Applied Intelligence (2021) 51:690–712
https://doi.org/10.1007/s10489-020-01823-z

A combined multiple action recognition and summarization for surveillance video sequences

Omar Elharrouss[1] · Noor Almaadeed[1] · Somaya Al-Maadeed[1] · Ahmed Bouridane[2] · Azeddine Beghdadi[3]

Published online: 26 August 2020

Abstract

Human action recognition and video summarization represent challenging tasks for several computer vision applications including video surveillance, criminal investigations, and sports applications. For long videos, it is difficult to search within a video for a specific action and/or person. Usually, human action recognition approaches presented in the literature deal with videos that contain only a single person, and they are able to recognize his action. This paper proposes an effective approach to multiple human action detection, recognition, and summarization. The multiple action detection extracts human bodies' silhouette, then generates a specific sequence for each one of them using motion detection and tracking method. Each of the extracted sequences is then divided into shots that represent homogeneous actions in the sequence using the similarity between each pair frames. Using the histogram of the oriented gradient (HOG) of the Temporal Difference Map (TDMap) of the frames of each shot, we recognize the action by performing a comparison between the generated HOG and the existed HOGs in the training phase which represents all the HOGs of many actions using a set of videos for training. Also, using the TDMap images we recognize the action using a proposed CNN model. Action summarization is performed for each detected person. The efficiency of the proposed approach is shown through the obtained results for mainly multi-action detection and recognition.

Keywords Video summarization · Human action recognition · CNN · HOG · TDMap

1 Introduction

Currently, video technologies are facing several challenges and difficulties, mainly attributed to the extraction of information in real-time from a large number. The extracted information can be useful to identify and detect many events that can help in many analyses, such as abnormal events and people's behavior, as well as to predict events that usually happen in the scenes. Recently, a number of researchers focused their studies on finding effective techniques to summarize useful information from videos. This research field is essential for the improvement of video surveillance systems that require large storage space and complex data analysis, considering that data is captured 24 hours a day and 7 days a week. Therefore, summarization of video data is required in such systems to simplify data analysis, facilitate information storage, and to improve the access to each time video. The summarization process can also be related to the type of scene (private or public) where the data analysis depends on whether the scene is dynamic or static, as well as whether it is crowded or uncrowded. Since the summarization process should consume less time for processing and less space for storage it might require pre-processing to enhance the process without losing any information before the feature extraction task [1–7]. Video summarization methods are generally classified into two main categories: scene-based (static, dynamic) and content-based. From a video containing changing scenes, static-based methods consist of selecting keyframes and dynamic-based methods consist of selecting short video clips. Since the scenes can change, the cameras can move, the summarization, in this case, is carried out by determining the video sequences (shots) that represent the same scenes [8–15]. This allows the keyframe to be selected using extracted features and appropriate clustering methods. The selection can lead to some redundant frames, called meaningless frames, requiring an operation for their removal. On the other

✉ Omar Elharrouss
elharrouss.omar@gmail.com

Extended author information available on the last page of the article.

hand, in content-based methods, the summarization is made using semantics and content of the video. Several types of summarizations can be found in this category, including motion-based, event-based, and action-based methods. Figure 1 illustrates the classification of video summarization, including the subcategories of each method. Content-based video summarization is based on video content that requires pre-processing, for example, motion-based methods utilize the results of motion detection or the trajectories of objects to summarize the video. In addition, the summarization using this technique can be static(keyframes) or dynamic(short clips) [15–22]. The challenge in video summarization is related to the video understanding and the classification of important sequences of the video, and the importance depends on the types of actions or objects that must be summarized. The complexity and variety of scenes in a video making the foundation of generic methods impossible [40–56].

This paper proposes an action-based video summarization by recognizing actions made by each person in the scene and then summarizing these actions at the end of the video. So it is an action recognition and summarization approach where human body actions are first detected using a proposed background-subtraction-based approach and then recognized. The proposed method provides the allows detecting and recognizing many human body actions, unlike other methods that capture the action of only one person in the scene. For that, two methods are proposed the first one uses the Cosine similarity measure of the HOGs of Temporal Difference Map (TDMap) and the second one is a CNN classification of actions from TDMap images. The rest of the paper is organized as follows. Section 2 provides an overview of research studies that have been conducted in the area of video summarization. Section 3 details our proposed approach including action detection and recognition. The results of the implemented work are discussed and analyzed in Section 4. Finally, conclusion and future works are discussed in Section 5.

2 Related works

The growth of video technologies has led to the creation of efficient tools to manipulate this type of data. Summarization aims to generate a short version of a video as a representation, using keyframes of important subsequences. This summarization provides a rapid view of the information contained in a large video. It also provides a good evaluation for users of the video and provides knowledge regarding the topic and the most important content in the video. Considering the information contained in each video, many methods have been developed using several techniques. Each technique summarizes the video using a specific feature, such as trajectories, moving objects, abnormal detection, and many others. These categories of techniques can be classified into two general categories, scene-based (i.e., static [1–7, 9, 17, 21, 22], dynamic [8, 12, 13, 16, 18–20] and content-based approaches, and the content-based approaches can be further decomposed into three types related to the content of the video including motion-based [10–16, 20], action-based [21, 22] and event-based [11, 15, 17–19], as shown in Fig. 1. Video summarization is a short version of the longer video sequence. The static video summarization is a collection of frames (keyframes) selected from the original video. The proposed approaches extract the keyframes using many features [1, 2]. In general, a video contains many parts, called shots, which represent different sequences. Each sequence represents a scene captured by a fixed or moving camera. The general idea of these methods is to classify these shots using clustering techniques [2], after which keyframes from these shots would be extracted. The meaningless frames that are similar are removed [3]. In the same context, [4] proposed a video keyframe extraction method using Jensen– Rényi divergence (JRD), Jensen — Shannon divergence (JSD), and Jensen– Tsallis divergence (JTD) to measure the difference between neighboring video frames, segmenting a video clip into shots and then possibly into sub-shots, and choosing keyframes in each

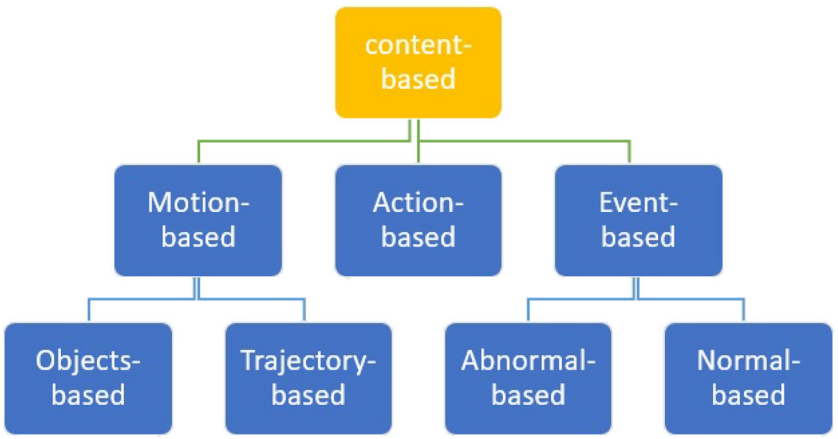

Fig. 1 Video-summarization-based methods

shot. This is computationally inexpensive and yet effective. In [5], authors utilized sparse dictionary selection to extract keyframes directly, developed an online version to summarize the video in real-time and provided a guide for users to obtain a summary with an appropriate length. Video summarization is a reduced representation for fast video retrieval. In another work [6],a temporal- and spatial-driven approach was proposed. In this study, Optimum-Path Forest (OPF) clustering was used to automatically determine the number of keyframes and extract them to compose the final summary. To generate a video summary, [7] used a graph-based hierarchical clustering method. Called HSUMM, the proposed approach adopts a hierarchical clustering method to generate a weight map from the frame similarity graph in which the clusters can easily be inferred. In the same context and to generate an efficient summarization, the authors in [8] proposed a divide-and- conquer-based framework. In this work, the original video data is divided into shots, where an attention model is computed from each shot in parallel. Viewer attention is based on multiple sensory perceptions. From the deployment of surveillance cameras, a large amount of data is produced, and the intelligent systems can extract several types of information from videos. A monitoring system can analyze videos and extract any information regarding the content of the covered areas, including information about objects (motion, action, and trajectory) and the event that happened in the scene. This information can help any system to understand the content of the video [9]. According to the purpose of the system and the tasks to be handled, the system needs to learn and extract just the needed information, thus, video summarization is a good solution to abstract the content of any video [10, 11]. In the following, we describe each category of methods that utilize content-based summarization, including motion-based (object-based and trajectory-based approaches), action-based and abnormal-event-based methods. Detection of moving objects provides a good understanding of the content of each scene covered by cameras for video surveillance systems. The information about motion also represents an effective feature for video summarization. In some methods, the motion of objects is used to summarize the video content. Based on the extraction of moving objects in video sequences, [12] combined adaptive fast-forwarding and content truncation to summarize the content of videos.

In another work [10], the authors used background subtraction, clustering techniques, and a noise algorithm to summarize the content of videos. In [13], the authors ' surveillance video was converted into a temporal domain image (temporal profile). This technique makes it easy for human operators to search within a long video. Most video summarization methods use a single view captured by a single camera. Some researchers try to use the advantage of multiple views of a scene covered by several cameras to summarize videos. For example, Panda et al. [14] exploited multi-view videos of a scene in video summarization using the sparse selection as selected shots. The trajectories can be a good solution for the recognition and summarization of the activity of objects during their presence in the scene. A good number of proposed methods use the object trajectory that is obtained by the tracking operation [15]. The authors used trajectories for abnormal event detection, which in turn was exploited for the generation of video summaries. Similarly, a framework has been developed for multiple-scene understanding and scene activity summarization [16]. The authors proposed a representation, using motion flow, of shared areas of interest in scenes covered by multiple cameras to understand the activity and behaviors in each scene. Objects ' trajectories can be an efficient solution for many situations in video surveillance and a good technique for activity understanding and video summarization tasks.

Video surveillance systems play an important role in ensuring people 's safety. In addition, the detection of abnormal events and unusual activities can be useful for these systems. The summarization of these events and activities is good support for each system to learn and understand the content of the covered area. Consequently, based on activity and event detection, many methods have been proposed for video summarization [17]. An overview of video summarization methods based on abnormal event detection can be found in [18]. The main steps of these methods are the detection of unusual events followed by their summarization. For example, in [19], the authors proposed a visual surveillance briefing system (VSB) that retrieves abnormal events using object appearances and motion patterns and adopted a video summarization algorithm. Some authors have proposed patch-based methods to model the key regions in the scene and learn the normal activity patterns in it [20]. After that, based on previous feature results, the unusual activities are detected. As a final step, a summarization of all abnormal activities is developed to create a short-period summary from a long video. In the same context, [15] proposed a novel approach for large-scale surveillance video summarization on the basis of event detection. The detection of trajectories of vehicles and pedestrians has been achieved. Using these features, abnormal event detection is developed. The video summarization step exploits the event detection results to summarize the short period that contains unusual activities in the scene.

Human action recognition is an important task for many applications, including video surveillance systems, video indexing and retrieval, sports applications, and multimedia. Action detection, recognition, and summarization can be

exploited to support many other tasks. For example, in sports applications, recognizing and understanding player poses allows the judges to make a good decisions in the case of player fouls, especially for football games, which require precise decisions in many situations. Several methodologies have been proposed for the summarization of actions. [21] proposed a sports pose summarization method for self-recorded RGB-D videos. The authors chose games to test the performance of the methods because they contain a succession of complex actions. The extended version of this approach uses deep neural networks to extract two types of action-related features and classify video segments into interesting or uninteresting parts [22]. The authors proposed a method to recognize actions, which can lead to a good selection of meaningful informative summaries. In addition, there is a reciprocal task that recognizes the actions of the generated video summary. For that, the authors used the latent structural SVM framework combined with an algorithm for inferring the action.

Video summarization methods in [1–9] select key frames or short clips without analyzing the content of the video can lose information's if the purpose of the summarization is not specified. Also, for these methods the summarization made on the videos which contains different scenes changing during the video periods like movies video. For that the summarization based on the scene's variation still insufficient if the videos contain some important information.

For content-based approaches, video summarization is performed for some cases that specify the goal of summarization like summarization of abnormal event that can happened in a monitored scene [11, 15, 17, 18]. But find a generic method that can handle all event categories still a difficult task. For action-based summarization method in [21, 22] which there are a few research papers for this goal, we find most methods limited to the summarization of sport action. Also, the summarization of multiple actions in the same time is not processed.

2.1 Action recognition

In literature, many methods have been proposed to classify human actions [3, 23–31, 33–61]. The proposed methods can be split into three categories: motion-based methods, appearance-based methods, and space-time-based methods. Motion-based methods consist of computing parametric and generic optical flows before comparing results with motion templates. For appearance-based methods, images' motion history is extracted to be compared with the active shape models. In space-time approaches, space-time features with the training results are used in the space-time domain. In the same context, authors in [23, 24] used the concept of Compact Descriptors for Visual Search (CDVS). The use of local features and global data structure provided by CDVS is useful when it comes to real-time feature extraction in real-time especially with the use of computing optimization. Dasari et al. [25] classified human actions by tracking CDVS feature trajectories of the human body. Authors in [26] start by selecting regions (patches) in the video that can be described as actions. Then, they generate boxes contain ing the detected motions and each one of these boxes a discrimination score assigned. For action recognition, the authors applied a clustering technique to each box to identify different actions.

El-Henawy et al. [27] proposed a technique for human action recognition using fast HOG3D and Smith-Waterman of the partial shape matching of each frame. First, the foreground of video subsequences is extracted from the input stream. Then, the keyframes of the current subsequence are blended before the extraction of the contour of the resulting frame. To classify the HOG3D features, the author utilizes a non-linear SVM decision tree.

Using human motion for action recognition purposes, Xu et al. [28] exploited wearable sensors to extract human motion using natural physical properties. Extracted features are used to classify related actions. Human action recognition from video surveillance data can be viewed from different angles. In particular, 2D analysis of action recognition for human-computer interaction requires a good exposition of the human body for video acquisition. Zhang et al. [29], propose a new algorithm which starts with a pre-training phase based on synthetic data to extract view-invariance between 3D and 2D videos. And to encode extracted trajectories from 3D videos, they introduce a new feature named as 3D dense trajectories.

For video surveillance systems, human body is incomplete in some cases, which represent a challenge for human action recognition. The proposed methods suffer from some limitations and especially in the case of occlusions and highly crowded scenes. Indeed, it is rather hard to detect and recognize multiple human bodies in crowded scene. One solution to overcome these limitations is to apply some pre-processing and learning process in order to cope with the various occlusions and crowded scenarios. Furthermore, it is worth mentioning that to the best of our knowledge there is no publicly available datasets for person detection in highly crowded scenes that could be used for the learning process.

There are other alternatives to handle this problem like the use of a Kinect camera [57], that can be helpful for recognition of the action as well as the summarization of it. But, the occlusion and action recognition in a crowd scene still represent a challenge [58, 62–64].

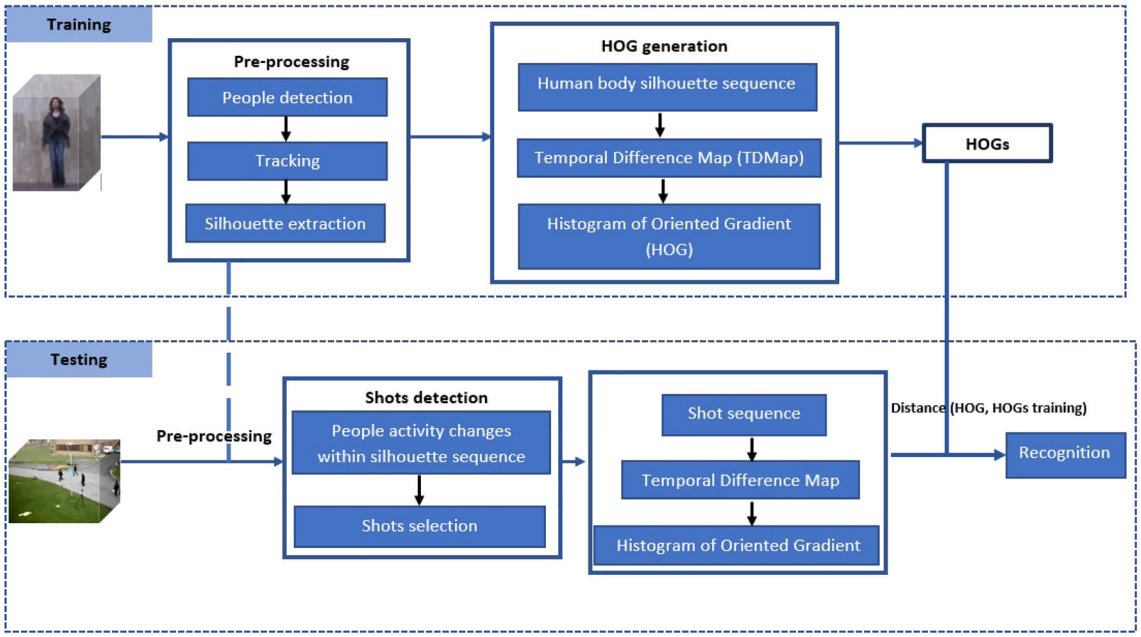

Fig. 2 Flowchart of the proposed method

3 Proposed method

In this paper, we propose a new approach for a combined multiple human action recognition and summarization technique. In our method, we start with detecting human bodies using a proposed background-subtraction-based approach. Then, each one was tracked separately to generate a corresponding video sequence of each person during his presence in the scene. We designed the training part to represent all categories of human actions by a set of Histograms of Oriented Gradient (HOG) of the Temporal Difference Map (TDMap) that represents the motion history of the target. For the action recognition step, we extracted from the scene a sequence of each moving person. Then, shots representing homogeneous parts of the sequence were selected using a similarity histogram between frames. Peaks of the histogram represent the transactions in the sequence, a shot is defined by the subsequence between each pair of peaks. Next, each shot was used to identify the considered action based on a training set which is a set of HOGs of each action. The action classification is performed using two methods: (1) Cosine similarity measure for comparing the HOG of the current action h those of the training set. (2) convolutional neural network (CNN) model to classify actions using TDMap images as input.

After the recognition stage, the summarization of actions from the scene is performed by representing the timeline of the actions for each person in the scene from each shot (representing an action). The flowchart of our proposed method for human action detection is depicted in Fig. 2. Recognition and summarization steps using detected human body sequences are detailed in Fig. 3.

3.1 Pre-processing

Existing methods for action recognition are designed to detect human actions from a scene involving one person. In this paper, the detection, and recognition of multiple human bodies is proposed. The idea is to extract the silhouette of each person during his/her presence in the scene. Next, we use motion detection using a background-subtraction-based method to identify all the persons present in the scene and generate detection masks of detected objects.

For the background subtraction method, we start by initializing the background model using the N first frames from the video, based on the decomposition of each frame into blocks of 16×16 pixels. After the generation of the background model, the persons are detected by using background subtraction and object segmentation in each frame of the video.

Based on detected masks, each person is tracked through a bounding box. In this work, the Kalman-based tracker [44] is used, which applies the Kalman filter to predict the centroid of each track in the current frame and update its bounding box accordingly. While most methods only track one object at a time, the extended version of [44] is tailored to track multiple moving objects; During their presence in the scene, the silhouette of each detected person is extracted to form a new sub-sequence of the considered silhouette.

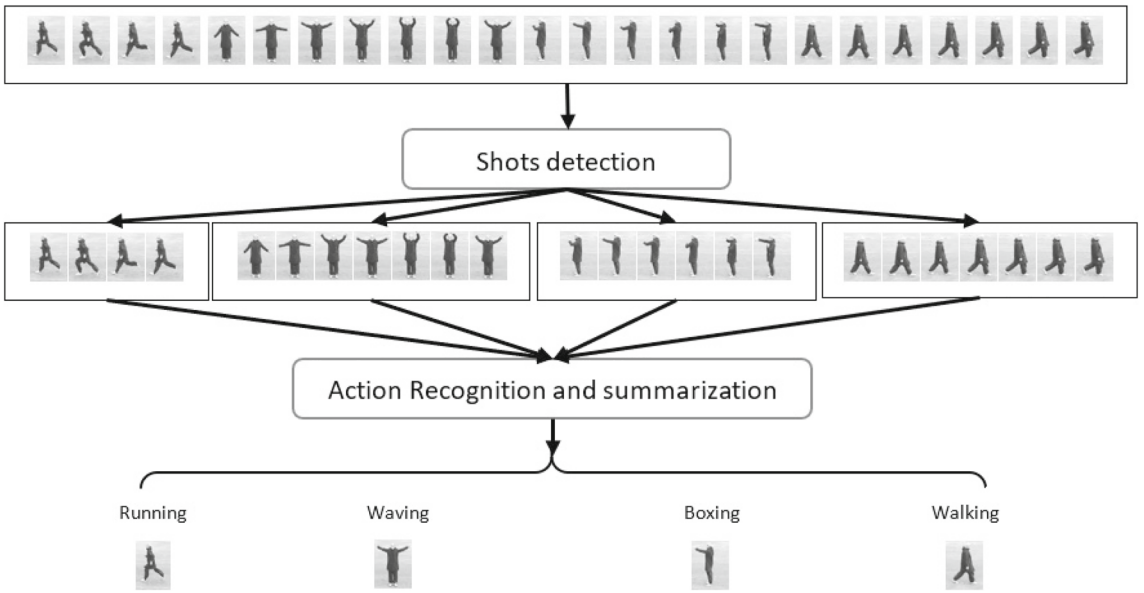

Fig. 3 Summarization steps based on recognized actions

Figure 4 illustrates the proposed method for motion detection, tracking until Oriented Gradient images extraction.

3.1.1 Motion detection

Before staring the gait recognition process and in order to ensure proper detection of the silhouette of the human body moving in the scene, a background-subtraction-based method is proposed. Background subtraction methods are the most used techniques for motion detection. The main operation for these methods is the background modeling which consists of extracting the unchanged pixels and region during the video. For that, we proposed a method for modeling the background by computing the similarity between blocks during a small period of the video representing here by the 100 first images of the video.

The modeling starts by dividing each frame into $w \times w$ blocks, then computing the Sum of Similarity (SS) of each consecutive block during T frames of the video. The SS values are computed using the following expression:

$$SS_{b(i,j)} = \sum_{t=1}^{T-1} \sum_{i=1}^{w} \sum_{j=1}^{w} \cos ine(I_t^{(i,j)}, I_{t+1}^{(i,j)}) \quad (1)$$

Where $b(i,j)$ represents the block background of the coordinates (i,j). Cosine similarity defined in [30] for computing the similarity between two vectors where the result is the interval [0,1] for the positive values. The cosine similarity between two vectors a and b is computed by the following expression:

$$\cos ine(a,b) = \frac{\sum_{i}^{n} a_i b_i}{\sum_{i}^{n} a_i^2 \sum_{i}^{n} b_i^2} \quad (2)$$

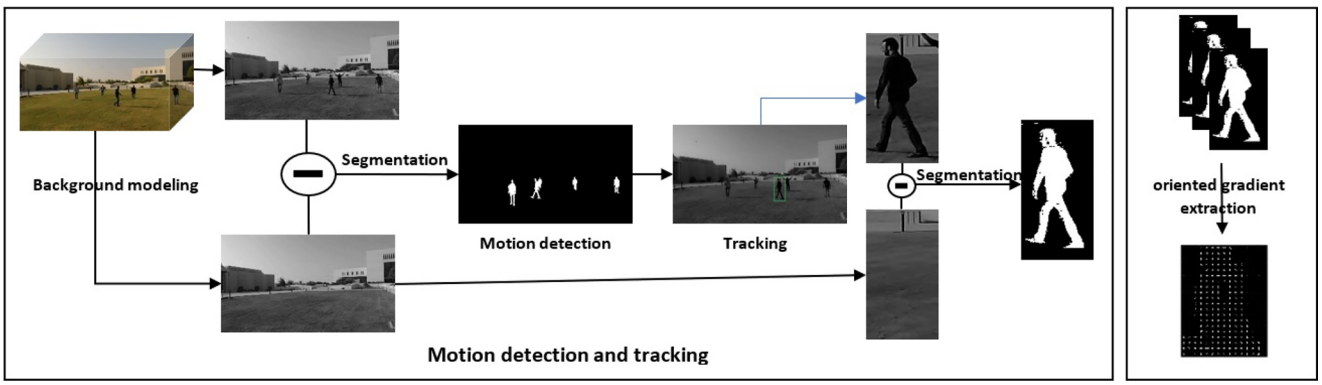

Fig. 4 Pre-processing steps for extracting each human silhouette

Where a and b here represent blocks of two consecutive frames.

The background model is generated using the *SS* value of each block. By collecting the maximum values of the sum of similarity of each block *(i,j)*. Regions of the blocks that did not change a lot during the 100 frames will have the most significant values because the value is 1 where two clocks are similar. The generated background model is defined based on *SS* values by the following expression:

$$B^{(i,j)} = Argmax\{SS_{(i,j)}\} \tag{3}$$

After the generation of the background model, the background subtraction is the next step to subtract the background from each current frame of the video suing absolute difference. Then, based on the subtraction results, a segmentation operation is performed to classify the pixels belong to the background and those belong to the foreground or to the moving objects. This operation uses a threshold where the most method tests a set of thresholds that after that choose the one that gives the best results. In this paper we propose a segmentation method for selecting this threshold adaptively using the exponential function of the absolute difference between the current frame and background frame:

$$T(i,\ j) = 1 - e^{-|I(i,j)-B(i,j)|} \tag{4}$$

Where T values have to be in the range of [0,1] and It is the current frame and *Bt* denotes the background image.

The threshold value converges to 0 when the background subtraction result goes to 0, and the threshold values tend to 1 when the background subtraction value is significant.

The computation of moving object at each time in the video represented by a binary images is performed using the selected threshold. The binary frame at time t of the video is computed using the following expression:

$$D(i,\ j) = \begin{cases} 255 \ if\ T(i,\ j) \simeq 1 \\ 0 \qquad otherwise \end{cases} \tag{5}$$

3.2 Data preparation

The training phase aims to select a number of features from the training videos. First, a pre-processing is applied to the training video to select the silhouette of the human bodies. For the selection of features, we applied the Histogram of oriented gradient (HOG) and we computed the Temporal Difference Map (TDMap) between frames (each pair of consecutive frames) for each subsequence. TDMap is used to efficiently generate motion history of object s' movement in all video regions.

The existing method uses the entire video region to extract the MHI of person action, which is not tolerated when the scene contains many people acting. For that, we used just the region where there motion. So our representation of data corresponds to the sequences of human body regions during his motion in the video. The new data is just an extraction of the regions containing the human body and not all the scenes. Accordingly, we cannot extract the trajectories of the body or the MHI because we do not use the entire video region. An example of the generated data and the corresponding TDMap is represented in Fig. 5. Human action history can have different structures from one activity to another. And so, we use HOG to extract information related to the structure of each action obtained from the oriented gradient representation. Then, the Histogram of Oriented Gradient is computed and the corresponding set of histograms for each action is collected. The same process is reproduced for all actions.

3.3 Shots detection

After the extraction of human silhouette from the original video, we obtain a sequence of silhouettes of each person in the scene. In this sequence, a person can perform many actions : the person can be in a walking state before starting to run or wave his hand. Transitions between actions change the appearance in the sequence. To detect the changes, we traced the histogram of similarity values between each pair of consecutive frames in the sequence using a cosine similarity measure defined in [30] and in the (2).

The peaks in the histogram represent the change between each action. Thus, using a smoothing operation and local maxima selection through a threshold, we extracted the subsequence between these pics that represents shots. The threshold value is taken to be 0.15 after an extension experimental evaluation. Each shot in the same sequence is exploited to recognize the action in it.

3.4 HOG based recognition

For the training data, we used the Weizmann, KTH, and UCF-ARG datasets, since they involve several human actions, namely, walking, running, hand waving, jacking, jumping, bend and side. When it comes to human interactions we used videos form UT-Interaction and INRIA XMAS (IXMAS) datasets for training. Figure 5 represents the flowchart of our training analysis; Fig. 5a represents the computation of HOGs of one subsequence. Therefore, for each action, we computed a set of HOGs that represent each action. Figure 5b illustrates the HOGs of four categories of videos, and each video represents an action. After detecting the human's body silhouette and extracting the shots for each detected person, the histogram of oriented gradient of the temporal difference map between each pair of consecutive images is computed. Then, the HOG within each shot is compared to all HOGs formed in the training phase. The comparison is made using the

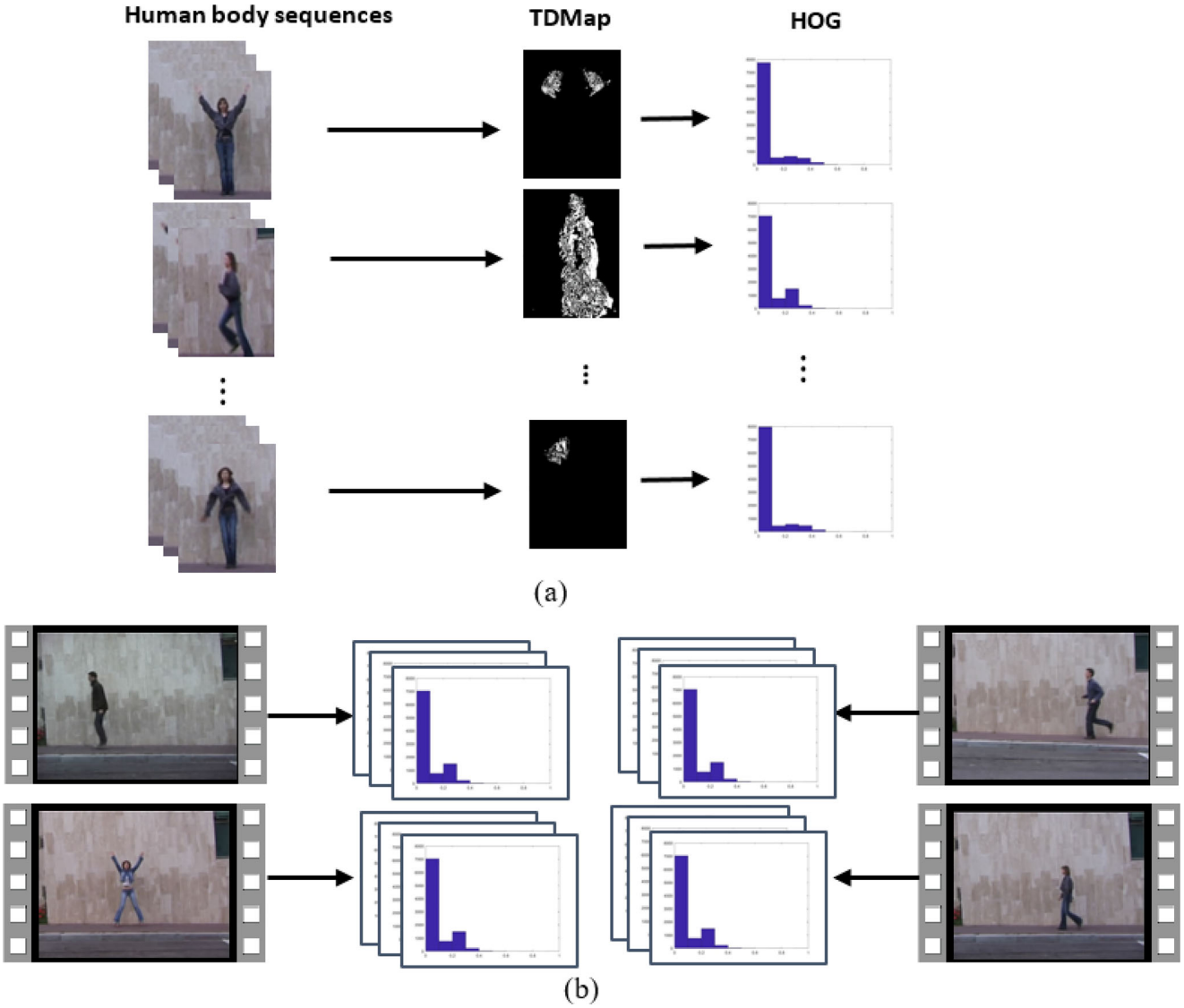

Fig. 5 Training process with the generation of HOGs of each action. **a** Histogram of oriented gradient of one video training of an action. **b** HOGs of each action video

distance computation between histograms, and the smallest is selected to detect each shot. The recognition can be formulated as follows:

$$Action\_index = Argmin\{dist(HOGs_{training}, HOG_{current})\} \quad (6)$$

The distance between histograms used in this paper is the same in (2) and defined in [30]. Here, A and B are two vectors representing histograms, a_i and b_i are histogram values.

The recognition of such action is made by comparing the computed distance between HOG of the current action and all HOGs of the training phase. Then the minimum distance indicates the real action.

3.5 CNN-based recognition

Depending on the application, the selection of the optimal CNN architecture is challenging. The proposed deep-learning-based approach involves the preprocessing of action videos before feeding them to the convolution neural network. The preprocessing consists of extracting the target region that contains human bodies in action, followed by an extraction of TDMap and then resizing the data before creating NumPy. A Convolution Neural Network (CNN), which is a supervised learning with multistage deep learning network, is implemented. CNN could learn multiple stages of invariant features from input images. Convolution and pooling are the main layers in a CNN

model. Any complex CNN can be constructed with a convolution-pooling combination.

The architecture of our model, as illustrated in Fig. 6, composes of two convolution-pooling units, with six convolutional layers and four MaxPooling layers, one flattened layer and two fully connected layers. The output layers comprise ten neurons that represent the number of actions. We introduced convolution neural network with the following notations: I(x,y) as an input image with size of $x \times y$ and d the temporal depth; Conv(x,y,f) is the convolutional layer and pooling Mpool(x,y,k) where x and y are image dimension, f number of channels, and k number of kernels. PReLUs indicates Parametric Rectified Linear Unit, FC(n) is a fully connected layer with n neurons, and D(r) is a dropout layer with a dropout ratio r. Using the notations, the proposed CNN model can be described as follows:

I(120,120,1), conv(119,119,32), conv(118,118,32), Mpool(59,59,32), conv(58,58,64), conv(57,57,64), Mpool(28,28,64), conv(27,27,128), Mpool(13,13,128), conv(12,12,128), Mpool(6,6,128), flatten(2304), FC(128),D(0.65), FC(number of identities).

The input of our system is a TDMap image with a resolution of 120×120 pixels. For the training and testing we used the preprocessed data from KTH, Weizmann, and UCF-ARG datasets. The model is trained using CrossEntrpy with a batch size of 128 examples, Adam as optimizer with a learning rate of 1e-3.

As an activation function we use the Parametric Rectified Linear Unit (PReLU), which is a generalized parametric formulation of ReLU. This activation function the parameters of rectifiers are learned adaptively and improves the accuracy with a negligible extra computational cost [53]**.** Only positive values are fed to the ReLU activation function, while all negative values are set to be zero. PReLU assumes that a penalty should be applied for negative values, and it should be parametric. The PReLU function can be defined as :

$$f(y_i) = \begin{cases} y_i & if y_i > 0 \\ a_i y_i & if y_i < 0 \end{cases} \tag{7}$$

Where a_i controls the slope of the negative part. When a_i = 0, it operates as a ReLU; when a_i is a learnable parameter, it is referred to as a Parametric ReLU (PReLU). If a_i is a small fixed value, PReLU becomes LReLU (a_i = 0.01). As shown in [53], PReLU can be trained using the backpropagation concept.

3.6 Summarization of actions

Once the detection and recognition of action is completed , the summarization is carried out by recording each action made by each person. The shot detection is the operation of splitting the frames of each action from the succession of actions made by each person. So, summarized actions of each person are represented by a timeline from each shot. The representation of the summarization can be performed also by selecting a frame that represent each action. Herin, we use the two representations.

Before choosing the frame from the shot, we performed a shot detection operation which defines the similar frames in each shot. So that the frame in each shot represents the same action. Then, a random selection of frames is performed since all frames in the shots represent the same action.

In this work, a video summary is defined as a concatenation of action labels and keyframes of the video where the actors perform a specific action.

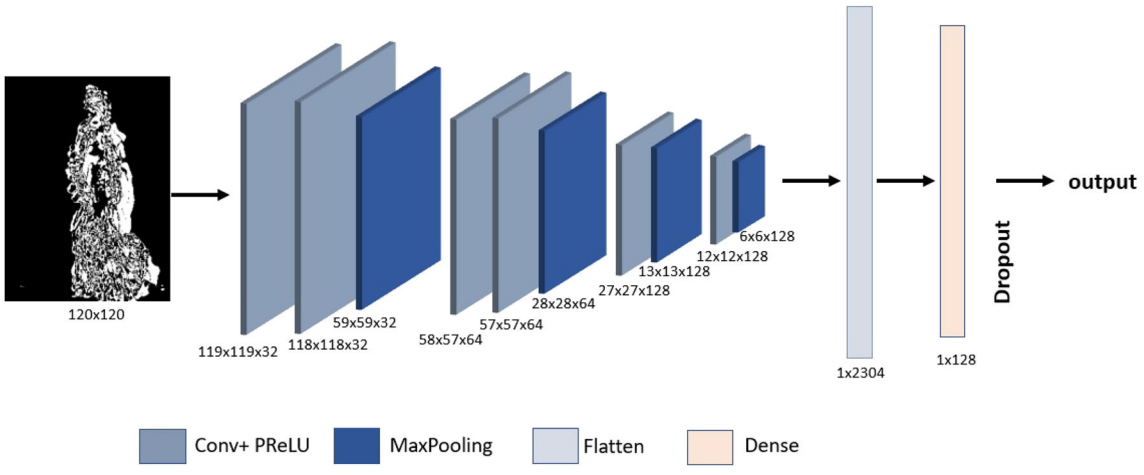

Fig. 6 CNN model trained on TDMap images

3.7 Illumination change detection and video enhancement

Low lighting, uneven illumination or any change of illumination in the observed scene are among the sources of degradation that strongly affect video quality and consequently the process of scene analysis and understanding and particularly object detection and visual tracking performance [65, 66]. It is therefore useful to detect the illumination changes and apply the appropriate pre-processing before performing high-level vision tasks such as moving object detection and tracking. In this work, we propose an illumination change detection technique before enhancing the video quality using a Retinex-based perceptual Contrast Enhancement method using Luminance Adaptation (RCELA) [31]. This method is adapted to our problem to make it appropriate for real-time processing.

Illumination change detection is an active research topic in computer vision [67, 68]. In this work, we use a simple and efficient method for illumination change detection based on the entropy associated with the gray-level histogram of the pixels. Indeed, any change in luminance significantly affects the grayscale histogram of an image. This variation is much more pronounced in the entropy associated with the distribution of pixel gray-levels.

The entropy is defined as follows:

$$E_t = \sum_k P_k \log P_k \tag{8}$$

where P_k is the probability distribution of gray-levels ($0 \leq k \leq K$) in the input frame F_t

According to the characteristic cited below, and to detect the illumination changes at any time during the video, we use the entropy of each frame. Thus, the current image of the video can be enhanced when the absolute difference $|E_t - E_{t-1}|$ between the entropy of the current image E_t and the entropy of the previous background model E_{t-1} is greater than a threshold T which is set to 0.7 after the experiment.

The proposed illumination changes detection is used to detect the occurrences ofillumination changes at any time in the video: for example, when the light is switchedoff in a period in the video. If the video has a bad illumination from the beginning, oran illumination degradation is detected, we can apply the enhancement before starting therecognition process.

4 Results and discussion

In this section, we evaluate the performance of the proposed method. For the proposed background modeling technique, the evaluation is made on SBI dataset and comparing with two background-subtraction-based approaches. Also, the built multiple human action dataset is presented with a description.

For evaluating the performance of the multiple action recognition approach, Weizmann and UCF-ARG human action recognition datasets utilized to train the proposed approach, when the proposed dataset as well as PET09 dataset are used for testing. The performance of the proposed action recognition methods are presented also. The recognition of detected actions is performed using two method : (1) Action classification named (CS-HOG-TDMap) using Cosine similarity measure between HOGs of TDMap in the training set and the current action to recognize. (2) Action classification named (CNN-TDMap) using the proposed CNN model trained on TDMap images.

Summarization of action is tested on the generated video considering the lack of videos containing many person acting many actions in the same scene. The proposed method was compared with state-of-the-art methods using the same datasets.

4.1 Experimental

In the pre-processing phase, which has a main contribution in our approach, we used the simplest motion detection method based on the background subtraction technique. Based on object detection results, a tracking of moving objects, is directed for further analysis of events. In other words, object tracking aims to trace a moving object (i.e., person) and recognize their actions. Human activity recognition methods can barely recognize actions if more than one person is present in a scene. Using our proposed method we aim to overcome such obstacle. Herein, moving objects are tracked using a bounding box enclosing each tracked object. To each bounding box is associated a label to designates every human present in a scene. Next, we extract the moving person and record all movements with an image resolution of 320x240 pixels. The absolute differences between each extracted box are then computed to generate the motion mask. Binary frames are used to compute the motion zone in each box by computing the absolute temporal difference between frames. The number of frames (N) used in the training phase is set to 40. Then, HOG is computed for each action. In the testing phase, this histogram is computed and then compared with all action histograms using the local soft cosine measure in [30] the described in the proposed method section.

To summarize the action, the change of appearance between the frames is computed using a similarity measure to extract the class within each video by selecting the shot. From the histogram of similarities, the local maxima are calculated to detect the shots in order to recognize the action within the shot. Then, the video is summarized by

recording one frame from each shot. The proposed method is implemented using MATLAB R2018a on a computer with the following configuration: An Intel Core i5 processor running at 3.4 GHz and 8 GB of RAM.

The proposed method is an action-based video summarization approach. After the recognition of multiple action of the people in the scene, the summarization of each person action within a timeline is performed. The used datasets are suitable for summarization of the human action after recognition of each one them. While the other datasets like YouTube, UCF50 or Hollywood are not. So, the used dataset are the only ones in literature that are suitable for summarization of human actions in a private or public scene monitored by a

surveillance camera.

Datasets including HMDB51, UCF101, YouTube, and Hollywood cannot be used for recognizing actions with the proposed algorithm owing to the complexity of the videos that are collected from movies (eg.YouTube), videos captured by moving cameras, and video produced by a jitter camera such as the variation of the point of view, the variation of illuminations. Many approaches use the entire video to classify the action in it without analyzing the content of the videos. For that, we did not use this kind of datasets. Also, our method is to recognize and summarize multiple human actions for surveillance videos.

Using the presented representation of data that divide each video into short clips of 1 second, considering some videos that contains more than 10 second. The numbers of videos used are about 1500 clips used in the training and testing parts. The summarization of the number for each action is illustrated in the Table 1.

4.2 Datasets

The datasets used in the experiments, the Weizmann, KTH, UCF-ARG, UT-Interaction, IXMAS and MHAD (our dataset) are briefly reviewed in this section. The other types of video that are collected from movies (eg.YouTube), like HMDB51, UCF101, YouTube , and Hollywood cannot be used for recognizing actions with the proposed algorithm owing to the complexity of the videos [55]. Also, the videos are captured by moving cameras, and video produced by a jitter camera such as the variation of the point of view.

Table 1 Number of used videos for each action from each datasets

Action category	KTH	Weizmann	UCF-ARG	Total
Boxing	20	-	281	301
Carrying	20	-	221	241
Jogging	20	–	127	147
Running	20	20	88	128
Walking	20	30	243	293
Hand waving	20	18	223	261
Hand clapping	20	–	188	208
Digging	–	–	230	230
Throwing	–	–	260	260

The Weizmann's Actions as Space-Time Shapes dataset was recorded in 2005, aiming to test new algorithms for human action recognition [32] . Weizmann's dataset use s a space-time-based algorithm where each sequence represents only one person acting. The background is known, which makes it easy to remove. The collected dataset contains the main human actions, including (walking, running, jumping, galloping sideways, bending, one-hand waving, and two-hands waving, jumping in place, jumping jack and skipping). The dataset contains nine actions, and each action has nine videos that represent the situations of human actions made by nine different actors.

Similarly, the KTH dataset contains a set of human actions including walking, running, boxing, hand waving, hand clapping, and jogging [33] . In this paper, the video contains four different scenarios which represent many states of objects and scenes, including outdoors and indoor videos, different scales of human body and cloth e s with different colors. This dataset contains 2391 images with a resolution of 160x120 pixels captured by a static camera.

The Multi-view Human Action dataset UCF-ARG is a set of videos recorded from different angles and classified into three categories: a ground camera, a rooftop camera at a height of 100 feet, and an aerial camera mounted into the payload platform. Each of these subsets contains 10 actions acted by 12 actors, representing most situations possible for each action, including 4 iterations by each actor in different directions.

Because of the similarity between the KTH and UCF-ARG videos and UIUC dataset [3] videos, w e use UIUC dataset just for testing. The dataset consists of 532 high-resolution video sequences of 14 human action classes, and every action is performed by eight persons. All the video sequences are recorded indoor scenes.

For human-human interaction the UT-Interaction [46] and IXMAS [47] datasets are used. The dataset contains 6 classes including shake-hands, point, hug, push, kick and punch. There is a total of 20 video sequences whose lengths are around 1 minute. For IXMAS datasets we choose Material class that represent sequences for human-human interaction.

For multiple human action recognition, we built our dataset named Multiple human action dataset (MHAD[1]). MHAD, as the naming reflects, to provide s a new dataset that contains many actions made by many actors in the same

[1] https://drive.google.com/open?id=1pfnnansy4VAejLRKNhCA8fn9IABarwwz

video. In one hand and related to video surveillance needs, each actor can act many actions during his presence in the scene. This can represent a rich data for many tasks in computer vision including video summarization methods based on human action, motion detection and tracking methods, people detection and recognition and people counting.

On the other hand, many persons can be found in the same video in action. Compared to the existing video surveillance dataset (that contains moving objects in the scene but with one action like walking), our dataset provides many persons acting different actions in the same video.

The proposed dataset can help computer vision researchers, especially those working on video summarization, motion detection and tracking, real-time human action recognition and many related tasks. By the following we present the characteristics of the proposed dataset in details.

Dataset characteristics: The proposed dataset includes a set of human actions representing usual human activities. MHAD composed of 10 actions, including: boxing, walking, running, hand waving, hand clapping, jogging, carrying, standing, backpack carrying, and two persons fighting.

Generated videos contain from 3 to 5 persons acting in the scene. The duration of each video is 2-3 minutes and the duration of each action is from 2 to 3 minutes. In addition, three of the videos are outdoor and one is an indoor. The background is generated for each video and annotations of each moving actor is provided.

In the current work, we only used datasets that are captured by a fixed camera because our approach consists of a modification of the background and the detection of moving human bodies before tracking each one of them. For the same purpose, we have built our dataset containing three videos. For each video we can find persons acting different actions. Also, consecutive actions are performed by each person during his presence in the scene.

The accuracy of summarization is related to the recognition accuracy. So, if actions are well recognized, the summarization is just a representation of these actions by one image for each person 's action.

The ground truth of the action in our dataset depends on the succession of actions in each video. Unlike the other summarization methods, wherein the scene changes every time, our dataset is for multiple action recognition and the summarization is rather based on detection. Figure 7 represents the succession of actions for each person in two videos from our datasets.

4.3 Action recognition and summarization results

In order to evaluate the proposed method for background modeling, SBI dataset is used. Figure 8 represents the generated background using the proposed approach. The obtained results are convincing and the using our method the background is built without artificial ghost for all videos. For Foliage and People&Foliage sequences the background the proposed method success to estimate the background with good results even the sequences a full of moving object during all time of the videos.

In order to consolidate the visualized results, we used different metrics, including Gray-level Error (AGE), Total number of Error Pixels (EPs), Percentage of Error Pixels (pEPs), Total number of Clustered Error Pixels (CEPs), Peak-Signal-to-Noise-Ratio (PSNR), MultiScale Structural Similarity Index (MS-SSIM), Color image Quality Measure (CQM).

These metrics are presented in Table 2 that illustrate the obtained results comparing with two background modeling methods IMBS-MT [54] and [43] respectively. As shown, the proposed method succeed to modelized the background with good results compared to the other method in the most of dataset videos including *HighwayI*, *Hall&Monitor, sallen, and Foliage.*

To evaluate the proposed method, obtained results are compared with state-of-the-art methods. Most action recognition methods perform the recognition on the original data that contains one person in action. The presence of more than one person in a scene can reduce the performance of the recognition process. In addition, most methods use the silhouette of the moving object to recognize the action. Hence, the accuracy of the recognition can be influenced by a bad binarizaion or segmentation. To overcome all these problems, our method recognizes multiple actions of multiple actors. The mask of each moving human body in action is computed using the background. Then, each one of the extracted silhouettes is used to compute the histogram of oriented gradient (HOG), which is compared with all histograms formed in the training phase.

The recognition of human actions is attained by computing the distance between HOGs of each extracted target based on training-phase results. In the training phase, the histogram of oriented gradient of the temporal difference map frame of the video is computed. Therefore, for each action we have a set of HOGs where each HOG represents one situation of one action is obtained. In the testing phase we compute the HOG of the detected sequence. A comparison of HOG with all HOGs of the training phase using the distance listed above is then performed. The minimum value of all distance values represents the action. To evaluate the effectiveness of the proposed technique, KTH, Weizmann, and UCF-ARG datasets are tested against each other. In addition, we perform a test between the video of within each dataset.

The extracted silhouette for each human body during a video sequence might contain more than one action. The

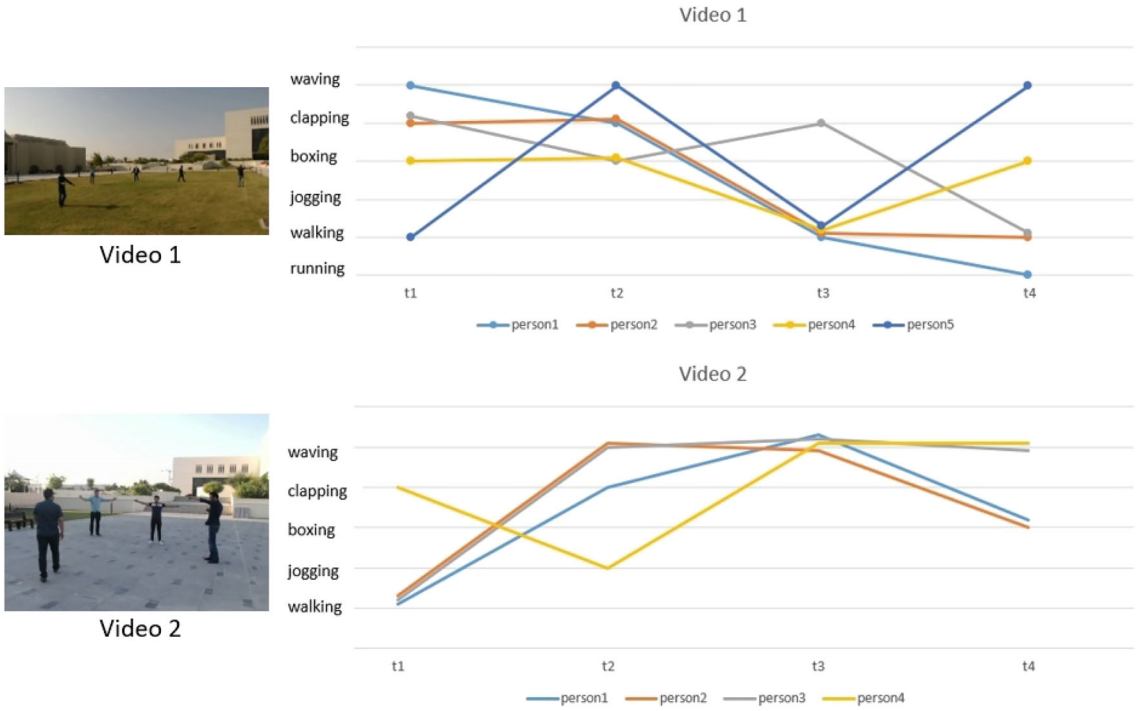

Fig. 7 Succession of actions for each person in each video

proposed technique for detecting shots, which are similar images that contain the same action, is presented in Fig. 9. The histogram of similarity between each two consecutive images is computed. Then, a filter is applied to extract the shots using the difference between each two consecutive values. The histogram presented in Fig. 9b shows the transaction between actions. From that, we can observe the transaction caused by the change from one action to another.

Tables 3, 4, 5, 6 and 7 illustrate the similarity distances between HOGs of the actions of the three datasets: within the KTH dataset, within the Weizmann dataset, within the UCF-ARG dataset, and between the KHT and UCF-ARG datasets . The distance is computed using cosine similarity measure of equation (4). The results tabulated in Table 2, which represent the similarity distances within the actions of the KTH dataset, reveal that the distance between similar actions such as walking, and jogging is smaller than that between very different actions such as hand-waving and running. It can be clearly seen in Table 3 that the distance between running and walking is smaller than that between

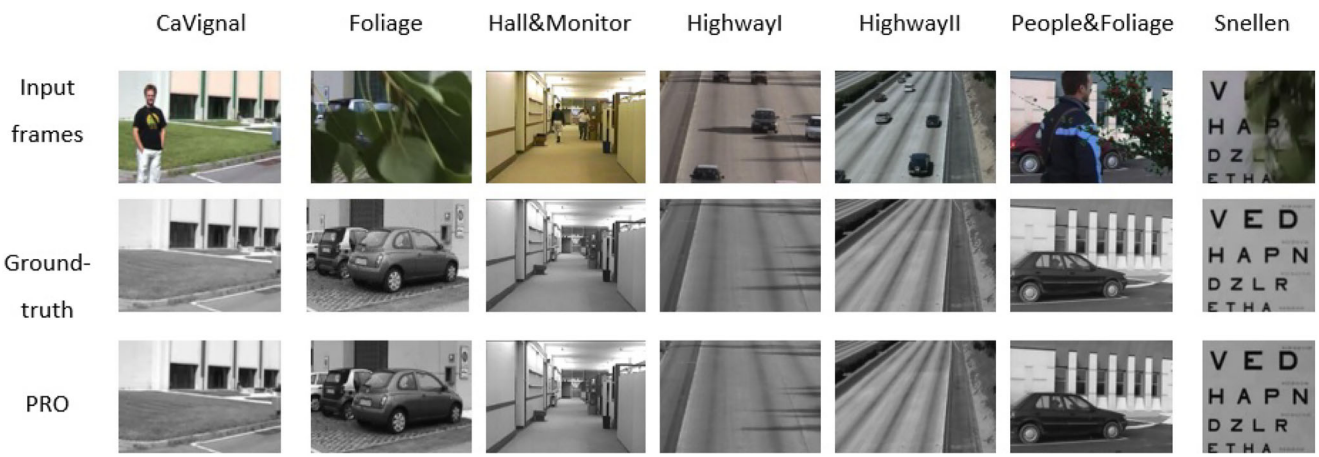

Fig. 8 Background results on the SBI dataset using theproposed approach

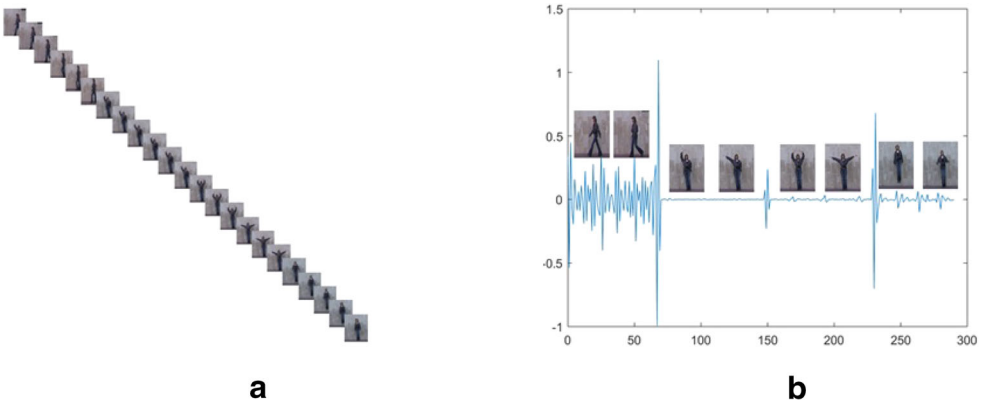

Fig. 9 Actions sequence and the histogram of similarity between each two consecutive frames. **a** Sequence of actions. **b** Histogram of similarity values

other actions, as well as the jumping action. In addition, waving with two hands in Hand wave 1 is close to Hand wave 2, which is an action of two hands waving. The evaluation of actions of the Ground dataset and Rooftop dataset of the UCF-ARG dataset, which represent categories of videos captured by a ground camera and a rooftop camera, respectively, are shown in Tables 4 and 5. From tabulated results, it can be clearly seen that similar or close

Table 2 Performance results the compared methods on SBI dataset

Method	AGE	pEPs%	pCEPS%	MSSSIM	PSNR	CQM
			CaVignal			
IMBS-MT	**0.7692**	**0.0147**	**0.0000**	**0.9982**	**45.9202**	**57.1044**
[43]	3.8855	0.0041	**0.0000**	0.9933	34.8725	54.5813
Ours	2.4732	0.0521	0.0005	0.9826	36.7687	56.8224
			Foliage			
IMBS-MT	7.5809	9.8507	3.1319	0.9090	22.7278	34.0028
[43]	8.5594	0.4313	**0.0000**	0.9892	27.7099	39.6381
Ours	**2.4569**	**0.1586**	0.0034	**0.9964**	**36.3866**	**43.2006**
			HallAndMonitor			
IMBS-MT	1.5350	**0.0923**	**0.0000**	**0.9954**	38.6214	48.5224
[43]	2.3878	0.1567	0.0102	0.9934	37.9820	61.3861
Ours	**1.3945**	0.1554	0.0112	0.9926	**40.1342**	**62.6213**
			HighwayI			
IMBS-MT	1.4913	0.0612	0.0026	0.9939	14.7728	58.8328
[43]	3.0301	0.1855	0.0085	0.9880	35.0837	59.7762
Ours	**1.3173**	**0.0042**	**0.0000**	**0.9956**	**42.7145**	**63.4632**
			HighwayII			
IMBS-MT	1.8684	0.0260	0.0000	0.9960	**40.1098**	48.80094
[43]	2.3279	0.1113	0.0000	**0.9967**	38.9867	**49.7341**
Ours	**2.1022**	**0.0120**	**0.0000**	0.9939	36.2130	45.1972
			PeopleAndFoliage			
IMBS-MT	8.3982	7.3568	3.2305	0.8514	20.0658	32.5231
[43]	5.7884	0.1974	0.0034	0.9885	35.7556	47.2501
Ours	**1.5720**	**0.0067**	**0.0000**	**0.9944**	**41.1009**	**47.4375**
			Snellen			
IMBS-MT	14.4480	25.3279	19.7290	0.8668	19.7436	40.115
[43]	3.7620	**0.0163**	**0.0000**	0.9951	37.1563	49.3740
Ours	**1.7584**	0.0214	0.5691	**0.9981**	**38.1033**	**50.3010**

Table 3 HOG distances between actions in the KTH dataset

Walking	Running	Hand waving	Hand clapping	Boxing	Jogging	
Walking	0	11.6660	13.3649	12.6702	12.7134	8.1761
Running		0	13.5564	12.5536	12.8201	10.2712
Hand waving			0	11.1547	11.8629	12.6148
Hand clapping				0	9.5362	11.7075
Boxing					0	11.7063
Jogging						0

Table 4 HOG distances between actions in the Weizmann datasets

Walk	Run	Hand wave1	Hand wave2	Jump	
Walk	0	7.6265	14.2200	14.6942	10.1359
Run		0	14.2022	13.5661	8.9414
Hand wave1			0	8.6277	14.0759
Hand wave2				0	14.4815
Jump					0

Table 5 HOG distances between actions in the Rooftop dataset (UCF_ARG)

Walking	Jogging	Carrying	Running	Clapping	Waving	Boxing	
walking	**0**	5.7182	6.0824	6.5012	9.5363	9.4992	8.8397
jogging		**0**	5.5433	5.7861	9.7584	9.581	9.0902
carrying			**0**	6.2462	9.1589	9.0042	9.295
running				**0**	9.3355	9.2795	8.9813
clapping					**0**	9.1084	9.7879
waving						**0**	9.8457
boxing							**0**

Table 6 HOG distances between actions in the Ground dataset (UCF_ARG)

Walking	Jogging	Carrying	Running	Clapping	Waving	Boxing	
Walking	**0**	8.1535	7.3251	8.2226	11.7833	12.7892	11.3271
Jogging		**0**	8.1356	7.1703	10.8356	11.5909	10.0972
Carrying			**0**	8.1052	11.2039	12.2782	10.9788
Running				**0**	11.733	12.5412	11.1658
Clapping					**0**	9.6713	9.7428
Waving						**0**	8.9645
Boxing							**0**

Table 7 HOG distances between actions of the KTH dataset and the Ground of the UCF_ARG dataset

Walking	Jogging	Running	Clapping	Waving	Boxing	
Walking	9.1847	9.5707	9.4803	13.188	13.434	12.79
Running	8.763	8.837	8.3031	12.9458	12.7355	13.3534
Jogging	13.9284	13.7113	14.1774	12.8062	9.6436	11.7898
Hand waving	12.6371	12.7246	13.0104	11.6685	10.7648	13.3327
Hand clapping	12.5395	11.973	12.0112	10.3149	10.7344	12.0164
Boxing	12.1038	12.3846	12.8088	11.6816	10.4053	11.616

actions such as walking, jogging, carrying and running have a minimum distance between them because of the similarity of the appearance in terms of the direction and moving component of the human body. The same applies for hands waving and clapping; the distance is close for many situations. Similarly, for the recognition of KTH actions in Ground(UCF-ARG) actions represented in Table 6, we can observe that the proposed method failed to recognize the jogging actions. In addition, the boxing action of the KTH dataset is recognized as a waving action in the Ground dataset; this is due to the similarity in the appearance of boxing and waving in many situations.

The errors rate in Table 7 represents 2% of the tested data, and between the closest actions like jigging and running. Also, in some cases, actions can be similar, and so the summarization using images can be useful to spot the differences.

For the sake of comparison of the proposed method's results with some of the state-of-the-art methods, the accuracy of each approach is presented in Table 8. The two proposed method for action recognition are named respectively by: (1) the classification using cosine similarity measure between HOGs of the training set and the current HOG of the current action (CS-HOG-TDMap).(2) the recognition using CNN model (CNN-TDMap). The accuracy for the state-of-the-art methods are the same values reported in the papers. While the accuracy rate obtained using the proposed method is the ratio between the number of recognized action s and the total number of actions. The KTH, Weizmann and UCF-ARG datasets have been used by many methods in the literature for the last three years. The obtained results using the proposed approach for all the datasets are convincing and robust. However, the proposed method succeeds in recognizing over 98% of actions in the KTH and Weizmann, and UCF-ARG datasets using CS-HOG-TDMap and 99% using CNN-TDMap, because of the simplicity of the data, which contains a simple background; in addition, the actions are clear and in normal situations. For instance, in the UT-interaction dataset, the proposed method reach 87% and 98% recognition rates. For the IXMAS dataset, a large dataset with many situations of each action, the proposed method had a successful recognition rate of 99%.

Compared to state-of-the-art-methods related to multiple human action recognition, that use the same category of

Table 8 Recognition rate comparison using single action

Method	Accuracy (%)
KTH dataset	
Kaminski et al. 2017 [24]	92%
Dasari et al. 2018 [25]	87%
El-Henawy et al. 2018 [27]	95%
Sreeraj et al. 2015 [36]	95%
Shao et al. 2014 [37]	95%
Yong et al. 2015 [38]	96%
Cheng et al. 2016 [34]	97%
Liu et al. 2017 [39]	94%
Sharif et al. 2017 [41]	**99%**
CS-HOG-TDMap	**98%**
CNN-TDMap	**99.82%**
Weizmann dataset	
Jiang et al. 2015 [40]	95%
Zhang et al. 2016 [35]	**98%**
Kaminski et al. 2017 [24]	81%
Sharif et al 2017 [41]	95%
CS-HOG-TDMap	**98%**
CNN-TDMap	**99.85%**
UCF-ARG dataset	
Qazi et al. 2017 [42]	90%
CS-HOG-TDMap	**97%**
CNN-TDMap	**99.73%**
UT-interaction	
Saho et al. 2019 [50]	76%
Jalal et al. 2018 [51]	83%
CS-HOG-TDMap	**87%**
CNN-TDMap	**98.9%**
IXMAS	
Liu et al. 2018 [52]	**99.9%**
CS-HOG-TDMap	**98%**
CNN-TDMap	**99.6%**

Table 9 Accuracy comparison of multiple human action recognition with state-of-the-art-methods

Methods	Accuracy (%)
Jin et al. 2017 [48]	96% (KTH)
Akula et al. 2018 [49]	87.44% (Infrared videos)
CS-HOG-TDMap	**97.5%**
CNN-TDMap	**98,9%**

datasets represented, Table 9 represents the accuracy rate for each method, it can be observed that the proposed method results are improved and more effective. Obtained results are related to the use of the new representation of the data.

Detection and recognition of multiple human action using the proposed method can be implemented in real-time via an extended version. The proposed approach is tested on our dataset by extracting the sequence of each person in the scene and apply the proposed algorithm. The obtained results are showed in the Fig. 10 illustrating the detected person and their actions on MHAD and PET datasets. The visualized results represent one example from PET09 dataset and three videos from our dataset.

The proposed approach is validated in three major steps including human detection, subsequence extraction, recognition and summarization of each person actions. UIUCI dataset is use also for testing the trained actions. Figure 11a shows some obtained results. These videos are not included in the training phase because they contain the same category of action in the used datasets. For UIUC dataset video the recognition rate reach 98%.

For the human-human action we use two datasets such as IXMAS and UY-interaction Fig. 11b and c represents some obtained results of detection and recognition of human interactions. For example, in IXMAS dataset we use some video where two persons fighting in the training phase. The obtained results shown in Fig. 11b shows the recognition in four videos captured from different field of view (FOV). Also, the results in Fig. 11c illustrates some results from UT-interaction dataset representing some recognized actions.

The accuracy of the proposed algorithm is related to the detection tracking and the segmentation of the human body in the scene. Additionally, the detected person may be in an invariant position that requires a large number of videos in the training phase to represent all of the actions in several positions.

In this paper, we defined a video summary as a concatenation of action labels and keyframes of the video where the actors perform that action. The summarization using human actions presented in this paper consists of splitting the actions of each person present in the scene. Based on the results of the recognition and the extraction of each action, the summarization using the entire image is generated. Figure 12 presents the results of the proposed method's application on our dataset, which includes two persons and many actions for each of them. The body silhouette of a person is extracted before detecting the shots that represent one action. For each shot, our algorithm selects frames that contain the corresponding body silhouette. The same process takes place for each person in the scene. If a person

Fig. 10 Action recognition results tested on three of MHAD videos and on video from PET09

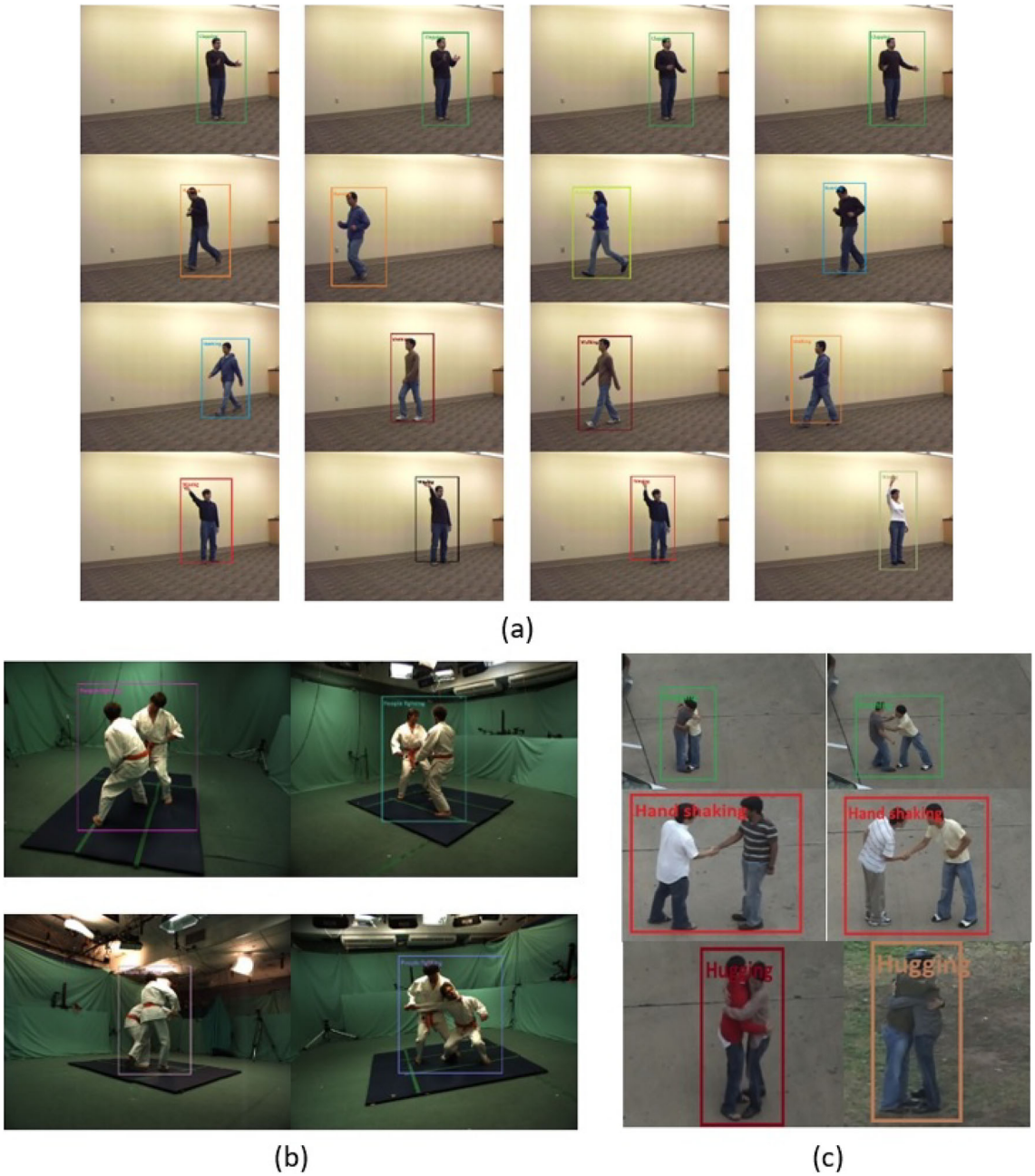

Fig. 11 Recognition results of UIUC, IXMAS and UT-Interaction datasets. **a** UIUC dataset. **b** IXMAS dataset. **c** UT-Interaction dataset

enters the surveilled zone and performs one action, walking into the scene for example, the summarization of this is one image. The proposed method can help video surveillance system operators show just the most important times in the video, noting that most areas are empty at most times.

As presented in Fig. 13, in the testing part a sequence of each detected person is generated each time5 and the action in each is recognized using the proposed model. The succession of analysis (motion detection, motion tracking and the proposed architecture) provides a multi human action recognition. After that, each action can be represented at the original video. Also, in order to summarize the detected and recognized actions during the entire time of a video, Fig. 13 represents a summarization

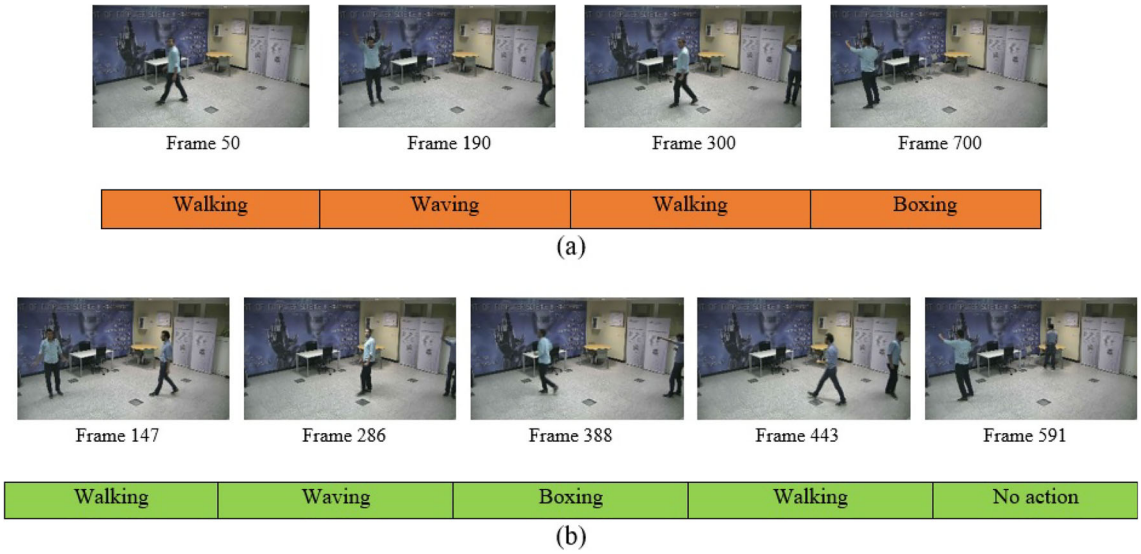

Fig. 12 Action-based summarization of each person in the scene. **a** Summary of person 1 using extracted shots of the silhouette detection. **b** Person 2 summary during his presence in the scene

using graphs of the recognized actions of each person during his presence in the scene.

The action recognition results can be influenced by the illumination changes. The detection of any illumination changes in the scene can be useful to enhance the video captured before recognizing the action. The proposed illumination changes detection methods allow to apply the video quality enhancement just in the case when there is any change. The following Fig. 14 illustrate the results of enhancement for two videos (LightSwitch and Lobby) from Star dataset. After the illumination changes detected the enhancement of next frames is takes place.

The enhancement is an additional part in the work that allows us to enhance the video if there is any illumination change during the video. But the novelty and the difference of the proposed method from the existing methods is the consideration of multiple human action(s) recognition. In addition to the combination of recognition and summarization of actions is not used for action recognition methods in literature.

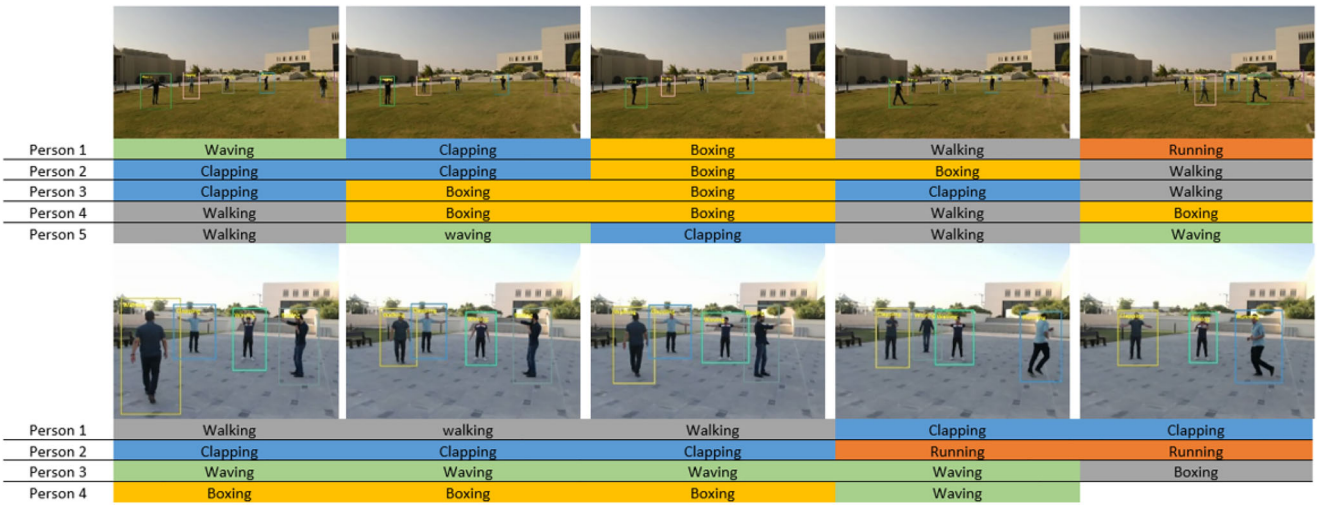

Fig. 13 Summarization of actions made by each person during his presence in the scene. First row : actions recognition and summarization for video 1. Second row: actions recognition and summarization for video 2

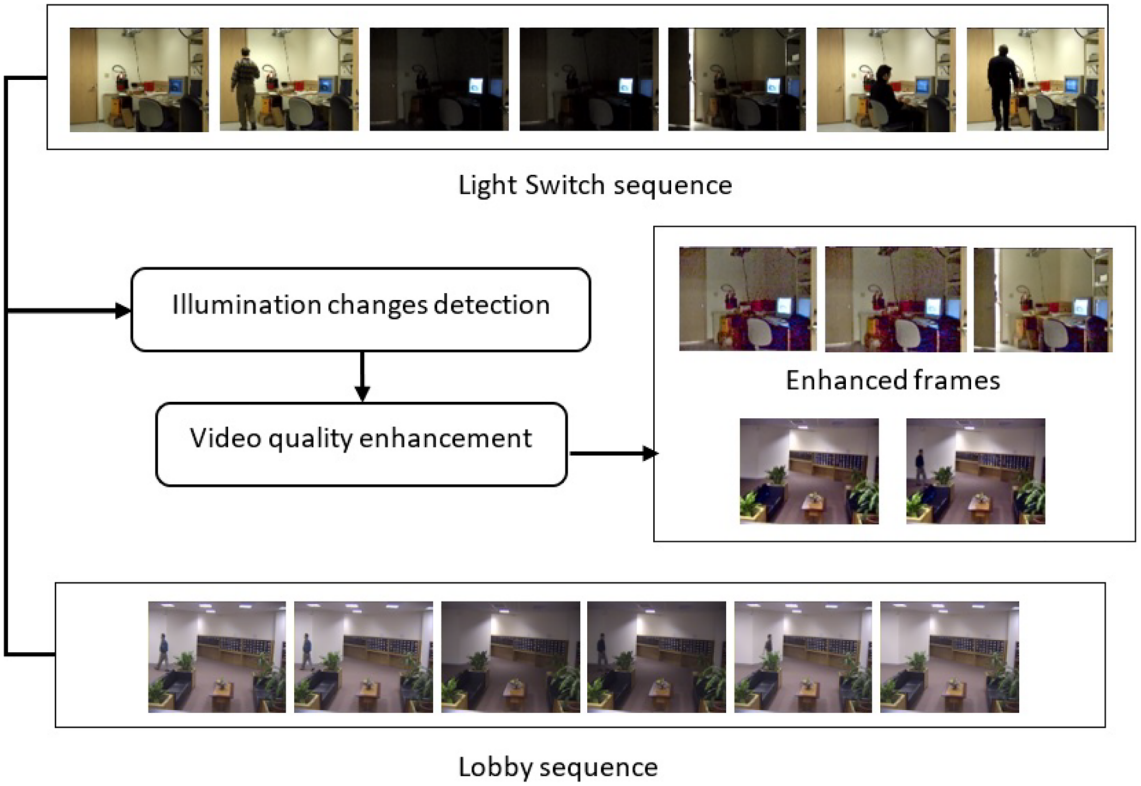

Fig. 14 Sub-sequence quality enhancement after detection of illumination changes

5 Conclusions

In this work, a novel approach for multiple human detection, recognition and summarization was developed, where actions of a person presenting in a scene are summarized. Herein, for a specific scene, motion/actions of each person are detected, tracked, and a sequence of each human body silhouette is generated. Upon recognizing and summarizing each action within a shot (i.e. each shot represents a subsequence that represent one action), shots detection operation was developed in order to determine the set of actions in the generated sequence. Shots that represent the homogeneous part of the sequence were selected using the cosine similarity measure of consecutive frames. The recognition of each action was based on two steps. First, using a training dataset of HOGs of TDMaps that was generated in each sub-sequence. The generated HOGs, which represent different situation of each action, were computed and then used in the testing phase. The recognition is made using the distance between the HOG of the current action and the HOGs of the training by selecting the minimum one. in addition, the computed TDMaps images are used in a CNN model to classify also the action. Summarization was made by representing all shots with an image for each detected person during his presence in the scene. Using the proposed algorithm, multiple detection and recognition of human action could be achieved. This algorithm can be also used in real-time due to its simplicity and the number of features used in this method.

Acknowledgments This publication was made by NPRP grant # NPRP8-140-2-065 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Author Contributions Omar Elharrouss carried out the experimentation, generated the results and was responsible for writing the paper. Noor Almaadeed designed the approach and supervised the writing of the paper. Somaya Al- Maadeed analyzed the results and helped in the writing of the paper. Ahmd Bouridane and Azeddine Beghdadi reviewed the approach and the results to further improve the quality of the paper.

Funding Information Open Access funding provided by the Qatar National Library. This publication was made by NPRP grant # NPRP8-140-2-065 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Availability of data and materials Our recorded dataset can be found at the following link: https://drive.google.com/open?id=1pfnnansy4VAejLRKNhCA8fn9IABarwwz

Compliance with Ethical Standards

Competing interests The authors declare that they have no competing interests.

Abbreviations •, HOG: histogram of the oriented gradient; •, TDMap: Temporal Difference Map.

References

1. Almeida J, Torres R. d. S., Leite NJ (2010) Rapid video summarization on compressed video. In: 2010 IEEE International Symposium on Multimedia (ISM). IEEE
2. Almeida J, Leite NJ, Torres R. d. S. (2012) Vison: Video summarization for online applications. Pattern Recognit Lett 33(4):397–409
3. Almeida J, Leite NJ, Torres R. d. S. (2013) Online video summarization on compressed domain. J Vis Commun Imag Represent 24(6):729–738
4. Xu Q et al (2014) Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence. Inform Sci 278:736–756
5. Mei S et al (2015) Video summarization via minimum sparse reconstruction. Pattern Recognit Lett 48(2):522–533
6. Martins GB, Papa JP, Almeida J (2016) Temporal-and spatial-driven video summarization using optimum-path forest. In: 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE
7. os Santos Belo L et al (2016) Summarizing video sequence using a graph-based hierarchical approach. Neurocomputing 173:1001–1016
8. Mehmood I et al (2016) Divide-and-conquer based summarization framework for extracting affective video content. Neurocomputing 174:393–403
9. ujatha C et al (2014) Multilevel Framework for Summarization of surveillance videos. In: 2014 Fifth International Conference on Signal and Image Processing (ICSIP). IEEE
10. Elharrouss O, Al-Maadeed N, Al-Maadeed S (2019) Video Summarization based on Motion Detection for Surveillance Systems. In 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, pp 366–371
11. Zhang S, Zhu Y, Roy-Chowdhury AK (2016) Context-aware surveillance video summarization. IEEE Trans Image Proc 25(11):5469–5478
12. Chen F, De Vleeschouwer C, Cavallaro A (2014) Resource allocation for personalized video summarization. IEEE Trans Multimed 16(2):455–469
13. Bagheri S, Zheng JY, Sinha S (2016) Temporal mapping of surveillancevideo for indexing and summarization. Comput Vis Image Understand 144:237–257
14. Bagheri S, Zheng JY, Sinha S (2016) Temporal mapping of surveillancevideo for indexing and summarization. Comput Vis Image Understand 144:237–257
15. Song X et al (2016) Event-based large scale surveillance video summarization. Neurocomputing 187:66–74
16. Xu X, Hospedales TM, Gong S (2017) Discovery of shared semanticspaces for multiscene video query and summarization. IEEE Trans Circ Syst Vid Technol 27(6):1353–1367
17. Xu X, Hospedales TM, Gong S (2017) Discovery of shared semanticspaces for multiscene video query and summarization. IEEE Trans Circ Syst Vid Technol 27(6):1353–1367
18. Kalaivani P, Roomi SMM (2017) Towards Comprehensive Understanding of Event Detection and Video Summarization Approaches. In: 2017 Second International Conference on Recent Trendsand Challenges in Computational Models (ICRTCCM). IEEE
19. Yun S et al (2014) Visual surveillance briefing system: Event-based video retrieval and summarization. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE
20. Lin W et al (2015) Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. Neurocomputing 155:84–98
21. Tejero-de-Pablos A et al (2016) Human action recognition-based video summarization for RGB-D personal sports video. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE
22. Tejero-de-Pablos A, Nakashima Y, Sato T, Yokoya N, Linna M, Rahtu E (2018) Summarization of user-generated sports video by using deep action recognition features. IEEE Trans Multimed 20(8):2000–2011
23. Wang L et al (2017) Video enhancement using temporal-spatial total variation retinex and luminance adaptation. In: 2017 International Conference on Progress in Informatics and Computing (PIC). IEEE
24. Kamiński Ł, Maćkowiak S, Domański M (2017) Human activity recognition using standard descriptors of MPEG CDVS. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE
25. Dasari R, Chen CW (2018) MPEG CDVS Feature Trajectories for Action Recognition in Videos . In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE
26. El-Masry M, Fakhr MW, Salem M. A.-M. (2017) Action recognition by discriminative EdgeBoxes. IET Comput Vis 12(4):443–452
27. El-Henawy I, Ahmed K, Mahmoud H (2018) Action recognition using fast HOG3D of integral videos and Smith–Waterman partial matching. IET Imag Process 12(6):896–908
28. Xu C, He J, Zhang X (2017) DFSA: A classification capability quantification method for human action recognition. In: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE
29. Zhang J et al (2018) Action recognition from arbitrary views using transferable dictionary learning. IEEE Trans. Image Proc 27(10):4709
30. Sidorov G et al (2014) Soft similarity and soft cosine measure: Similarity of features in vector space model. Comput Sist 18(3):491–504
31. Xu K, Jung C (2017) Retinex-based perceptual contrast enhancement in images using luminance adaptation. In: 2017 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). IEEE
32. Blank M et al (2005) Actions as space-time shapes. In: null. IEEE

33. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: 2004. ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition. IEEE
34. Cheng S et al (2016) Action recognition based on spatio-temporal log-Euclidean covariance matrix. Int J Sig Process Imag Process Pattern Recog 9(2):95–106
35. Zhang S, Zhang W, Li Y (2016) Human Action Recognition Based on Multifeature Fusion. In: Proceedings of 2016 Chinese Intelligent Systems Conference. Springer
36. Sreeraj M (2015) Multi-posture Human Detection Based on Hybrid HOG-BO Feature. In: 2015 Fifth International Conference on Advances in Computing and Communications (ICACC). IEEE
37. Shao L et al (2014) Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Trans Cybernet 44(6):817–827
38. Yang J, Ma Z, Xie M (2015) Action recognition based on multi-scale oriented neighborhood features. Int J Sig Process Imag Process Pattern Recog 8(1):241–254
39. Liu H et al (2017) Study of human action recognition based on improved spatio-temporal features. In: Human Motion Sensing and Recognition. Springer, 233–250
40. Jiang J et al (2015) Human action recognition via compressive-sensing-based dimensionality reduction. Optik-Int J Light Elect Opt 126(9-10):882–887
41. Sharif M et al (2017) A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. EURASIP J Imag Vid Proc 2017(1):89
42. Qazi HA et al (2017) Human action recognition using SIFT and HOG method. In: 2017 International Conference on Information and Communication Technologies (ICICT). IEEE
43. ELHARROUSS O, ABBAD A, MOUJAHID D et al (2017) Moving object detection zone using a block-based background model. IET Comput Vis 12(1):86–94
44. Lefloch D, Cheikh FA, Hardeberg JY, Gouton P, Picot-Clemente R (2008) Real-time people counting system using a single video camera. In: Real-Time Image Processing 2008 vol. 6811. International Society for Optics and Photonics, pp 681109
45. Tran D, Sorokin A (2008) Human activity recognition with metric learning. In: European conference on computer vision. Springer
46. Ryoo MS, Aggarwal JK (2009) Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV. Citeseer
47. Weinland D, Boyer E, Ronfard R (2007) Action recognition from arbitrary views using 3d exemplars. In: ICCV 2007-11th IEEE International Conference on Computer Vision. IEEE
48. JIN C-B, Shengzhe LI, Hakil etKIM (2017) Real-Time Action Detection in Video Surveillance using Sub-Action Descriptor with Multi-CNN. arXiv:1710.03383
49. AKULA A, SHAH AxK, et GHOSH R (2018) Deep learning approach for human action recognition in infrared images. Cogn Syst Res 50:146–154
50. Sahoo SP, Ari S (2019) On an algorithm for human action recognition. Expert Syst Appl 115:524–534
51. Jalal A, Mahmood M, Sidduqi MA (2018) December. Robust spatio-temporal features for human interaction recognition via artificial neural network. In: IEEE Conference on FIT
52. Liu Y, Lu Z, Li J, Yang T (2018) Hierarchically Learned View-Invariant Representations for Cross-View Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology
53. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
54. Bloisi DD, Pennisi A, Iocchi L (2017) Parallel multi-modal background modeling. Pattern Recogn Lett 96:45–54
55. Wang H, Oneata D, Verbeek J, Schmid C (2016) A robust and efficient video representation for action recognition. Int J Comput Vis 119(3):219–238
56. Masoumi M, Amiri S (2013) A blind scene-based watermarking for video copyright protection. AEU-Int J Electron Commun 67(6):528–535
57. Li G, Li C (2020) Learning skeleton information for human action analysis using Kinect. Signal Processing: Image Communication, pp 115814
58. Wang L, Huynh DQ, Koniusz P (2019) A comparative review of recent Kinect-based action recognition algorithms. IEEE Trans Image Process 29:15–28
59. Megrhi S, Jmal M, Souidène W, Beghdadi A (2016) Spatio-temporal action localization and detection for human action recognition in big dataset. J Vis Commun Image Represent 41:375–390
60. Megrhi S, Jmal M, Beghdadi A, Souidene W (2015) Spatio-temporal action localization for human action recognition in large dataset. Proceedings of SPIE 9407, Video Surveillance and Transportation Imaging Applications 2015, pp 94070O. https://doi.org/10.1117/12.2082880.4
61. Megrhi A, Beghdadi W (2014) Souidene, Trajectory feature fusion for human action recognition. IEEE–EUVIP2014, Paris
62. BOUTTEFROY PLM, BOUZERDOUM A, PHUNG SL et al (2008) Abnormal behavior detection using a multi-modal stochastic learning approach. In: 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing. IEEE, pp 121–126
63. Khan SD, Ullah H, Ullah M, Conci N, Alaya-Chekh F, Beghdadi A (2019) Person Head Detection Based Deep Model for People Counting in Sports Videos, AVSS The 16-th IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS), Taipei, pp 18–21
64. Khan SD, Ullah H, Ullah M, Alaya-Cheikh F, Beghdadi A (2019) Dimension invariant model for human head detection. EUVIP2019, Rome, pp 28–31
65. Moujahid D, Eharrouss O, et Tairi H (2018) Visual object tracking via the local soft cosine similarity. Pattern Recogn Lett 110:79–85
66. Elharrouss O et al (2016) Moving object detection using a background modeling based on entropy theory and quad-tree decomposition. J Electron Imaging 25.6:061615
67. Lou J-G, Yang HT, Hu W, Tan T (2002) An Illumination Invariant Change Detection Algorithm, ACCV2002: The 5th Asian Conference on Computer Vision, Melbourne, pp 23–25
68. Do QB, Beghdadi A, Luong M (2013) Color Mismatch Compensation Method Based On a Physical Model. IEEE Trans Circ Syst Video Technolo 3(2):244–257

Dr. Omar Elharrouss received his master's degree in 2013 from the Faculty of Sciences, Dhar El Mehraz, Fez, Morocco. In 2017, He received his PhD in the LIIAN Laboratory at USMBA-Fez University. His research interests include pattern recognition, image processing, and computer vision.

Dr. Noor Almaadeed received her Ph.D. from Brunel University, London, UK, in 2014. Her areas of research include speech signal detection, speaker identification, audio/visual speaker recognition, etc. She received her Bachelor's in Computer Science from Qatar university in 2000, and Master of Science in Computer and Information Sciences from City University, UK, in 2005. She is an Assistant Professor in the Department of Computer Science & Engineering in Qatar University. She was awarded the Qatar Education Excellence Day Platinum Award -New PhD Holders Category 2014-2015.

Somaya Al-Maadeed received the Ph.D. degree in computer science from the University of Nottingham, U.K., in 2004. She is currently the Head of the Computer Science Department, Qatar University, where she is also the Coordinator of Computer Vision Research Group. She enjoys excellent collaboration with National, International institutions, and industry. She was a Visiting Academic with Northumbria University, U.K. She has authored extensively in computer vision and pattern recognition. She is a Principal Investigator of several funded research projects. She attended workshops related to higher education strategy, assessment methods, and interactive teaching. She published and delivered workshops on teaching programming for undergraduate students. She supervised students through research projects related to community and industry. In 2015, she was elected as the IEEE Chair for Qatar section. She and her team received the Best Performance at ICDAR2011 and ICDAR2015 Signature Verification. She organized several workshops and competitions related to biometrics and computer vision. She was selected as a Participant at the Current and Future Executive Leaders Program, Qatar Leadership Centre (2012-2013) established in 2008 by an Emiri Decree.

Ahmed Bouridane received the "Ingenieur d'Etat" degree in electronics from "Ecole Nationale Polytechnique" of Algiers (ENPA), Algeria, in 1982, the M.Phil. degree in electrical engineering (VLSI design for signal processing) from the University of Newcastle-Upon-Tyne, U.K., in 1988, and the Ph.D. degree in electrical engineering (computer vision) from the University of Nottingham, U.K., in 1992. From 1992 to 1994, he worked as a Research Developer in telesurveillance and access control applications. In 1994, he joined Queen's University Belfast, Belfast, U.K., initially as Lecturer in computer architecture and image processing and later on he was promoted to Reader in Computer Science. He is now a full Professor in Image Engineering and Security at Northumbria University at Newcastle (UK), and his research interests are in imaging for forensics and security, biometrics, homeland security, image/video watermarking and cryptography. He has authored and co-authored more than 300 publications and one research book. Prof. Bouridane is a Senior Member of IEEE.

Azeddine Beghdadi is Full Professor at the University Sorbonne Paris Nord. He published more than 280 international refereed scientific papers. His research interests include image quality enhancement/assessment, image/video compression, multimedia security, bio-inspired models for image analysis and processing, and physics-based image analysis and processing. Dr. Beghdadi is the founder of the European Workshop on Visual Information Processing (EUVIP). He is associate editor of "Signal processing: Image Communication", Journal, Elsevier, European journal on image and video processing, Springer Verlag, Journal of Electronic Imaging, SPIE Digital Library, and Mathematical Problems in Engineering, Journal, Hindawi.

Affiliations

Omar Elharrouss[1] · Noor Almaadeed[1] · Somaya Al-Maadeed[1] · Ahmed Bouridane[2] · Azeddine Beghdadi[3]

Noor Almaadeed
n.alali@qu.edu.qa

Somaya Al-Maadeed
s_alali@qu.edu.qa

Ahmed Bouridane
ahmed.bouridane@northumbria.ac.uk

Azeddine Beghdadi
azeddine.beghdadi@univ-paris13.fr

1 Department of Computer Science and Engineering, Qatar University, Doha, Qatar

2 Department of Computer and Information Sciences, Northumbria University at Newcastle, Newcastle, UK

3 Galilée Institute Sorbonne Paris Nord Universite France, Paris, France