to action recognition', In Proceedings of the 15th ACM international conference on Multimedia, 2007, pp. 357-360

[30] Nour el houda Slimani, K., Benezeth, Y., Souami, F.: 'Human interaction recognition based on the co-occurence of visual words', In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 455-460

[31] Mukherjee, S., Biswas, S. K., Mukherjee, D. P.: 'Recognizing interaction between human performers using'key pose doublet", In Proceedings of the 19th ACM international conference on Multimedia, 2011 ,pp. 1329-1332

[32] Ryoo, M. S.: 'Human activity prediction: Early recognition of ongoing activities from streaming videos', In Computer Vision (ICCV), 2011, pp. 1036-1043

[33] Ji, X., Wang, C., Zuo, X., Wang, Y.: 'Multiple Feature Voting based Human Interaction Recognition', International Journal of Signal Processing, Image Processing and Pattern Recognition, 2016, 9 (1), pp. 323-334

[34] Ji, X., Wang, C., Li, Y., 'A view-invariant action recognition based on multi-view space hidden markov models', 2014, International Journal of Humanoid Robotics, 11(01), p. 1450011.

[35] Schuldt, C., Laptev, I., Caputo, B.: 'Recognizing human actions: A local SVM approach', In Pattern Recognition, ICPR 2004, 3, pp. 32-36

[36] Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: 'Behavior recognition via sparse spatio-temporal features', In Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65-72

[37] Ryoo, M. S., Rothrock, B., Matthies, L.: 'Pooled motion features for first-person videos', In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 896-904