Neural Computing and Applications (2020) 32:16387–16400
https://doi.org/10.1007/s00521-018-3951-x

IAPR-MEDPRAI

Human activity recognition via optical flow: decomposing activities into basic actions

Ammar Ladjailia[1] · Imed Bouchrika[2] · Hayet Farida Merouani[1] · Nouzha Harrati[2] · Zohra Mahfouf[2]

Received: 24 July 2018 / Accepted: 18 December 2018 / Published online: 1 January 2019

Abstract

Recognizing human activities using automated methods has emerged recently as a pivotal research theme for security-related applications. In this research paper, an optical flow descriptor is proposed for the recognition of human actions by considering only features derived from the motion. The signature for the human action is composed as a histogram containing kinematic features which include the local and global traits. Experimental results performed on the Weizmann and UCF101 databases confirmed the potentials of the proposed approach with attained classification rates of 98.76% and 70%, respectively, to distinguish between different human actions. For comparative and performance analysis, different types of classifiers including Knn, decision tree, SVM and deep learning are applied to the proposed descriptors. Further analysis is performed to assess the proposed descriptors under different resolutions and frame rates. The obtained results are in alignment with the early psychological studies reporting that human motion is adequate for the perception of human activities.

Keywords Action recognition · Motion descriptor · Optical flow · Decomposing activities

1 Introduction

Much scientific research in computer vision is dedicated to the arena of human motion analysis. These studies are supported by the large number of applications where automated analysis of human motion is deemed very crucial including biometrics, smart automated surveillance, sports arbitration and human–machine interaction. As we are becoming more digital natives in such a modern era, the recognition of human activities is becoming an interesting research area with the potency to be integrated within various realistic human-centric contexts [1, 6]. Additionally, because of the unprecedented increase in multimedia data produced continuously from security cameras, movie production and Web uploads, it is now becoming an important necessity to analyse such video content semantically via automated methods. This would be a major milestone to facilitate the process of indexing, search and retrieval of multimedia content. The deployment of automated vision systems to recognize human activities can stand as an innovative solution to increase the adoption and usability for such smart visual applications.

The process of extracting and recognizing human actions via automated marker-less methods are two separate tasks that are affirmed to be cumbersome and complex. Devising an automated solution can be difficult to develop and generalize to different settings due to various reasons that can be linked to either: acquisition settings, subject or activity context. Earlier approaches proposed for this task, depending on special equipment mounted on the person including sensors [29]. On the other hand, vision-based solutions are not in a mature state mainly because of the high degree of freedom for the human body in tandem with the unpredictable appearance variability. This would exacerbate additional challenges within the feature extraction stage [33]. Difficulties can stem from the acquisition environment which includes illumination,

✉ Ammar Ladjailia
a.ladjailia@univ-soukahras.dz

Imed Bouchrika
imed@imed.ws

[1] Department of Computer Science, University of Annaba, 23000 Annaba, Algeria

[2] Faculty of Science and Technology, University of Souk Ahras, 41000 Souk Ahras, Algeria

background clutter, viewpoint and camera movement as well as self-occlusion or occlusion made by other objects. For the last factor, people can perform the same activity in various fashions and ways [57]. This is dependent on the culture, context or people themselves. Moreover, a specific activity carried out by different subjects can have totally unrelated and different semantics. More challengingly, most human activities are performed in parallel being interleaved within each other as we do rarely behave or interact in a sequential fashion. For example, a subject can use a desktop whilst drinking or talking on the phone at the same time.

Because of the incontestable role of automated systems in security surveillance for human activity recognition, we describe in this paper a marker-less motion-based descriptor using optical flow features for the automated recognition of human activities. The proposed approach does not dependent on background subtraction due to the intricate nature of outdoor footage which is often subjected to various challenging environmental conditions. Alternatively, motion features are derived from estimating optical flow from a triplet of consecutive images to generate discretized index value which describes the temporal orientation at a locus point. A histogram is constructed from consecutive frames such that kinematic-based data from optical flow describing the global and local properties are considered. Two different classification paradigms are considered. The first concerns the classification based on the feature selection using the simple K-nearest neighbour (KNN) classifier. Further, deep learning is considered during this research as a more advanced classifier which is based an autoencoder neural network. Experimental results performed on the Weizmann and UCF101 datasets affirmed the potentials of the proposed approach to better distinguish between different human basic actions. This a milestone to extend the proposed procedure to recognize further composite actions and activities. To compare our results against state-of-the-art methods in computer vision and machine learning, two recent studies on the use of deep learning for human activities recognition are assessed using the same dataset. Further experiments are performed to explore the performance of motion features for human action recognition under different scenarios including lower resolution and decreased frame rates.

This paper is organized as follows. The next section outlines the previous approaches for the *marker-less extraction and recognition of human activities*. The theoretical description of the presented method for extracting and reconstructing a motion-based activity signature is detailed in Sect. 3. The following section introduces the experimental results performed on the Weizmann and UCF101 datasets. Conclusions and future work are drawn at the end.

2 Related work

Based on major studies within the literature, the two terminologies “action” and “activity” are mentioned interchangeably and contentiously with some overlap [38]. *Action* can be defined as a very basic activity or simple movement carried out by a subject within a short interval lasting for a few seconds. This can include, for instance, bending, sitting and waving hands. Poppe [38] explained further the word *action primitive* as an atomic movement at the limb level. An *activity* can be described as a sequence of basic actions performed by an individual or group of people. Cases of activities include complex actions such as leaving an unattended bag, assaulting a pedestrian or shaking hands. A vision-based system for human activity recognition is composed of three principal phases: detection, tracking and the interpretation of the *activity* or *action*. The automated detection of human activities plays a vital role in various applications and innovative systems including smart homes and visual surveillance. Although there has been a considerable body of research devoted to analysing human motion, classification of human activities and pose reconstruction, recent research focus is moving towards using nonintrusive and marker-less computer vision methods for the detection of human activities from natural and realistic complex scenes rather than using laboratory settings [36]. There are a number of research studies within the arena of computer vision on the use of deep learning for human activity recognition. In [14, 28, 50], research studies have reported the suitability of employing deep learning approaches to classify human actions whilst they stressed on the difficulty in treating the temporal dimension for video sequences. Asadi-Aghbolaghi et al. [4] have recently surveyerd deep learning methods for human activities proposing a taxonomy of three major classes which are 3D models, motion-based input features and temporal methods.

2.1 Representation of features

For the categorization of existing methods in the area of automated detection of human activities, there is a consensus among major surveys [12, 38, 46] to have two broad categories based on the representation of features including either global or local representation. The global features are derived from a person as a whole after applying foreground segmentation. The estimation of such features is based on low-level data including edges, interest points or optical flow. Poppe [38] have argued that methods based on the global representation are prone to different factors including occlusion, noise and camera viewpoint variations. Weinland and Boyer [52] proposed a compact global representation for human activity recognition which can be matched against a set of prebuilt discriminative pose templates. The representation is based directly on edge data without the need for background subtraction whilst the matching process is performed via the Chamfer distance. For the local representation of features, smaller regions or patches are derived independently from a given image in order to produce a feature vector. The main merits of using local representation are its invariance to appearance variations as well as background clutter. Further, the requirement for a good localization of the region of interest can be relaxed. Kliper-Gross et al. [27] deployed the local representation descriptor proposed by Yeffet and Wolf [54] for the automated classification of human activities using bag of visual words combined with the use of the support vector machine. Their proposed descriptor is based on matching local patches against neighbouring regions within the previous and next frames. In a different study, Oshin et al. [36] deployed the distribution of interest points at the spatiotemporal level for the classification of human actions in an unconstrained environment.

2.2 Optical flow-based descriptors

The use of optical flow has been considered as a strong low-level feature within various vision-based applications. This is because motion-based features are considered as a strong visual attention cue for the perception of scenes [22]. Chaudhry et al. [7] proposed the Histogram of Oriented Optical Flow (HOOF) descriptor reporting its invariance to motion orientations and scale. The descriptor is constructed by estimating optical flow features on every frame without the need for background segmentation or the localization of the subject. Subsequently, the Binet–Cauchy kernels are applied for matching nonlinear histograms. Their approach was evaluated on the Weizmann dataset with a reported correct classification rate of 95.66%. Martínez et al. [32] deployed the optical flow in order to estimate the velocity vector at each pixel. For each frame, an accumulated local histogram is constructed containing the motion orientations for the optical vectors which are discretized uniformly into 32 directions. The global histogram for the human action is composed of 192 bins by concatenating 6 local consecutive histograms. Based on the Weizmann database, a correct classification score of 95% is attained using support vector machine. Wang et al. [49] introduced the *optical flow image* as an ordered and compact representation from optical flow data from consecutive frames. Colque et al. [9] proposed the *Histograms of Optical Flow Orientation and Magnitude* descriptor via estimating optical flow from cuboids regions taking into account the temporal and spatial dimensions. Their proposed descriptor was applied for the detection of abnormal activities in surveillance scenarios.

2.3 Mining for basic actions

There is a recent trend within the research community towards the process of mining for basic human actions for the automated recognition of human activities within complex scenes. Yi et al. [55] captured the evolution of human motion for the classification of complex actions. The temporal structural information is derived based on key frame selection where a hierarchical video representation is proposed based on trajectory sheaf to encode video clips at different levels. Feng et al. [18] exploited the use of a mining process for spatial–temporal patterns in order to construct a data-driven-based human motion denoising method. Detection of basic actions and motion patterns is conducted using a dictionary learning method where multiple compact and representative motion keywords are learned from training data. Alfaro et al. [2] proposed a method for reducing a video to a set of key sequences representing significant atomic acts of each action class. Zhu et al. [56] proposed an approach called key volume mining deep framework for the application of human action recognition. The framework is based on mining key or rudimentary volumes for each human action class.

3 Proposed approach

The system proposed for the analysis and classification of human actions is composed of three main building blocks. The main assumption for the system is that the video should contain only motion related to a human action. Initially, the optical flow is estimated through a consecutive set of frames. In the next stage, the feature vector is constructed as a histogram from the motion descriptors for the frames being considered. Action classification is performed using simple classifiers based on a subset of features derived through the training phase. Figure 1 shows the flow diagram for the proposed approach for human action recognition. For people detection, the Histograms of Oriented Gradients based on the pedestrian detector from [10] can be utilized. This detector attains the state-of-the-art performance on full-body pedestrian detection. In order to increase the Recall of person detection in difficult conditions, a simple approach for person tracking is deployed. The bounding box of each detected pedestrian is propagated subsequently to the next frame [5].

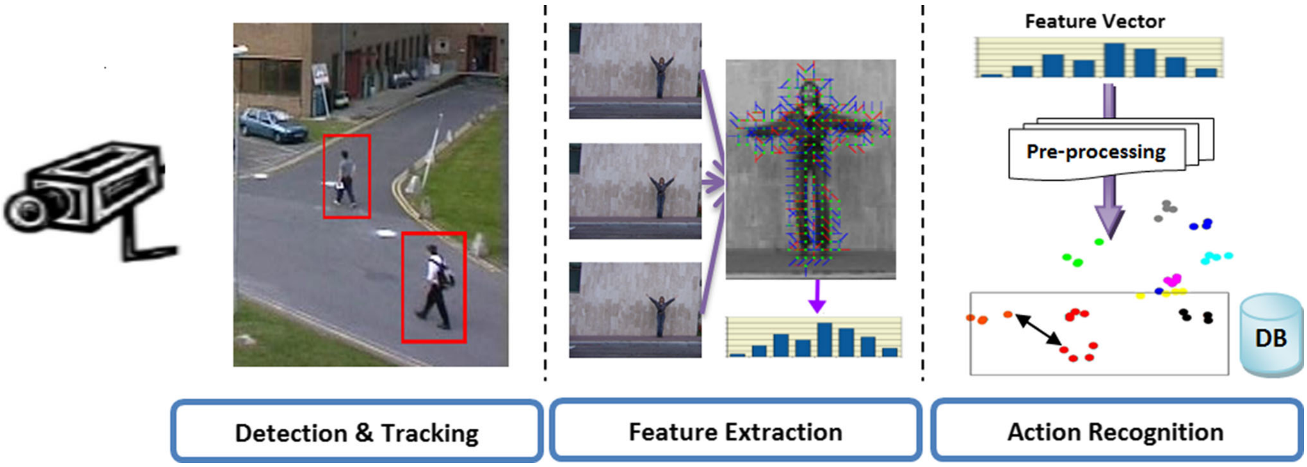


Fig. 1 Overview of the proposed system for human action recognition

3.1 Optical flow estimation

The proposed approach is based on estimating optical flow in order to generate a feature vector from a sequence of consecutive frames to describe a human activity. The approach relaxes totally the requirement for foreground segmentation since it is computationally a prohibitive process to update the background model for real-time outdoor surveillance systems where the process to update the background model is affected by various factors including weather conditions, background clutter and other environmental conditions. Based on the work of Kliper-Gross et al. [27] who introduced the Motion Interchange Pattern to recognize human actions in tandem with the confirmed merits and effectiveness for using local feature representation, we describe a local descriptor based on computing optical flow where features are derived at a local level. Given the existence of motion within a small area between two consecutive frames t and $t+1$, the displaced patch should be located easily within the neighbouring regions at the new frame. The introduced descriptor composes a feature vector describing the displacements of local patches across consecutive frames. In fact, it is impractical to conduct a brute force search to estimate the displacements of every patch using similarity matching operators due to two main reasons: Firstly, poor matching results can be produced due to the high self-similarity around surrounding regions. Secondly, displaced patches can have their appearance changed either due to the flexibility of the human body or to the environmental conditions. Alternatively, the optical flow is considered in this research study in order to extract the motion displacement information from consecutive frames. Optical flow has attracted considerable interest from the research community in computer vision because of its pivotal role for numerous applications including autonomous vehicle, security surveillance and defence systems. The estimation of optical flow is based on observing the movement of intensity values from frame to the next [19]. The flow vector can result either from an object moving within the monitored scene or instead due to the movement of the camera. Given a frame I_t at time t, the basis of optical flow considers that the intensity for a moving object of coordinates $I(x, y, t)$ stays constant as elaborated below:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{1}$$

such that δx and δy are the displacement values in the x-direction an y-direction, respectively, in time δt. The intensity constancy basis is described using the differential form as expressed in (2):

$$\frac{\mathrm{d}I}{\mathrm{d}t} = 0 \tag{2}$$

The application of the Taylor series in (2) becomes:

$$\begin{aligned} I(x + \delta x, y + \delta y, t + \delta t) &= I(x, y, t) \\ &+ \delta x \frac{\partial I}{\partial x} + \delta y \frac{\partial I}{\partial y} + \delta t \frac{\partial I}{\partial t} + \epsilon \end{aligned} \tag{3}$$

where ϵ represents the second- and higher-order terms for the Taylor series and $\frac{\partial I}{\partial x}$ is the partial derivative of frame I with respect to the x variable. As both sides of (1) are equal, (3) can be simplified to produce the following equation:

$$\delta x \frac{\partial I}{\partial x} + \delta y \frac{\partial I}{\partial y} + \delta t \frac{\partial I}{\partial t} + \epsilon = 0 \tag{4}$$

Horn–Schunck [21] proposed dividing Eq. (4) by δt to obtain:

$$\frac{\delta x}{\delta t}\frac{\partial I}{\partial x}+\frac{\delta y}{\delta t}\frac{\partial I}{\partial y}+\frac{\partial I}{\partial t}+O(\delta t)=0 \tag{5}$$

where $O(\delta t)$ is a term of order δt. In the limit $\delta t \to 0$. Therefore, (5) is written as :

$$\frac{\delta x}{\delta t}\frac{\partial I}{\partial x}+\frac{\delta y}{\delta t}\frac{\partial I}{\partial y}+\frac{\partial I}{\partial t}=0 \tag{6}$$

Differential techniques are the most used methods to find the solution for (6) with the two unknowns $\frac{\delta x}{\delta t}$ and $\frac{\delta y}{\delta t}$. In this research, the approach proposed by Horn–Schunck [21] is employed for computing optical flow between consecutive images.

3.2 Motion-based descriptor

The human activity descriptor is constructed by considering a set of frame triplets. For a single triplet of consecutive frames denoted, respectively, as *previous*, *current* and *next*, a descriptor value d at the pixel level is computed for every point within the locus frame via estimating the optical flow images for $v_{prev}:\{previous, current\}$ and $v_{next}:\{current, next\}$. In order to retain only informative features, thresholding process is employed using the magnitude of optical flow to filter out values which are less than the value of $\tau = 0.5$. In order to produce a histogram-based descriptor, the orientation for the optical flow vectors is discretized by dividing the polar coordinate system into 8 even sections numbered from 1 to 8. Based on the angle value for the optical vector, the *index* within the produced circular sections is considered as the discretized value for the flow vector. This is formally explained in (7). For cases where there is no motion or the vector is filtered out during the thresholding process, the *index* value is set to zero.

$$index_{a,b}(x,y)=\left\lfloor\frac{Angle_{a,b}(x,y)\times 8}{2\times\pi}\right\rfloor+1 \tag{7}$$

such that $\lfloor\ \rfloor$ represents the integer part a real number. a and b are two successive frames. As two optical flow images are computed for a triplet of frames, the resulting *index* at each point within the previous and next frames is joined together to produce a number at base 9 which is subsequently converted to base 10 as expressed in (8). The produced number is the descriptor value for the point at the coordinate (x, y)

$$desc_t(x,y)=index_{t_1,t_2}(x,y)+index_{t_2,t_3}(x,y)*9 \tag{8}$$

such that t_1, t_2 and t_3 are the first, second and third frames within a triplet t. The produced number using the $desc_t(x,y)$ function is the descriptor value for the point at the coordinate (x, y). Based on empirical experiments, a basic action can be represented sufficiently using a set of 15 consecutive frames for the case videos recorded with a frame rate of 25 frames/second. The encoding process is conducted to include seven triplets for each human action in order to produce the histogram of optical flow orientation as expressed in (9).

$$H_t(i)=\sum_{x,y}\begin{cases}1 & \text{if } desc_t(x,y)==i\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

Figure 2 illustrates the full procedure to generate the histogram of optical flow orientation features from local features. There is a Boolean basic function which returns 1 for true cases and 0 otherwise. H_t refers to the histogram obtained at the tth triplet. In this study, a number of motion features which can embed more distinctive traits on the human action are further constructed by applying simple fusion techniques such as arithmetic and statistical fusion methods being performed on the set of orientation histograms produced from (9). In (10), the equation expresses the produced feature vector by concatenating the different histograms. The standard deviation is abbreviated as STD. The resulting action vector is composed from features describing solely local dynamic-based features without considering information related to the global spacial structure of the activity neither the anthropometric nor anatomical data.

$$F_{local}=\left[H_1\ldots H_7\ Mean\ (H_1\ldots H_7)\right.\\ \left. STD(H_1\ldots H_7)\sum_{t=1}^{7}H_t\right] \tag{10}$$

In order to extract the global spatial features that better describe the geometric properties of the motion cues, every image of the optical flow taken from a pair of consecutive frames is stripped both vertically and horizontally into adjacent bars of similar width as shown in Fig. 3, in contrast to most studies which are based on splitting the region of interest into a grid of cells bounded to the location of the subject. Because people may move and it is an essential requirement to capture the spatial displacement across frames, two histograms are constructed from the sequence of optical flow images that should express the spatial movement vertically and horizontally. The optical flow vectors contained within every bar are accumulated together into their respective bin of histogram regardless of the motion orientations which are better considered during the local features extraction.

3.3 Classification process

In this step, we considered two methods for the classification. The first concerns the classification based on the feature selection using the simple K-nearest neighbour (KNN) classifier followed by the use of different

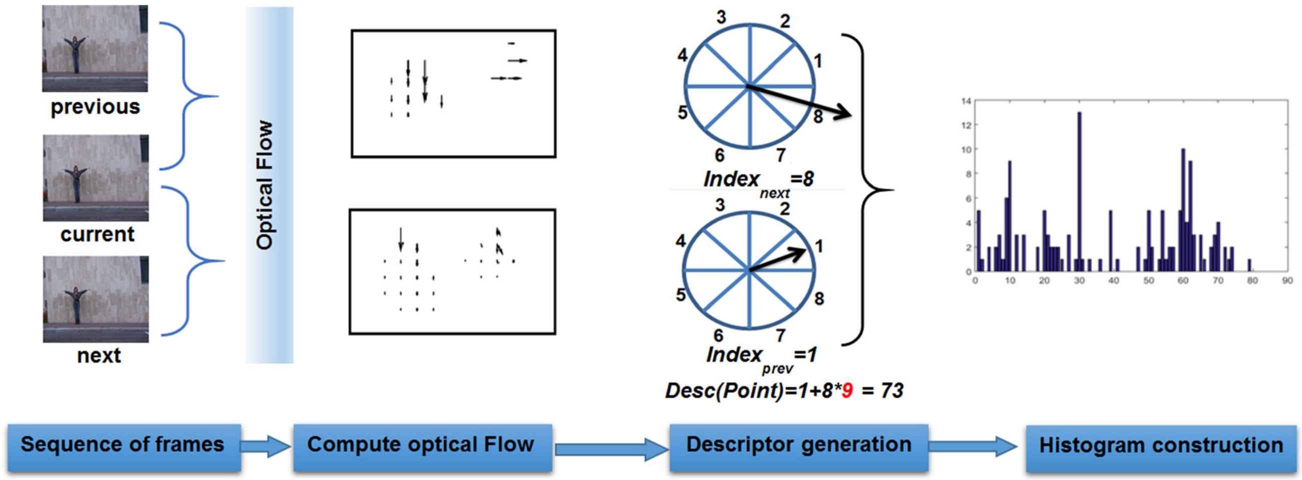


Fig. 2 Histogram construction from local motion descriptors

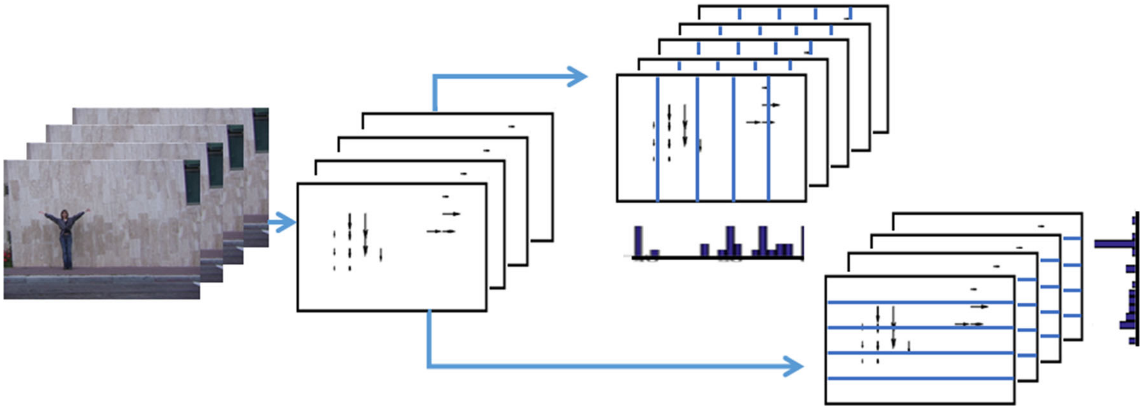


Fig. 3 Derivation of global motion features

classification methods including decision tree and support vector machine (SVM). The KNN and SVM are considered as baselines to compare against recent major studies whilst other advanced derivatives of the KNN [20] and SVM [24] can be deployed. Furthermore, deep learning is considered during this research as a more advanced classifier which is based on an autoencoder neural network. For the first classification paradigm, a feature selection procedure is devised in order to derive distinctive and representative features for the application of human activity recognition, as it is impossible to search in an exhaustive or brute force fashion for all combinations of subsets to find the most discriminative subset of features. This is due to the dimensionality of the raw feature vector. Alternatively, the Adaptive Sequential Forward Floating Selection algorithm is utilized to extract a subset of features. In this research, an objective function is proposed as an evaluation metric that evaluates the distinctiveness of each raw vector or set of features in order to extract the optimal features for human activity recognition. The validation-based criterion is deployed to select the representative features which minimize the mis-classification and maximize inter-class separation among different classes of human activities. Advanced filtering approach can be potentially deployed to optimize the separation between different classes [30]. In this research, a similar voting procedure for the nearest neighbour classifier is utilized. The evaluation criterion uses a coefficient w which reflects the importance of the nearest neighbours belonging to the same class. A score value for a given instance s to belong to a class c is expressed in (11). The *winner-take-all* approach can be potentially deployed within the selection procedure in the same way to derive the optimal subset of features having the highest score [44, 48].

$$P(s,c) = \frac{\sum_{i=1}^{N_c-1} z_i(s,c)w_i}{\sum_{i=1}^{N_c-1} w_i} \tag{11}$$

such that N_c is the number of objects within class c, and the coefficient w_i for the ith nearest instance is inversely related to their nearness or proximity as given:

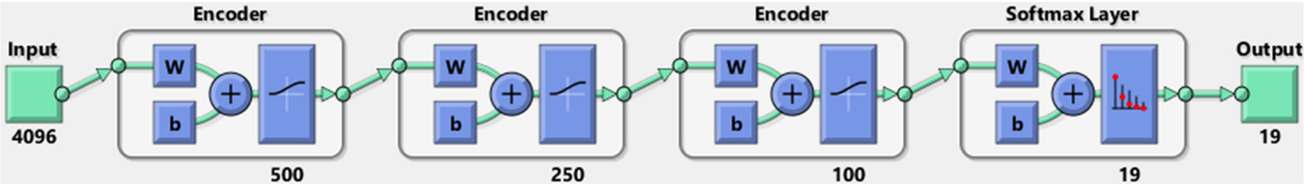


Fig. 4 Classification with deep learning

$$w_i = (N_c - i)^2 \tag{12}$$

z_i is computed as:

$$z_i(s,c) = \begin{cases} 1 & \text{if } nearest(s,i) \in c \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The $nearest(s, i)$ retrieves the ith candidate in terms of proximity to the instance s. To infer the distance and proximity level, the Euclidean distance is computed between different instances. To evaluate whether a subset of features has the most potency to classify human activities, the feature selection procedure is integrated with a validation-based metric which is computed using the leave-one-out cross-validation. In simple terms, the final activity descriptor with the optimal subset of features is composed from features *subset* among the raw space F such that the maximum validation value is achieved as the average sum of all computed values across the N candidates as explained in (14):

$$Action = \underset{subset \in F}{\arg\max} \left(\frac{\sum_{x=1}^{N} L_{subset}(x)}{N} \right) \tag{14}$$

such that L is the leave-one-out cross-validation function. For the second classification paradigm, deep learning is deployed on the same dataset within this research to assess the potentials of using optical flow features for the recognition of human activities. The deep learning classifier is constructed from a set of three Autoencoders with a softmax network layer in order to process the produced features at the classification stage. The new representation of features is carried out by passing an input of 2D raw features into a sequence of three Autoencoders composed of different sizes neurons as shown in Fig. 4.

4 Experimental results

4.1 Human action datasets

In order to assess the use of motion-based features derived from optical flow using the proposed descriptor, two separate datasets are considered for the evaluation process.

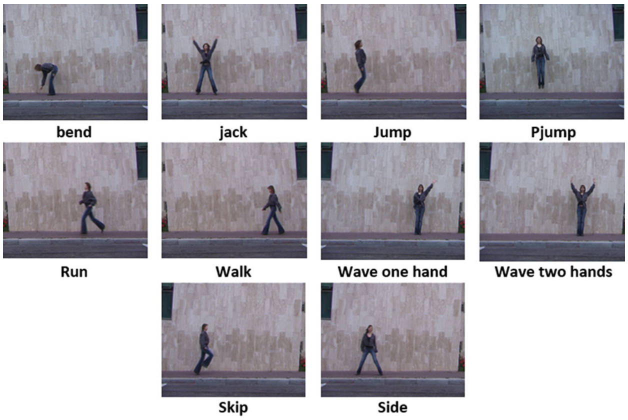


Fig. 5 Basic actions from Weizmann dataset

- **Weizmann dataset:** is made of 90 videos recorded with resolution of 180 × 144 at a frame rate of 25 frames per second. The dataset contains 9 different subjects who are instructed to performing 10 different basic activities as depicted in Fig. 5. For this research, a new dataset composed of 241 video sequences is constructed from the original Weizmann dataset by manually annotating videos to search for 19 different basic actions. Each sequence of basic actions runs for 15 frames which are all manually verified and checked to represent a full human action. The list basic actions include: running, walking, siding jumping and skipping from left to right (*LTR*) and vice versa. Additional actions include waving both hands, one hand and bending. The activity of *pjump* is considered as an action as it can be contained within 15 frames.
- **UCF101 dataset** [42]: 72 samples for 23 different classes that describe basic actions carried out by different users are taken from the UCF101 dataset. The collected videos from this dataset are chosen such that there is no movement for the camera. As addressing the camera motion can be easily compensated using off the shelf tools such as the conventional motion compensation (MOCO) method [15] or camera motion compensation by real time [23], the focus for this research was devoted to the detection of basic actions from still videos. Samples from the UCF101 dataset are shown in Fig. 6.

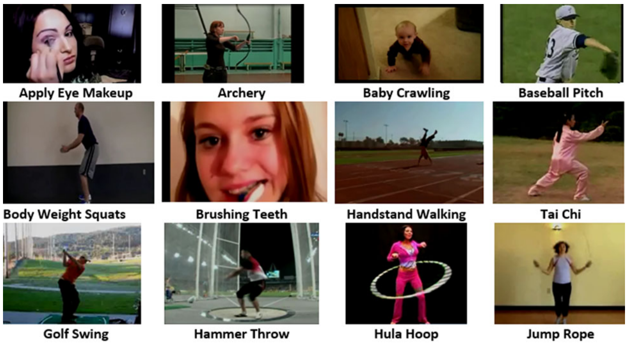


Fig. 6 Sample actions from UCF101 dataset

4.2 Human action classification

Using the feature selection procedure for the classification process, an optimal signature for human actions is derived containing 41 features. Two thirds of the dataset is used purely for feature selection whilst the remaining third is considered as the probe dataset. The K-nearest neighbour (KNN) classifier is used to compute the correct classification rate using different values of $k \in 1, 3, 5$ using the leave-one-out cross-validation. We considered the use of the KNN classifier for the classification stage due to its simplicity in addition to the convenience of comparing the obtained results against earlier techniques being applied on the same dataset. The Cumulative Match Score (CMS) is computed in order to assess the classification after different iterations or ranks. A high Correct Classification Rate (CCR) of 98.76% for the 19 basic actions is achieved at rank $R = 1$, and meanwhile, a CCR of 100% is reported at rank $R = 2$, respectively. The CMS curve is depicted in Fig. 7 for the classification process applied on the Weizmann dataset. Table 1 shows the results obtained for the various classifiers with different spacial bars applied to the Weizmann dataset. The number of *bars* reflect the compactness of the global representation for features.

The obtained classification results are encouraging since the recognition process is based solely on motion data from optical flow. The spatial global features are derived by dividing the image vertically and horizontally into a set of $b \in 5, 10, 20$ bars. To explore the performance of the features using various classifiers, the decision tree is employed without any feature selection as it has its own embedded selection method of features based on information entropy. A classification rate of 85.95% is attained via the use of decision tree for 10 bars and 94.21% for 5 bars. Multi-class support vector machine is also employed in this experiment with a reported CCR of 89.67% for 10 bars and 90.08 for 20 bars. It can be observed that the nearest classifier regardless of its simplicity, feature selection, has proved useful in achieving higher recognition rates compared to other classifiers. The execution time for each classifier is shown in Table 2.

4.3 Action similarity matching

In order to illustrate the verification results for inferring the similarity level between two different human actions across all pairs, the receiver operating characteristics (ROC) are computed as depicted in Fig. 8. At the verification phase, all human actions from the constructed dataset are verified sequentially against each other

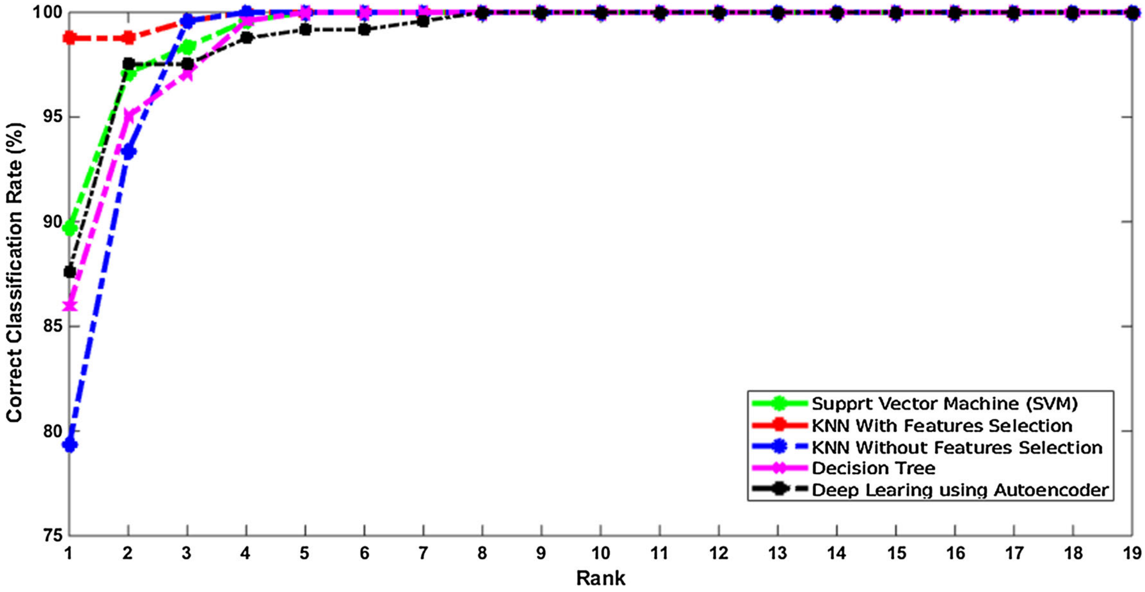


Fig. 7 Cumulative Match Score using different classifiers

Table 1 Effect of different spacial bars on the classification results

Classifiers	Bars		
	5	10	20
KNN1 with FS	97.93	95.45	95.04
KNN3 with FS	**98.76**	**95.86**	**95.62**
KNN5 with FS	97.52	95.04	95.04
KNN1 without FS	83.47	85.12	81.40
KNN3 without FS	79.75	85.12	79.75
KNN5 without FS	78.51	84.71	78.51
SVM	85.95	89.67	90.08
Decision tree	94.21	85.95	82.64

Bold values reflect higher CCR rates

Table 2 Classification time for different classifiers applied on the Weizmann dataset

	Bars		
	5	10	20
KNN			
$K = 1$	2.2497	2.2552	2.2618
$K = 3$	2.2524	2.2574	2.2634
$K = 5$	2.2555	2.2582	2.2646
SVM	0.1651	0.1664	0.1675
Decision tree	0.0915	0.0917	0.0996
Deep learning	0.2095		

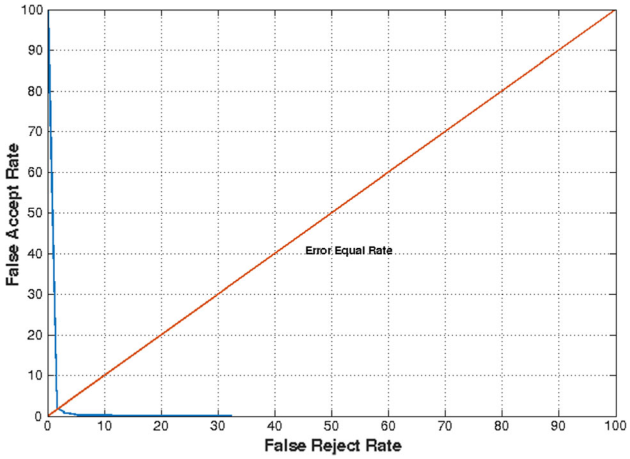


Fig. 8 Receiver operating characteristic (ROC) plot: verification results for similarity matching on the Weizmann dataset

checking if a given pair has the same class label or not. The matching process which is based on the Euclidean distance with a thresholding value described in the feature selection phase is employed to assess whether the two human actions have the same semantics. In order to estimate the False Acceptance Rate (FAR) projected against the False Rejection Rate (FRR), various thresholding values are being deployed. Using the human action signature derived from optical flow features using the histogram-based descriptor, the system reached a satisfactory equal error rate of 1.89%. Further, similarity matching is conducted via analysing the distribution between the distances for pairs belonging to the same classes versus different classes using Daughman decidability index metric [11]. The following decidability index values of 0.8205 and 1.6136 are reported for feature selection and raw features, respectively, during the intra- and inter-classes matching of instances. This clearly shows that the process of recognizing if two actions are the same based on pair matching is a more challenging task.

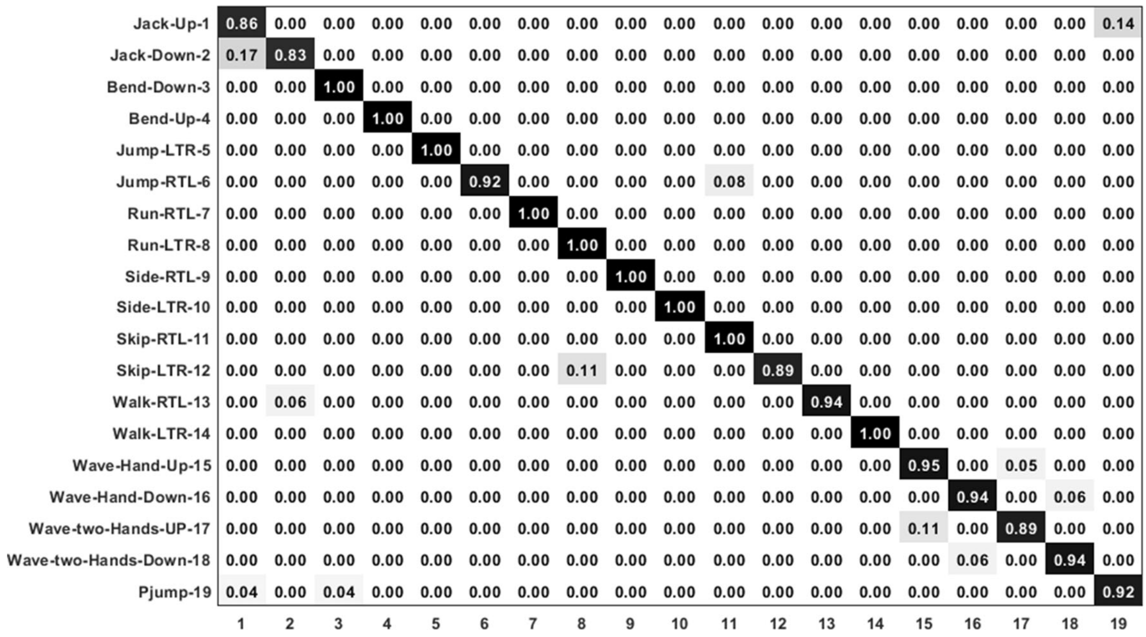
To better visualize the verification and separation results between the different human action classes, the confusion matrix is drawn in Fig. 9. The lighter squares signify greater separation values and thus better discriminability between different classes. The dark diagonal line is the zero distance when comparing a class to itself. The Euclidean distance is computed to deduce the separation level between two classes as the average between all matched pairs. As all features are already normalized during the preprocessing stage between 0 and 1. It is observed that some actions tend to be almost the same when using motion features as for the case of waving hands and *Pjump* and other events. Meanwhile, there are certain similarities between running, walking and side walking.

4.4 Mining for basic actions in complex scenes

Based on the basic actions detected from the Weizmann dataset, further experiments are conducted to detect these actions from different datasets with more realistic and complex scenes. We have manually annotated 1400 video sequences from the UCF101 [42] and KTH [40] datasets for basic actions. Afterwards, the classification procedure is performed on the dataset to search for basic actions using a predefined threshold which was set based on the experiments for the similarity matching. Figure 10 depicts the paradigm being deployed for the detection of basic actions using the proposed descriptor.

The evaluation process is based on the Recall and Precision using the obtained results compared against the manually annotated data. Results are summarized in Table 3 such that:

- **True Positive (TP)**: Both system and annotator detect the same action
- **False Positive (FP)**: The system does not detect the same action labelled by the annotator.
- **False Negative (FN)**: The system does not detect whilst the annotator detects a human action..



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Jack-Up-1	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14
Jack-Down-2	0.17	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bend-Down-3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bend-Up-4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump-LTR-5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump-RTL-6	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Run-RTL-7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Run-LTR-8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Side-RTL-9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Side-LTR-10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Skip-RTL-11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Skip-LTR-12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Walk-RTL-13	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00
Walk-LTR-14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Wave-Hand-Up-15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.05	0.00	0.00
Wave-Hand-Down-16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.06	0.00
Wave-two-Hands-UP-17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.89	0.00	0.00
Wave-two-Hands-Down-18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.94	0.00
Pjump-19	0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92

Fig. 9 Confusion matrix for human action recognition: results for cross-matching of different classes. *Lower values reflect higher discriminability*

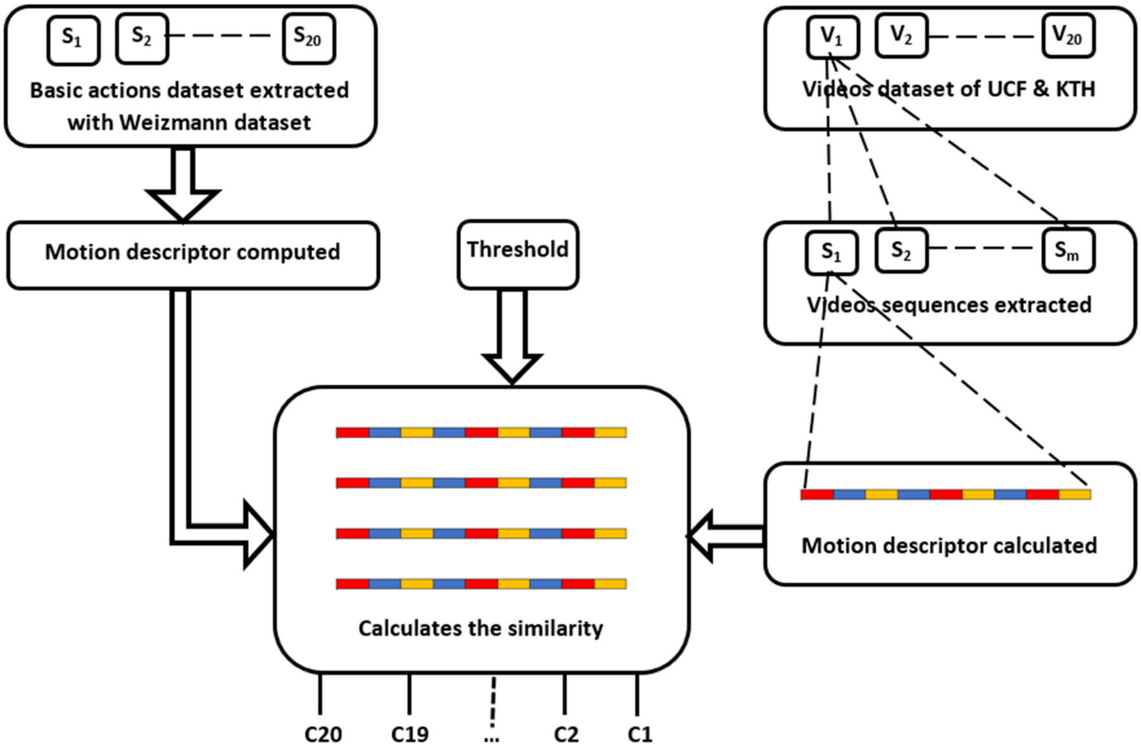

S1
S2
S20
Basic actions dataset extracted with Weizmann dataset
Motion descriptor computed
Threshold
V1
V2
V20
Videos dataset of UCF & KTH
S1
S2
Sm
Videos sequences extracted
Motion descriptor calculated
Calculates the similarity
C20
C19
...
C2
C1

Fig. 10 Decomposing video into simple action sequences using a motion descriptor

- **True Negative (TN)**: Both system and annotator do no detect a human action in the processed sequence

The estimated metrics are computed as follows:

- **Precision**: *Precision* or *Specificity* measures the proportion of negatives that are correctly identified as:

Table 3 Statistical results for the decomposition of videos

TP	TN	FP	FN	Precision	Recall	F1 Score
682	184	356	178	65.70%	79.30%	71.87%

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

- **Recall**: *Recall* or *Sensitivity* measures the proportion of positives that are correctly identified as:

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

- **F1 Score**: *F1 Score* or *F-measure* is computed as :

$$F1_{Score} = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{17}$$

4.5 Comparative analysis

Comparative analysis for the proposed method compared to existing approaches which are applied recently for human activity recognition on the Weizmann and UCF101 datasets is shown in Tables 4 and 5, respectively. The obtained results reflect the potency in dealing with the difficult area of human action recognition via decomposing into basic actions. For the same purpose and to enrich the comparison process, the work described by Dobhal et al. [13] is applied on Weizmann dataset. Their work is based on the representation of the human motion features in video by joining videos into a single frame called the binary motion image whilst deep learning is used the classifier. A correct classification rate of 87.60% is obtained using deep learning with autoencoder for 19 atomic classes and 97.11% for 10 classes. This is due in the first place to the small number of elements of each class and the difficulty in differentiating between the two models of two different movements as for the case of raising and lowering hands.

Table 5 Comparative results for UCF101 dataset

Method	CCR (%)
Our method: motion descriptors	70.00
Bag of words [42]	44.50
Spatiotemporal ConvNet [25]	65.40
Improved dense trajectories (IDT) [47]	85.90
IDT with higher-dimensional encodings[37]	87.90
Two-stream model (fusion by SVM) [41]	88.00
Long-term temporal convolutions[45]	92.70
TS-LSTM + temporal-inception [8]	94.10
Temporal segment networks [51]	94.20

4.6 Performance analysis

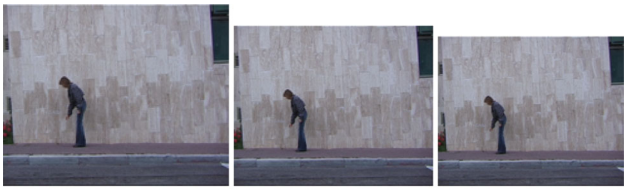
To evaluate the performance of the proposed optical flow descriptor for the classification of human actions, we have considered exploring two factors: frame drops and reduced resolution. For the first experiment, frame numbers are being dropped incrementally and we compute the correct classification rate for each round. Frames are dropped from all instances in the testing dataset and matched against the original training dataset. Table 6 expresses the performance relationship between the number of dropped frames to the classification rate. The system achieves an

Table 4 Comparative results for the Weizmann dataset

Method	CCR (%)
Our method: motion descriptors	98.76
Our method: deep autoencoder with 10 classes	97.11
Our method: deep autoencoder with 19 classes	87.60
Binary motion image and deep learning with 5 classes[13]	100.00
Multi-channel correlation filters [26]	97.80
Interest points + SIFT filters [34].	96.66
Binary motion descriptor [16]	95.81
Shape, motion and texture features [39]	94.44
Pose primitive [43]	94.40
Sequence alignment and shape context [3]	92.22
Hough transform-based voting framework [53]	92.20
Learning mid-level motion features [17]	90.50
Multiple features [31]	90.40
Spatial–temporal [35]	90.00

Table 6 Effect on frame drop for human action recognition

KNN	Frames dropped							
	0	1	2	3	4	5	6	7
$k = 1$	97.93	86.11	59.72	58.33	54.44	52.83	50.61	50.00
$k = 3$	98.76	79.16	59.72	57.33	44.98	44.79	44.44	44.04
$k = 5$	97.52	66.66	44.44	43.05	40.27	38.88	34.72	34.12



(a) 70%:101x126 **(b)** 60%:87x108 **(c)** 50%:72x90

Fig. 11 Different resolutions for the Weizmann dataset

acceptable success rate of 79.16% when dropping a single frame. However, a low recognition rate of 59.72% is reported when dropping two of the frames. This is because the classification is purely based on detecting motion from the consecutiveness nature of frames where dropping or missing frames can conceal these vital features.

To analyse the effect of resolution, we reduce the frame size for all data from 90 to 50% with decrements of 10% whilst the CCR is computed for every new resolution separately. It is common that in surveillance technology there is always low resolution. The produced resolution of the images is shown in Fig. 11. There are two experimental results being shown in Table 7. In the first one, the original subset of features is used for the reduced resolutions. The system shows that an acceptable recognition rate of 80% can be obtained even for the case of 116 × 144 . This shows indeed the potency of the proposed method for surveillance systems that rely mostly on poor quality resolution cameras. During the second experiment, the process of feature selection is applied separately at each level. Better results are obtained compared to the first experiment suggesting the need for devising an adaptive feature selection to cope with varying resolutions.

4.7 Features analysis

An exploratory analysis is performed in order to study the distribution of optical flow features and determine what motion cues are pivotal for the recognition of human activities. The components of the human action signature made within the histogram are assessed separately to investigate their contribution and recognition potency during the classification process. We have considered to evaluate 17,293 subsets of features within this empirical study such that every subset has an reported classification of 98.76% using the KNN classifier. All feature subsets are having a size of features ranging from 28 to 100. Choosing such a large number of subsets would ensure unbiased and accurate results for the analysis. The distributions and human action classification results of the different types of features are shown in Table 8.

The distribution results show a clear indication of what type of feature is rudimentary, but it does not offer a measure of its discriminatory potentials. Alternatively, the recognition significance of optical flow features is approximated via the use of the correct classification rate.

Table 8 Features analysis for human action recognition

Features	Distr (%)	CCR (%)
Local features	00.08	88.42
Global features—*temporal*	90.58	93.39
Global features—*spatial*	09.34	65.29

Table 7 Effect of reducing resolution for human action recognition

KNN	Resolution					
	100%:144 × 180	90%:130 × 162	80%:116 × 144	70%:101 × 126	60%:87 × 108	50%:72 × 90
Without FS						
$k = 1$	85.12	84.30	80.16	73.96	71.48	71.07
$k = 3$	85.12	79.75	79.34	72.72	69.83	67.76
$k = 5$	84.71	77.68	76.03	72.72	69.01	66.94
With FS						
$k = 1$	97.93	93.80	90.08	89.67	83.47	77.27
$k = 3$	98.76	89.66	89.25	88.84	82.23	74.38
$k = 5$	97.52	88.84	88.42	86.36	80.57	72.72

The obtained results show the importance of global temporal features which contribute with an almost 90% for the feature vectors whilst attaining a recognition rate of 93%. For the case of local features which describe the optical flow vectors without temporal or spatial information, marginal contribution of 0.08% is reported with a surprising correct classification rate of 88%. The combination of the local and global cues shows considerable influence to boost the correct classification rate for human action recognition.

5 Conclusions

The deployment of automated computer vision methods to recognize human activity is of central importance for many applications as security surveillance, sports analysis and human–computer interaction. In this study, a motion interchange descriptor is introduced for the extraction of visual features based on optical flow applied on a set of consecutive frames for the classification of human basic actions. A histogram of motion features is produced taking into consideration the local and global traits embedded within optical flow. Feature selection based on the nearness and proximity is performed to derive the most discriminative features. To evaluate the proposed descriptors for the recognition of human activities, extensive experimental results are conducted on two publicly available datasets including the Weizmann and UCF101, affirmed the potentials of the proposed approach to achieve a high correct classification rate of 98.76% and 70%, respectively, for basic human action recognition. The obtained results are in alignment with the early psychological studies reporting that human motion is adequate for the perception of human activities. Additional empirical evaluations are carried out to explore the performance of the introduced descriptor to handle different resolutions and frame rates.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv (CSUR) 43(3):16
2. Alfaro A, Mery D, Soto A (2013) Human action recognition from inter-temporal dictionaries of key-sequences. In: Pacific-Rim symposium on image and video technology. Springer, pp 419–430
3. Almotairi S, Ribeiro E (2014) Action classification using sequence alignment and shape context. In: The Twenty-Seventh International Flairs Conference
4. Asadi-Aghbolaghi M, Clapés A, Bellantonio M, Escalante HJ, Ponce-López V, Baró X, Guyon I, Kasaei S, Escalera S (2017) A survey on deep learning based approaches for action and gesture recognition in image sequences. In: 2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017). IEEE, pp 476–483
5. Bouchrika I, Carter JN, Nixon MS, Mörzinger R, Thallinger G (2010) Using gait features for improving walking people detection. In: 2010 20th International conference on pattern recognition (ICPR). IEEE, pp 3097–3100
6. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633–659
7. Chaudhry R, Ravichandran A, Hager G, Vidal R (2009) Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, pp 1932–1939
8. Chen M, Kira Z et al (2017) TS-lSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. arXiv preprint arXiv:1703.10667
9. Colque RVHM, Caetano C, de Andrade MTL, Schwartz WR (2017) Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Trans Circuits Syst Video Technol 27(3):673–682
10. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, vol 1. IEEE, pp 886–893
11. Daugman J (2004) How Iris recognition works. IEEE Trans Circuits Syst Video Technol 14(1):21–30
12. Dhulekar P, Gandhe S, Chitte H, Pardeshi K (2017) Human action recognition: an overview. In: Proceedings of the international conference on data engineering and communication technology. Springer, pp 481–488
13. Dobhal T, Shitole V, Thomas G, Navada G (2015) Human activity recognition using binary motion image and deep learning. Procedia Comput Sci 58:178–185
14. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
15. Fan B, Ding Z, Gao W, Long T (2014) An improved motion compensation method for high resolution UAV SAR imaging. Sci China Inf Sci 57(12):1–13
16. Fangbemi AS, Liu B, Yu N, Zhang Y (2018) Binary proximity patches motion descriptor for action recognition in videos. In: Proceedings of the 10th international conference on internet multimedia computing and service. ACM, p 17
17. Fathi A, Mori G (2008) Action recognition by learning mid-level motion features. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
18. Feng Y, Ji M, Xiao J, Yang X, Zhang JJ, Zhuang Y, Li X (2015) Mining spatial-temporal patterns and structural sparsity for human motion data denoising. IEEE Trans Cybern 45(12):2693–2706
19. Fortun D, Bouthemy P, Kervrann C (2015) Optical flow modeling and computation: a survey. Comput Vis Image Underst 134:1–21
20. Gentile C, Li S, Kar P, Karatzoglou A, Etrue E, Zappella G (2016) On context-dependent clustering of bandits. arXiv preprint arXiv:1608.03544

21. Horn BK, Schunck BG (1981) Determining optical flow. In: 1981 Technical symposium east. International Society for Optics and Photonics, pp 319–331
22. Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194
23. Janschek K, Tchernykh V, Dyblenko S (2005) Integrated camera motion compensation by real-time image motion tracking and image deconvolution. In: Proceedings, 2005 IEEE/ASME international conference on advanced intelligent mechatronics. IEEE, pp 1437–1444
24. Kar P, Li S, Narasimhan H, Chawla S, Sebastiani F (2016) Online optimization methods for the quantification problem. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1625–1634
25. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732
26. Kiani H, Sim T, Lucey S (2014) Multi-channel correlation filters for human action recognition. In: 2014 IEEE international conference on image processing (ICIP). IEEE, pp 1485–1489
27. Kliper-Gross O, Gurovich Y, Hassner T, Wolf L (2012) Motion interchange patterns for action recognition in unconstrained videos. In: European conference on computer vision. Springer, pp 256–269
28. Koohzadi M, Charkari NM (2017) Survey on deep learning methods in human action recognition. IET Comput Vis 11(8):623–632
29. Lara OD, Labrador MA (2013) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor 15(3):1192–1209
30. Li S, Karatzoglou A, Gentile C (2016) Collaborative filtering bandits. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 539–548
31. Liu J, Ali S, Shah M (2008) Recognizing human actions using multiple features. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
32. Martínez F, Manzanera A, Romero E (2012) A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance. In: Wang FL, Lei J, Lau RWH, Zhang J (eds) Multimedia and signal processing. Springer, Berlin, pp 267–274
33. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 104(2):90–126
34. Moussa MM, Hamayed E, Fayek MB, El Nemr HA (2015) An enhanced method for human action recognition. J Adv Res 6(2):163–169
35. Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 79(3):299–318
36. Oshin O, Gilbert A, Bowden R (2014) Capturing relative motion and finding modes for action recognition in the wild. Comput Vis Image Underst 125:155–171
37. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput Vis Image Underst 150:109–125
38. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
39. Rahman S, See J, Ho CC (2015) Action recognition in low quality videos by jointly using shape, motion and texture features. In: 2015 IEEE international conference on signal and image processing applications (ICSIPA). IEEE, pp 83–88
40. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol 3. IEEE, pp 32–36
41. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Proceedings of the 27th International Conference on Neural Information Processing Systems, vol 1. MIT Press, Cambridge, MA, USA, pp 568–576
42. Soomro K, Zamir AR, Shah M (2012) Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402
43. Thurau C, Hlaváč V (2008) Pose primitive based human action recognition in videos or still images. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
44. Tymoshchuk PV (2009) A discrete-time dynamic k-winners-take-all neural circuit. Neurocomputing 72(13–15):3191–3202
45. Varol G, Laptev I, Schmid C (2018) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510–1517
46. Vishwakarma S, Agrawal A (2013) A survey on activity recognition and behavior understanding in video surveillance. Vis Comput 29(10):983–1009
47. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
48. Wang J (2010) Analysis and design of a *k*-winners-take-all model with a single state variable and the heaviside step activation function. IEEE Trans Neural Netw 21(9):1496–1506
49. Wang J, Cherian A, Porikli F (2017) Ordered pooling of optical flow sequences for action recognition. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 168–176
50. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4305–4314
51. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. Springer, pp 20–36
52. Weinland D, Boyer E. (2008) Action recognition using exemplar-based embedding. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–7
53. Yao A, Gall J, Van Gool L (2010) A hough transform-based voting framework for action recognition. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2061–2068
54. Yeffet L, Wolf L (2009) Local trinary patterns for human action recognition. In: 2009 IEEE 12th international conference on computer vision, pp 492–497
55. Yi Y, Cheng Y, Xu C (2017) Mining human movement evolution for complex action recognition. Expert Syst Appl 78:259–272
56. Zhu W, Hu J, Sun G, Cao X, Qiao Y (2016) A key volume mining deep framework for action recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1991–1999
57. Zhu Y, Nayak NM, Roy-Chowdhury AK (2013) Context-aware activity recognition and anomaly detection in video. IEEE J Sel Top Signal Process 7(1):91–101