Multimed Tools Appl
https://doi.org/10.1007/s11042-018-6273-1

Stacked sparse autoencoder and history of binary motion image for human activity recognition

Mariem Gnouma[1] · Ammar Ladjailia[2,3] · Ridha Ejbali[1] · Mourad Zaied[1]

Received: 3 October 2017 / Revised: 25 April 2018 / Accepted: 15 June 2018

Abstract The recognition of human actions in a video sequence still remains a challenging task in the computer vision community. Several techniques have been proposed until today such as silhouette detection, local space-time features and optical flow techniques. In this paper, a supervised way followed by an unsupervised learning using the principle of the auto-encoder is proposed to address the problem. We introduce a new foreground detection architecture based on information extracted from the Gaussian mixture model (GMM) incorporating with the uniform motion of Magnitude of Optical Flow (MOF). Thus, we use a fast dynamic frame skipping technique to avoid frames that contain irrelevant motion, making it possible to decrease the computational complexity of silhouette extraction. Furthermore a new technique of representations to construct an informative concept for human action recognition based on the superposition of human silhouettes is presented. We called this approach history of binary motion image (HBMI).Our method has been evaluated by a classification on the Ixmas, Weizmann, and KTH datasets, the Sparce Stacked Auto-encoder

✉ Mariem Gnouma
mariem21gnouma@gmail.com; mariem.gnouma.tn@ieee.org

Ammar ladjailia
l_ammardz@yahoo.fr

Ridha ejbali
ridha_ejbali@ieee.org

Mourad zaied
mourad.zaied@ieee.org

[1] Research Team on Intelligent Machines, National School of Engineers of Gabes, University of Gabes, Gabes, Tunisia

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

[3] Algeria Department of Computer Science, University of Annaba, Annaba, Algeria

Published online: 05 July 2018

(SSAE), an instance of a deep learning strategy, is presented for efficient human activities detection and the Softmax (SMC) for the classification. The objective of this classifier in deep learning is the learning of function hierarchies with higher-level functions at lower-level functions of the hierarchy to provide an agile, robust and simple method. The results prove the efficiency of our proposed approach with respect to the irregularity in the performance of an action shape distortion, change of point of view as well as significant changes of scale.

Keywords Human activity recognition · Silhouette extraction · History of binary motion image · Deep learning

1 Introduction

In recent years, Human Activity Recognition (HAR) has attracted much attention, especially in the fields of video analysis technology. Recognizing human actions is a crucial element in computer vision applications such as video control, human computer communication, video browsing and analysis of abnormal actions [20].

To monitor activities in public places, several applications have been proposed [41] and can be divided into five categories: approaches based on frequency, spatio-temporal, local descriptors, methods based on appearance and methods based on form.

Almost, a human activity belongs to a set of actions that can be performed sequentially by different body regions. These compositions can appear temporally, and they can affect interactions with the environment, other people, or particular objects. For instance, people can speak while walking, running or waving their hand. We note that different compositional arrangements of actions can yield different semantics at a higher level.

Frequency-based methods like the Discrete Fourier Transform (DFT) [51] or its derivatives have been used to extract features to capture the geometric structure but are disposed of for partial occlusions. Spatio-temporal methods include Spatio-Temporal Volumes [42], Spatial Trajectories [43] and Spatio-Temporal Points [52] which are used for the invariance of the speed of action but require a theoretical background. For example, in spatial-temporal approaches, a Motion History Image (MHI) and a Motion Energy Image (MEI) are produced from a sequence of images that present the locations of movements. Local descriptors are usually used when the scale, the translation and the rotation occur. However, these techniques are very expensive in terms of calculation, especially those using optical flow processes [33, 35]. Appearance-based methods [2, 32] are based on the use of sample images called patterns or copies of objects for recognition which differ according to several conditions such as a change in the direction of vision, changing in the size or the shape of the object and the change of lighting. However, the methods based on the form [31] are based on the extraction of the silhouette of a person. In [40] proposed an approach uses spatio-temporal body parts movement (STBPM) features extracted from foreground silhouette of the human objects. The STBPM feature estimates the movements of different body parts for any given time segment to classify actions. Generally, for most character extraction approaches [3], the K-Nearest Neighbors and vector support machines are used for classification. Our work suggests the implementation of a model that has a number of features like occlusions, local being and the robustness that helps us obtain a coherent local motion descriptor for recognizing human activities using RGB-D data. Moreover, we use a fast dynamic frame skipping technique to avoid frames that contain irrelevant motion, making

it possible to decrease the computational complexity of silhouette extraction. This article aims at presenting the following contributions:

- The use of a fast dynamic frame skipping technique to considerably reduce the computational complexity of silhouette extraction
- The improvement of silhouette detection rate by applying a new foreground segmentation technique by combining the advantages of the Gaussian Mixture Model with the magnitude of optical flow motion.
- A new system architecture for HAR in video stream
- A new concept of representation to construct an informative and discriminative semantic overview for HAR systems by the use of the History of Binary Image representation.
- The use of the Stacked Sparse Auto encoder (SSAE) that allows learning high level information from a large number of unlabeled image patches. Our method of classifying human actions is therefore basically different from a number of current methods which are based on low level image characteristics (edge, texture, color, scoreboard, etc.).
- By training the SSAE classifier with unmarked instances, the SSAE model uses a hierarchical architecture to transform the original pixel signal intensities of the input image patches into the corresponding high-level structural information. During the ranking phase, each image patch to be evaluated is introduced into the hierarchical architecture and represented by a high-level structured representation of patches representing the human action.
- A typical experimentation through three sets of action data to verify the performance of the proposed method

We organized the remainder of this paper as follows. Section 2 introduces the related work in the literature. Section 3 explains techniques of frame skipping and of background modeling. Section 4 introduces our technique of HAR using motion boundary information. Section 5 reports the experimental results and discussions. Finally, we draw conclusions in Section 6.

2 Related work

Recognizing a human activity from a video sequence is one of the most important applications of computer vision. Various approaches to recognition [16, 17, 29, 48] have been proposed in the literature. A key consideration is that feature representations were used from the video sequence volumes. In particular, they calculate optical-space-time gradients [10] and partial reproductions that detect the points of interest of space-time volumes [57] and several other typical intensity-based have a major limit when videos are captured at lower quality. Recently, most researchers have treated human silhouettes and used them as characteristics for the discovery of human activity [8, 30]. Many researchers use of the silhouette of the person is the fact that it contains very useful and detailed information on the shape of the body. Indeed, a sequence of silhouettes generates space-time forms that contain information about the dynamics of the temporal motion of the global body and instantaneous spatial information about the position of the local body.

The approaches for activity recognition based on silhouette are divided into 2D [28, 37, 39, 50] and 3D [11, 36, 46, 55] methods. The 3D methods are based on modeling to perform on the human silhouette; Nonetheless, it is hard to treat a high number of joints of the body with them. Furthermore, 2D methods are recognizable by using a human silhouette appearance. Because of the use of a fixed interior background, it is unnecessary to follow

all the complex process when creating skeleton in 3D methods. As well, we use ,in our approach, a method that allows to obtain excellent results in using a steady background since it is very difficult to treat a changeable background and it offers excellent work with a steady background. Furthermore, human action is presented on several successive images. For these reasons, we use this method.

Chang et al. [10] extended their 3D method by using three motion history images of the provided data by combining them with the three dimensional depth data of human body movements to obtain the motion history image in 3D. Chan-drashekhar and Venkatesh proposed another work [9] for the recognition of the activity using a 2D space-time representation by combining all the video images into a single image called Action Energy Image. Xiaofei et al. [37] proposed a representation of the spatial-temporal silhouette to characterize the properties of motion such as daily activities. They used the multi-class SVM as a classifier where each action consists of multiple views and scenarios of motion descriptors.

Several publications have appeared in recent years documenting human action recognition. One of these examples is presented in [20] which the authors used 2D models for feature extraction. Nevertheless, their matching process was long.

They used two distinct images, namely, MEI and MHI for the phase [7]. Unlike Bobick and David, we used a single binary motion image [6]. Binary Motion Image has been used by Tushar et al.

Recently, deep convolutional neural networks (CNN) have been proposed for large-scale image processing. Dobhal et al. [14] who worked on the idea of 2D representation by combining the image sequences into a single image to perform the activity recognition. For the classification part, they used the convolution neural networks CNNs, which made it possible to make a set of sub-convolution subsampling operations to learn the neurons. The limit of this work is that they did not work on all actions, only actions that define a vertical person movement such as walk, run, slide, skip and jump.

Another solution is described in [54] based on CNN, which propose a technique for action recognition named stratified pooling, which is based on deep convolution neural networks (SP-CNN). The process is mainly composed of fine-tuning a pre-trained CNN on the target dataset, frame-level features extraction; the principal component analysis (PCA) method for feature dimensionality reduction; stratified pooling frame-level features to get video-level feature; and SVM for multiclass classification.

However, there are several problems to directly apply existing deep CNN models to action recognition. Firstly, the structures of video-based CNN and image-based CNN are different, thus the video-based CNN weights must be trained on video dataset from scratch. Secondly, a CNN usually contains tens of millions of parameters, which rely on sufficient amount of training videos to prevent over-fitting. Thirdly, the network needs several weeks to train depending on the architecture. That's why we proposed a solution based on Staked Sparce auto encoder.

In this article, a method of recognizing the action independently of the view to overcome the problem of high dimensioning related to the representations of multi-camera actions based on the silhouette is proposed. This paper focuses on a 3D representation of data based on a 3D motion model.

Our method directly uses the History of Binary Motion Image as input to our SSAE. The main advantage of our approach is the ability to be easily extended to incorporate high-level information as well as speed invariance, partial occlusion and distortion that are introduced by the extraction of the characteristics, which makes it simple and effective to implement.

3 Proposed approach

Figure 1 provides an overview of our proposed method for Human Action Recognition.

To define a correct human action label for a given video clip input,, the proposed method makes use of three sequential modules:

1. Leap frames technique with the following of the points of interest in each selected frame, the overall complexity of the calculation decreases for efficient detection system.
2. On-line Foreground segmentation using GMM and the uniform motion of Magnitude of Optical Flow (MOF).
3. The construction of an informative and discriminative descriptor for HAR systems by the use of the HBMI representation.
4. Human action classification to obtain a proper human action label using the Sparce Stacked Auto-encoder (SSAE) combined with the Softmax classifier (SMC) .

3.1 Dynamic frame skipping

Current research on frame skipping like in [12, 19, 27, 45] are in focus by making use of the total optical flow of information(magnitude + orientation + axes x and y). We propose a new frame skipping technique to reduce the complexity of optical flow computation. We adopt a dynamic frame skipping, a technique that has been broadly used for effective adjustment of the bit rate in video transcoding.

However, in order to obtain an effective HAR, we tried to develop a new technique for image leaping.

In this paper, as an alternative, we make use of a simple and fast technique for image skipping based firstly on the difference between adjacent frames : we divide the frame into blocks.

For optical flow computation, several algorithms exist in the literature. The two most popular were proposed by Lucas and Kanade [38] and Horn and Schunck [26]. In our previous work [21] we demonstrated that the algorithm of Lucas and Kanade is more suitable to be applied for dense optical flow computation.

Assuming that the brightness of each moving pixel in the frame is constant,

$$\mathrm{MV} = \begin{bmatrix} d_x \\ d_y \end{bmatrix} \tag{1}$$

With d_x and d_y are its displacements along the axis.

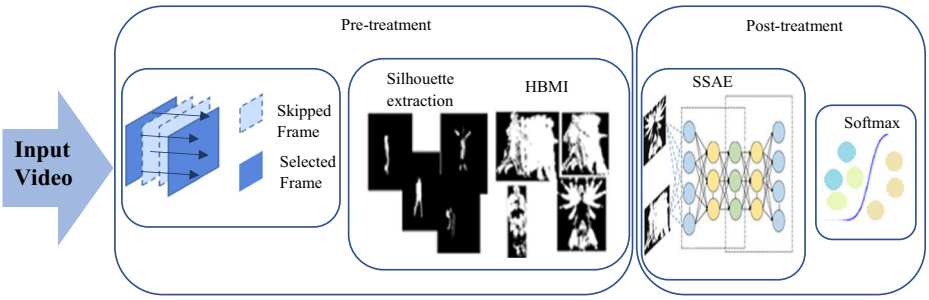


Fig. 1 Overview of our algorithm

To a given motion vector MV, we compute the magnitude Mg:

$$Mg = \sqrt{dx^2 + dy^2} \tag{2}$$

As previously mentioned, only optical flow information (magnitude) is used to construct the feature vector for each frame. To do this, we calculate the optical flow field between the two frames; then, we compute it (using the method of Lukas Kanade).

The choice of the use of the magnitude only as the main information for the computation of the optical flow since it indicates the total speed of movement of the pixel, then we do not need other information which can overburden the process of computation.

The motion field is a rich dynamic descriptor of flow that is related to the flow velocity, density, and motion pressure of each pixel. However, all velocity vectors are color-coded according to the vector displacement angle as shown in our old search [21].

Therefore, the speed can not be calculated without providing an additional information. To view the road of optical flow, we use the image of velocity vectors "Map motion" in Fig. 2.

This mapping is like building an image that combines angle and standard velocity fields at a point in space (RGB). Each block of the image obtained is a set of points of interest. We are interested only in the points of interest that are marked in white and that define an important movement.

With $p(mg(c, t)))$ is the probability of observing the current cell c at the time t.
We begin by the initialization of the the δ =0 at the t=1, then we calculate the sum index of white intensity of each cell after the processing of filtration of the second obtained image:

$$\delta = \sum_{1}^{n} \frac{1}{w * h} \delta + S \tag{3}$$

With S representing the sum of the white intensity of each cell in the image and the w and h denote the width and the height of the frame. Thus, we apply a filtering process to reject irrelevant frames that do not contain motion. The rejected frame contains a number of pixels of white luminance below a well-defined threshold.

The adequate threshold is chosen empirically for measuring the uniformity of selected frames which should satisfy the fact that light intensity should surpass this threshold.

The maximum number of frames to be skipped is limited to six in order to avoid motion fog. Therefore, we only extract frames that contain important information.

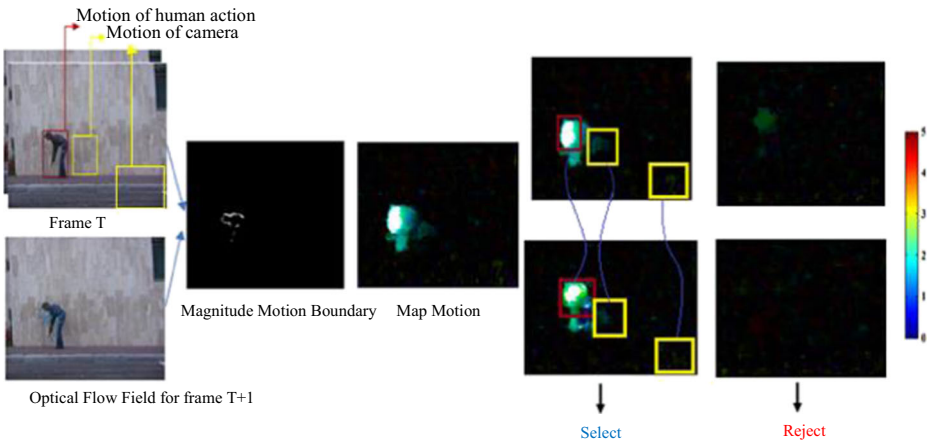


Fig. 2 Scheme for motion boundary information-based frame selected

By skipping frames with unimportant changes in motion, through the use of the dynamic frame skipping technique to track interest points in the frame, the overall complexity of computation decreases.

3.2 Silhouette extraction

In contrast to the aforementioned approaches for the classification of a motion in a public place, we used the KTH, IXmas and Weizmann datasets, in contrast to [14], which used the BMI to classify only five actions from the Weizmann database. We aim at developing methods that are generally useful to various videos. An overview of our technique for silhouette extraction is presented in Fig. 3.

In the first step, the frames to be used are selected by dynamic frame skipping as mentioned above. We extract only the frames that contain motion for efficiency reasons.

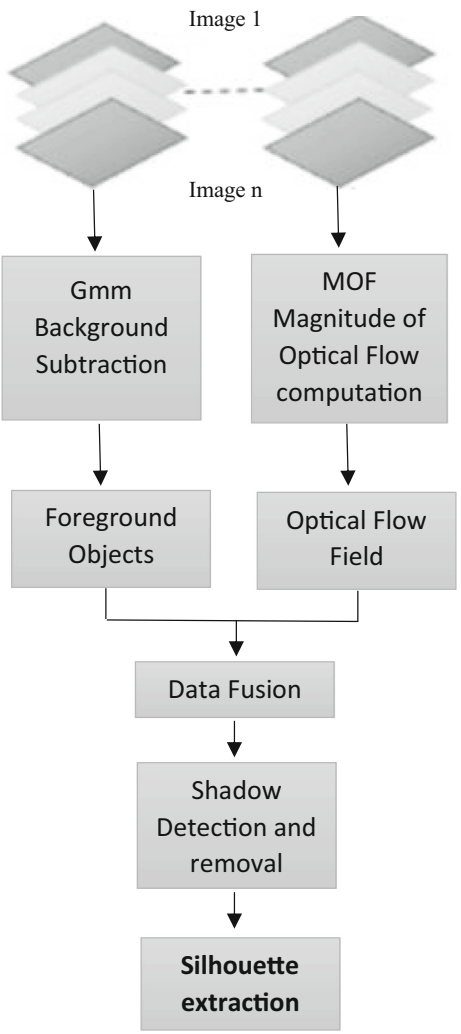


Fig. 3 The process of silhouette detection

Due to unexpected changes in short- and long-term dynamic scenes such as camera noise, light reflectance, darkness and shadow, the detection of the silhouette is a big problem to deal with. However, it is necessary to pay attention to the silhouette detection step to have good, robust and reliable recognition system. That is why we suggest a solution based on incorporating a uniform motion model into GMM [4] background subtraction used to each image slot selected from the first step. By considering these two suggestions, high accuracy of foreground segmentation for silhouette extraction is obtained.

We label the pixels of image M by comparing each pixel $X_1 ... X_t$ with a mixture of Gaussians. Then, the probability is determined using (4):

$$P(X_t) = \sum_{i=1}^{k} \omega_{i,t} \eta(X_t; \mu i, t, \sum_{i,t} \tag{4}$$

With the history of the intensity of each pixel ($X_0 ::: Y_0$) is given by (5) as:

$$X_1, X_2, ..., X_t = M(X_0, Y_0, ..., i) : 1 \leq i \leq t \tag{5}$$

Or $\omega_{i,t}$is an estimate of the weight (the portion of the data is recorded with this Gaussian) in the mixture at time t. $\sum_{i,t}$ and $\mu_{i,t}$are respectively the average and the covariance value of the Gaussian. μ is a Gaussian probability density.

We have

$\sum_{i=1}^{k} \mu_{i,t} = 1$

The average of the GM is $\sum_{i=1}^{k} \mu_{i,t} \mu_{i,t}$

With

K: the number of distributions

t: the time

Figure 2 shows the optical flow across two frames selected from the first processus of skipping frames at times t and t + 1. For each point P located at the image coordinate, the dense optical flow field provides a motion vector which is expressed as the velocities MV = (1).

From these components (x,y) of the velocity field, the optical flow of each point P can be defined by its magnitude as previously described.

$$\text{Optical Flow}(\mathrm{P}_{x,x,t}) = \text{Magnitude}(\mathrm{P}_{x,x,t}) \tag{6}$$

Where the Magnitude ($\mathrm{P}_{x,x,t}$) = (2).

After computing the optical flow on the selected frame, the detection of uniform motion is performed.

The goal of this integration is to improve the detection rate of GMM background subtraction without deteriorating the precision. Actually, pixels that belong to the background and that undergo changes are correctly classified as background entities by GMM. However, these pixels and using an optical flow are likely to be classified as foreground entities. Therefore, we start by the results provided by GMM, then, by using the measure defined for

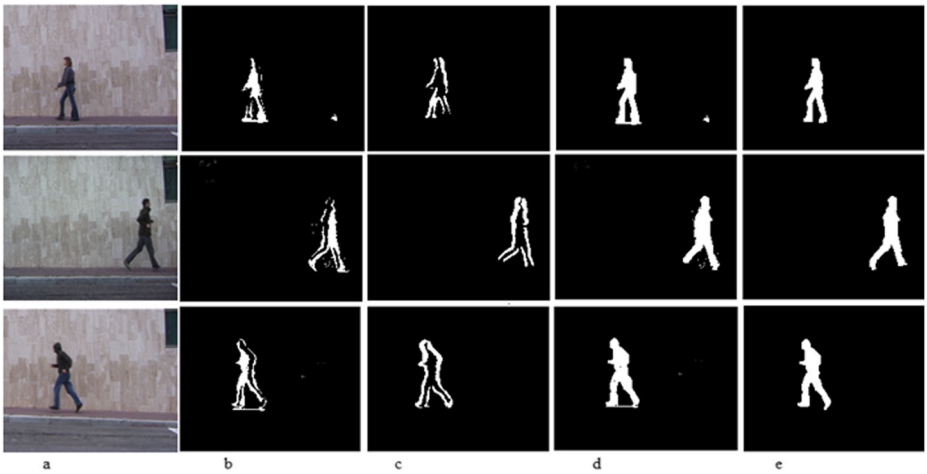


Fig. 4 Processing result: **a** original frames, **b** GMM, **c** Magnitude of Optical Flow computation(MOF), **d** Our approach without shadow removal and **e** Our approach after shadow removal

magnitude motion, the brand of each pixel is updated.This integration is an efficient way to correct the results and to avoid aberration caused by optical flow as well.

At this point of the process (time=t) there are two binary images extracted using described algorithms. The first image (Fig. 4b), extracts foreground pixels using GMM in an effective way, but it could contain shadows of moving objects that must be removed. The second image (Fig. 4c) includes only some parts of the moving objects.

The proposed idea in this paper is that any foreground region that corresponds to an object and does not exist in the image of MOF must be removed following steps that describe the proposed algorithm:

- A median filter, for the noise removal purpose, is applied to the binary image (Fig. 4d).
- A morphological close filtering is performed on the resulting image using a circular structuring element of 3-pixel diameter to fill the gaps and smooth outer edges (Fig. 4e).

The resulting image contains moving objects without their shadows. By employing the above-mentioned method, any sudden luminance change, like turning on a flash light in the scene, will not cause spurious foreground regions.

From these binary images obtained, the binary motion image (BMI) is then calculated, and is then fed to the auto-encoder model for training and testing.

The various frames are extracted from a video sequence and then the History of Binary Motion Image (HBMI) is calculated, which then serves the SSAE [1, 25] model for training and the Softmax SMC for the classification.

The main objective of image segmentation is the separation of the foreground and the background thus eliminating the noise using a filter. This pre-processing determines an object using the image collected from a camera. In addition, the retrieval functionality reduces the cost of recognition, since we do not use the full image.

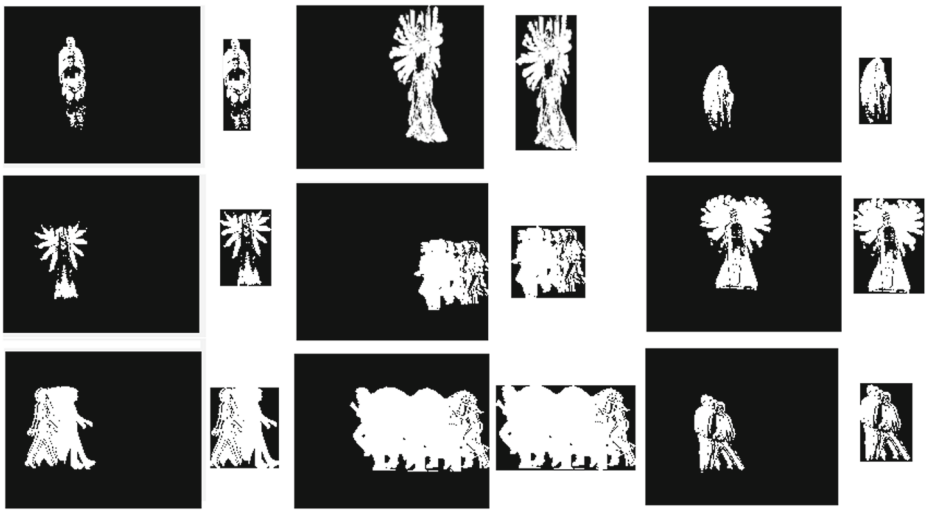


Fig. 5 An example of normalization operations

3.3 Feature extraction and normalization

For features, we used the HBMI which is more robust than other descriptor because of the use of depth information. The modeling of human actions makes it possible to modify the actions by the extraction of the different functions.

These input functions are subsequently converted into action sequences. Our algorithm begins with the development of a method that combines all the action sequences into a single image.

HBMI combines the binary image sequence using the (7)

$$HBMI(X, Y) = \bigcup_{t=1}^{n} f(t) M_{xy}(t) \tag{7}$$

Where HBMI(x,y) : the history of binary image.

$M_{xy}(t)$: the binary image sequence containing the ROI.

f(t) :is the weight function which gives higher performance to more recent frames.

n: the total number of frames.(video length) Points of interest (ROI) are obtained by delimiting the image and discarding the black background. Then, normalization operations are performed to obtain the final image input to the SSAE.

In action datasets, videos focus on human beings, but it is very common that humans do not all dominate the framework, which can be a problem for stock classification. lastly, a new proposition for frame normalization is presented in Algorithm 3.3 to extract only the region of interest in the image and to discard the black background.

We propose to eliminate all the unused information in the input image as shown in the following Fig. 5.

Algorithm 1 Normalization images dataset algorithms

Data: **ImagesTrain, ImagesTest, threshold, Width, Height**
Result: **ImagesTrainNorm, ImagesTestNorm**

▷ **ImagesTrain**: list of training images.
▷ **ImagesTest**: list of testing images.
▷ **threshold**: thresholding based on the magnitude of the velocity flow.
▷ **Width & Height**: the selected image size.
▷ **ImagesTrainNorm**: normalization of training images.
▷ **ImagesTestNorm**: normalization of testing images.

$ImagesTrainNorm \leftarrow \{\}$
$ImagesTestNorm \leftarrow \{\}$
$nbTrain \leftarrow numberOf(ImagesTrain)$
$nbTest \leftarrow numberOf(ImagesTest)$

▷ **Normalization for training images.**

for $k = 1, k <= nbTrain$ **do**
  $[x1, y1, x2, y2] \leftarrow getBoxSilhouette(ImagesTrain\,(k)\,, threshold)$
  ▷ **getBoxSilhouette**: return to the movement area.
  $image \leftarrow extractImage\,(ImagesTrain\,(k)\,, x1, y1, x2, y2)$
  $image \leftarrow resize\,(image, width, height)$
  $ImagesTrainNorm\,(k) \leftarrow image$
end

▷ **Normalization for testing images.**

for $k = 1, k <= nbTest$ **do**
  $[x1, y1, x2, y2] \leftarrow getBoxSilhouette(ImagesTest\,(k)\,, threshold)$
  $image \leftarrow extractImage\,(ImagesTest\,(k)\,, x1, y1, x2, y2)$
  $image \leftarrow resize\,(image, width, height)$
  $ImagesTestNorm\,(k) \leftarrow image$
end

3.4 Training and classification

3.4.1 SSAE

Learning unsupervised features helps learn discriminating and efficient characteristics from a large amount of unlabeled data [13]. In the field of recognition of human activities [34], labeled actions are difficult to acquire and require a specific and elaborate experimental framework. Therefore, unsupervised feature learning can provide an effective solution to activity recognition. In the proposed framework, the SSAE combined with a new technique applied to the input images is adopted to learn the characteristics for human activities. Then the learned functions are introduced into a Softmax classifier. The detail of the framework is illustrated in the following sections.

The basic structure of an automatic encoder is composed of an input layer, Hidden layer and Output layer. However, the intrinsic problems of the Autoencoder such as simply copying the input layer to a hidden layer make it inefficient to extract meaningful functions even though its output may be a perfect recovery of input data. As a learning algorithm, we use in our method a SSAE [18, 22–24] which benefits from all the advantages of a deep network [15] of greater expressive power.

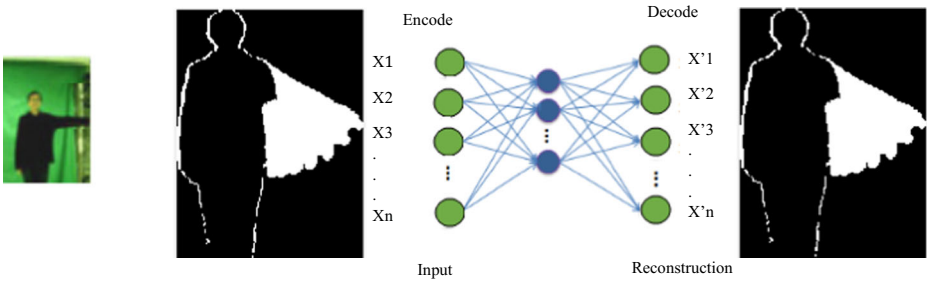


Fig. 6 Principle of Auto Encoder

We attempt to classify the video in a supervised way followed by an unsupervised learning using the principle of auto-encoder. Its role is to minimize the error of reconstruction of the data input as shown in Fig. 6.

3.4.2 SSAE+SMC

The approach presented in this paper (illustrated in Fig. 8) employs a full connection model for high-level feature learning. Autoencoder or Stacked Sparse Autoencoder (SSAE) is an encoder-decoder architecture where the "encoder" network represents pixel intensities modeled via lower dimensional attributes, while the "decoder" network reconstructs the original pixel intensities using the low dimensional features.

The stacked autoencoder is a neural network consisting of multiple layers of basic SAE (see Fig. 7) in which the outputs of each layer are wired to the inputs of each successive layer. We considered the two SAE layers, which consist of two hidden layers, and the Stacked Sparse Autoencoder (SSAE) to represent the two SAElayers.

SSAE is a full connection model which learns a single global weight matrix for feature representation. For our application, the size of the input image in our Softmax classifier differs from one image to the other since we have applied a normalization algorithm that extracts only the region of interest in the image, the overall complexity of computation process decreases.

In addition, each image patch may contain up to a single object that would be appropriate for construction of a full connection model. Therefore, we choose to use SSAE instead of other metrics in this paper.On the other hand, SSAE is trained, bottom up, in an unsupervised fashion, in order to extract hidden features. The efficient representations in turn pave the way for more accurate, supervised classification of the two types of patches. Moreover, this unsupervised feature learning is appropriate for images that contain several activities since we have a great deal of unlabeled image data to work with; image labels typically being expensive and laborious to come by. This higher level feature learning allows us to efficiently detect multiple actions from a large cohort of images.

Training an SSAE implies finding the optimal $\theta = (W, b_h, b_x)$ by minimizing the discrepancy between input and its reconstruction. After the optimal θ are made, the SSAE generates a function $f : R^{d_x} R^{d_h(2)}$ that transforms input pixel intensities of an image patch to a new feature representation $h^{(2)} = \mathrm{f(x)} \in R^{d_h(2)}$ of an action structures.

Each training patch x(k) of pixel intensities is represented by a high-level structured representation of patches $h^{(2)}(k)$ in the second hidden layer of the model. All the training patches can be written as $\left\{h^{(2)}(k), y^{(k)}\right\}_{k=1}^{N}$ where for each k $\in$ 1, 2, N, h(2) (k), y(k) is a pair of high-level features and its label. For the K class classification problem considered in this paper, the label of the k–th patch is y(k) $\in$ 1, 6 in the KTH dataset, y(k) $\in$1, 9 in

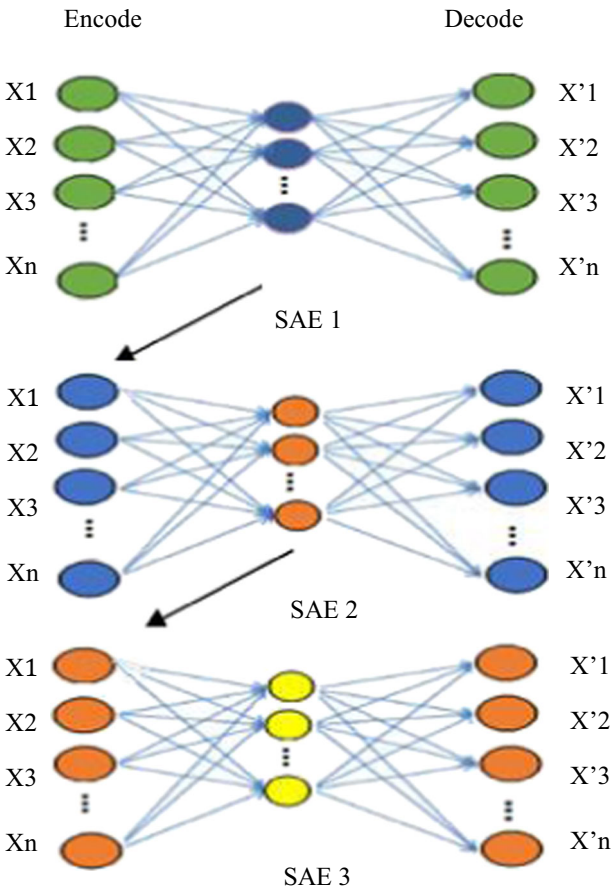


Fig. 7 The architecture of our Sparse auto encoder

the Weizmann database and y(k) ∈ 1, 11 in the IXMAS dataset. After the high-level feature learning procedure is complete, the learned high-level representation of an action structure, as well as its label $h^{(2)}$(k),y(k)$_{k=1}^{N}$, are fed to the output layer.

Our stacked auto encoder is a neural network [18, 53, 56] composed of three SAE base layers where, the hidden layer of each is wired to the inputs of the next layer (Fig. 7).

After extracting features for the training phase, a series of SAEs are used. In our network we have used a Softmax classifier (SMC) for the last layer. The Softmax classifier is a specialized activation function for classification networks. In the output, a linear transformation to the appropriate size of the last layer of the network was applied to the Softmax. When optimizing, Hidden layers learn how to transform entries for classes that are linearly separable. Eventually, for all hidden layers of the network we applied the fine tuning with the backpropagation algorithm to dramatically improve the performance of our SAE.

The SSAE +SMC architecture is illustrated in Fig. 8.

SMC is a supervised model which generalizes logistic regression as

$$f_{w^{(3)(ot)}} = \frac{1}{1 + ep(-w^{(3)^T} ot)} \tag{8}$$

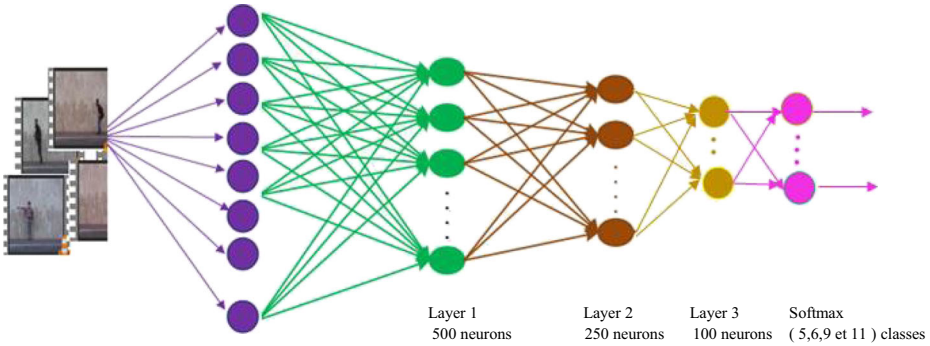


Fig. 8 The architecture of Stacked Sparse Auto-encoder and Softmax classifier

Where $f_{w^{(3)}}$ is a sigmoid function with parameters W(3). When the input ot of SMC is a high-level representation of an action structure learned by SSAE.

After training, the parameters θ of SSAE and W(3) of SMC are determined. Our SSAE+SMC is ready for Human Action detection.

4 Experimental results

We validate the effectiveness of our proposed model by testing the ability of our approach to discriminate simple and complex human actions on three benchmark datasets.

The proposed method is evaluated on the IXMAS dataset [49] as well as KTH [44] Dataset and Weizmann dataset [5].

The Inria Xmas Motion Sequence (IXMAS) is a set of data comparable to the current state of ”human action recognition” which contains 13 actions. Each action is presented three times by 10 actors (5 males/5 females).

The second dataset contains 10 types of action. In each class, a periodic action is carried out by 9 different persons and in a different manner and speed of execution. Each action sequence contains 40-90 frames outdoors with different clothes and indoors. The sequences are captured over homogeneous background with a static camera recording 25 f/s. Each video has a resolution of 160 $\times$ 120.

The Weizmann actions' dataset consists of 5,687 frames and 10 different categories of behavior classes: running, walking, jumping-in-place-on-two-legs (pjump), jumping-forward (jump), bending, jumping-jack (jack), galloping-sideways (side), skipping, waving-two-hands (wave2), waving-onehand (wave1). Video sequences in this dataset are captured with stationary camera and simple background. However, it provides a good experiment environment to investigate the recognition accuracy of the proposed method when the amount of behavior categories is large.

Using this datasets, both of qualitative and quantitative analysis of the results are presented with comparisons to the already cited methods.

All the experiments were carried out on an Intel Core i5 processor with 8 GB of RAM and NVIDIA Graphics Processor. The software implementation was performed using MATLAB 2017b.

The first, second and third layers consist of 500, 250 and 100 neurons, respectively. The last layer consists of n neurons, where n represents the number of types of actions (classes)

Fig. 9 Sample frames respectively from KTH, Weizmann and IXMAS datasets

to be recognized. The weights of the first three layers are initialized by pre-training. The last layer is a Softmax layer driven to generate a binary output with 1 for the predicted class and 0 for the rest. Figure 9 shows examples of different actions for the three datasets repectively in KTH, Weizmann and Ixmas: An example of HBMI results using several actions by eight different person is shown in Fig. 10.

Table 1 gives the action-confusion matrix provided by the kth dataset. A higher recognition rates have been reported for the actions Run, Walk, and Wave. The most confusion is between the three actions: jogging, boxing, and claping.

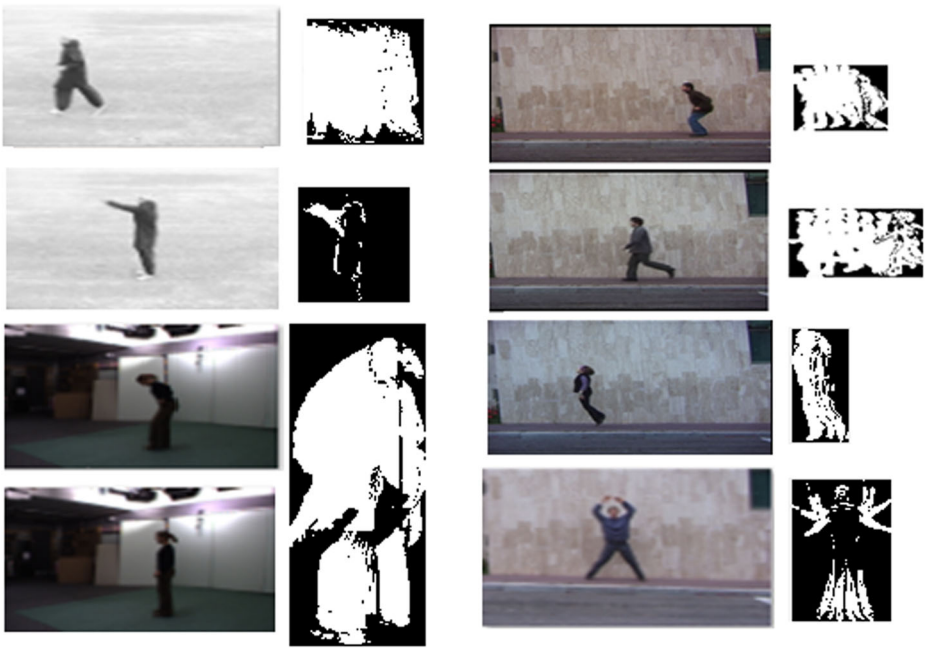


Fig. 10 Sample action performed by eight actors from different datasets

Table 1 Confusion matrix representation for the fusion-based action recognition (KTH dataset)

(%)	Walk	Joging	Run	Boxing	Wave	Clapping
Walk	100	0	0	0	0	0
Jogging	2	95	3	0	0	0
Run	0	0	100	0	0	0
Boxing	0	0	0	96	0	4
Wave	0	0	0	0	100	0
Clapping	0	0	0	4	0	96

Table 2 Testing results in portion scenario for KTH dataset

Scenario	3D SIFT	3D Sift + PDI	HBMI
SC1	0.96	0.96	0.985
SC2	0.8867	0.92	0.962
SC3	0.8542	0.9167	0.966
SC4	0.96	0.96	0.986

Table 3 Confusion matrix representation for the fusion-based action recognition (Weizmann dataset)

(%)	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave
Bend	100	0	0	0	0	0	0	0	0
Jack	0	100	0	0	0	0	0	0	0
Jump	0	0	96	0	0	0	4	0	0
Pjump	0	0	3	97	0	0	0	0	0
Run	0	0	0	0	100	0	0	0	0
Side	0	0	0	0	0	100	0	0	0
Skip	0	0	11	0	3	0	86	0	0
Walk	0	0	0	0	0	0	0	100	0
Wave	0	0	0	0	0	0	0	0	100

Table 4 Confusion matrix representation for the fusion-based action recognition (Weizmann dataset: 5 actions)

(%)	Walk	Jump	Run	Side	Skip
Walk	100	0	0	0	0
Jump	0	100	0	0	0
Runn	0	0	100	0	0
Side	0	0	0	100	0
Skip	0	0	0	0	100

Table 5 Recognition accuracy (%) on the Weizmann dataset

Method	Accuracy (%)
3Channel-CNN	76.00
Gray-CNN	86.46
QST-CNN-LSTM	96.34
HBMI- SSAE+SMC	97.66

In Table 2 the recognition experiments are performed by using combined features of HBMI on four scenarios (SC1, SC2, SC3 and SC4) according to KTH dataset, weizmann and xinria database. The experimental results for KTH dataset are shown in Table 2.

The experimental results in Table 2 show that the feature of HBMI has the better discriminative ability than PDI and SIFT feature [37]. SC1 and SC4 are more stable than the other two scenarios. We obtained almost the same recognition rate (98%). In SC2 and SC3, scenarios become more complex. Not only the human body exists scale variations with camera zooming in SC2, but also there are 45 degree view changes in jogging, running, walking. In SC3, human body shape changes with different wearing, and the phenomenon of non-homogenous background even emerges. These above situations make the silhouette area change obviously as well.

Table 3 gives the action-confusion matrix provided by the Weizmann dataset. A higher recognition rates have been reported for the actions bend, jack, run, side, Walk, and Wave. The most confusion is between the three actions: pjump, jump, and skip.

However, Table 4 presents the actions provided by 5 action (walk, run, jump, run, side,skip) in Weizmann dataset. This result is based on Leave-One-Person-Out cross validation.

Table 5 contains the performance comparison results for different recognition methods. Each test was repeated 20 times, and the maximum performance is shown. The results show

Table 6 Confusion matrix representation for the fusion-based action recognition (iXMAS dataset)

%	Check watch	Cross arms	Scratch head	Sit down	Get up	Turn around	Walk	Wave	Punch	Kick	Point	Pick up
Check watch	85	0	0	0	0	8	3	4	0	0	0	0
Cross arms	12	78	0	0	0	3	0	3	7	1	4	0
Scratch head	0	9	86	0	0	0	0	9	9	5	0	0
Sit down	0	0	0	100	0	0	0	0	0	0	0	0
Get up	0	0	0	0	98	2	0	0	0	0	0	0
Turn around	0	0	0	0	0	98	0	2	0	0	0	0
Walk	0	0	0	0	0	0	100	0	0	0	0	0
Wave	0	0	0	0	0	0	0	100	0	0	0	0
Punch	0	0	0	4	0	3	0	0	83	0	10	0
Kick	0	0	0	0	0	5	0	0	0	94	1	0
Pick up	0	0	0	7	0	0	0	0	1	0	0	92

Table 7 Results of the evaluation methods

Image	Methods	Execution time for 50 frames (milliseconds)	Memory used (bytes)	Recall	Precision	Similarity evaluation
Run	GMM [4]	24.33065	3.92e+008	0.7432	0.7022	0.781
	FCMOF [47]	30.31700	3.120e+008	0.7686	0.8059	0.924
	Ours	30.1752	3.15e+008	0.7997	0.8745	0.935
Jump	GMM [4]	20.38065	3.967e+008	0.8300	0.9321	0.848
	FCMOF [47]	22.76555	3.789e+008	0.8316	0.9427	0.952
	Ours	22.9524	3.742e+008	0.800	0.9635	0.964
Waving	GMM [4]	23.7857	5.436e+008	0.7475	0.8436	0.8743
	FCMOF [47]	25.4702	5.257e+008	0.8195	0.8323	0.9235
	Ours	25.3245	5.014e+008	0.7854	0.8654	0.9021

that the HBMI significantly outperforms the other for models. All of these methods have been tested under the same hardware configurations and software settings. One can see that HBMI achieved 97.66% and QST-CNN-LSTM achieved 96.34% recognition accuracy, whereas the accuracies of Gray-CNN were 86.46%, and 76.00%, respectively.

Table 6 presents the global evaluation metrics and the confusion matrix in Ixmas dataset. The major confusion occurs between cross arm and check match.

The computational complexity of our proposed method depends on: background modelling using both of GMM and the Magnitude of OF. Our proposed approach is compared to traditional algorithms and recent techniques. The execution time is calculated for 50 frames using tic-toc function and tabulated. Apart from time complexity, space complexity is also important. The usage of memory is found by memory function in MATLAB for 50 frames. The memory space used by this technique is compared to other methods and tabulated in Table 7.

Table 8 summarizes the overall performances of our action recognition systems in three datasets. we can see that our proposed method produces the best results for our implementation and for the test videos.

Table 8 Comparison between the proposed method and previous methods

Methods	Input	Actions	Recognition rate%
T Dobhal et al. [14]	Silhouette	9	98.5
T Dobhal et al. [14]	Silhouette	5	100
Chaaraoui et al. [8]	Silhouette	9	61.1
S Maity et al. [40]	Silhouette	9	95.06
Proposed (KTH dataset)	Silhouette	15	97.83
Proposed (Weizmann dataset)	Silhouette	10	97.66
Proposed (Weizmann dataset(5 actions))	Silhouette	5	100
Proposed (Ixmas dataset)	Silhouette	10	92.18

5 Conclusion

In this paper, a new method for the recognition of human action based on HBMI is introduced. We took as a basis a series of silhouettes of human actions as a representation of characteristics. The obtained human silhouette with background subtraction using a cpmbination between Gaussian Mixing Model (GMM) and the uniform motion of Magnitude of Optical Flow (MOF). Thus, we use a fast dynamic frame skipping technique to avoid frames that contain irrelevant motion, to decrease the computational complexity and to accelerate the process of silhouette extraction. Furthermore, Stacked Sparse Autoencoder framework is presented for automated human action detection. The Stacked Sparse Autoencoder model can capture high-level feature representations of pixel intensity in an unsupervised manner. These high-level features enable the classifier to work very efficiently for detecting multiple actition from a large cohort images. this new HBMI is used as the initial input in our classifier SMC. Our method is independent of the style of the individuals due to the use of binary foreground masks. Therefor, it is stable at the speed of action. Good recognition results have been obtained without compromising the relevance of the method. As a perspective and a continuation of this work, we could extend our algorithm for use in detecting abnormal behavior in a video stream.

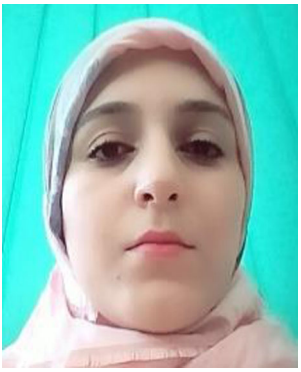
Acknowledgements The authors would like to acknowledge the financial support of this work by grants from General Direction of scientific Research (DGRST), Tunisia, under the ARUB program.

References

1. Abdessamad J, ElAdel A, Zaied M (2017) A sparse representation-based approach for copy-move image forgery detection in smooth regions. In: Ninth international conference on machine vision (ICMV 2016). International Society for Optics and Photonics, vol 10341, p 1034129
2. Abidine MB, Fergani B Evaluating a new classiffication method using pca to human activity recognition. In: Proceeding of International Conference on Computer Medical Applications (ICCMA). https://doi.org/10.1109/ICCMA.2013.6506158
3. Bellil W, Amar C, Ben ZM et al (2004) La fonction Beta et ses dérivées: vers une nouvelle famille d'ondelettes. In: First international conference on signal, system and design, SCS, pp 201–207
4. Benezeth Y, Jodoin PM, Kulkarni BM (2010) Histogram based foreground object extraction for indoor and outdoor scenes, ICVGIP
5. Blank M, Gorelick L, Shechtman E et al (2005) Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005. IEEE, pp 1395–1402
6. Bobick A, Davis J The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence
7. Bradski1 GR, Davis JW (2002) Motion segmentation and pose recognition with motion history gradients. Mach Vis Appl 13:174–184
8. Chaaraoui A, Climent-Prez P (2013) Silhouette-based human action recognition using sequences of key poses. In: Pattern Recogn Lett Elsevier, vol 34, pp 1799–1807
9. Chandrashekhar V, Venkatesh K (2006) Action energy images for reliable human action recognition. Action energy images for reliable human action recognition
10. Chang Z, Ban X, Shen JG (2015) Research on three-dimensional motion history image model and extreme learning machine for human body movement trajectory recognition. Mathematical Problems in Engineering
11. Chaudhry R, Oi F, Kurillo G, Bajcsy R (2014) Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 315 '13), pp 471–478
12. Chen C-Y, Hsu C-T, Yeh C-H, Chen M-J (2007) Arbitrary frame skipping transcoding through spatial-temporal complexity analysis. In: IEEE Conference on Region 10 Conference TENCON, pp 1–4

13. Cheriyadat AM (2014) Unsupervised feature learning for aerial scene classification. IEEE Trans Geosci Remote Sens 52:439–451
14. Dobhal T, Shitole V, Thomas G, Navada G (2015) Human activity recognition using binary motion image and deep learning. In: Proceeding of Computer Science Elsevier, vol 58. https://doi.org/10.1016/j.procs.2015.08.050
15. Ejbali R, et Zaied M (2018) A dyadic multi-resolution deep convolutional neural wavelet network for image classification. Multimed Tool Appl 77(5):6149–6163
16. Ejbali R, Zaied M, et Amar CB (2010) Intelligent approach to train wavelet networks for recognition system of arabic words. In: KDIR, pp 518–522
17. Ejbali R, Zaied M, et Amar CB (2013) Face recognition based on beta 2D elastic bunch graph matching. In: 2013 13th International Conference on Hybrid Intelligent Systems (HIS). IEEE, pp 88–92
18. ElAdel A, Ejbali R, Zaied M, Amar CB (2016) A hybrid approach for Content-Based Image Retrieval based on Fast Beta Wavelet network and fuzzy decision support system. Mach Vis Appl 27(6):781–799
19. Fast DCNN based on FWT, intelligent dropout and layer skipping for image retrieval
20. Gnouma M, Ejbali R, et Zaied M (2017) Human fall detection based on block matching and silhouette area. In: Ninth International Conference on Machine Vision (ICMV 2016). International Society for Optics and Photonics, p 1034105
21. Gnouma M, Ejbali R, et Zaied M (2018) Abnormal events' detection in crowded scenes. Multimedia Tools and Applications, 1–22
22. Hassairi S, Ejbal R, Zaied M (2015) A deep convolutional neural wavelet network to supervised arabic letter image classiffication. In: 15th International Conference on Intelligent Systems Design and Applications (ISDA). https://doi.org/10.1109/ISDA.2015.7489226
23. Hassairi S, Ejbal R, Zaied M (2016) Supervised image classiffication using deep convolutional wavelets network. In: 27th International Conference on Tools with Artifficial Intelligence (ICTAI). https://doi.org/10.1109/ICTAI.2015.49
24. Hassairi S, Ejbal R, Zaied M (2017) A deep stacked wavelet auto-encoders to supervised-feature extraction to pattern classiffication. In: Multimedia Tools and Applications. Springer. https://doi.org/10.1007/s11042-017-4461-z
25. Hassairi S, Ejbali R, et Zaied M (2016) Sparse wavelet auto-encoders for image classification. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, pp 1–6
26. Horn BKP, Schunck BG (1981) Determining optical flow. Artif Intell 17:185–203
27. Hwang J-N, Wu T-D, Lin C-W (1998) Dynamic frame-skipping in video transcoding. In: IEEE Conference on Works Multimedia Signal Processing, pp 616–621
28. Jalal A, Uddin M, Kim T (2012) Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. IEEE Trans Consum Electron 58(3):863–871
29. Jemai O, Ejbali R, Zaied M et al (2015) A speech recognition system based on hybrid wavelet network including a fuzzy decision support system. In: Seventh International Conference on Machine Vision (ICMV 2014). International Society for Optics and Photonics, pp 944–503
30. Jia K, Yeung D (2008) Human action recognition using local spatio-temporal discriminant embedding, IEEE Conference Computer Vision and Pattern Recognition
31. Karthikeyan S, Gaur U, Manjunath B (2011) Probabilistic subspace-based learning of shape dynamics modes for multi-view action recognition. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops)
32. Ke S, Thuc H, Lee Y, Hwang J, Yoo J (2013) A review on video based human activity recognition. https://doi.org/10.3390/280computers2020088
33. Khatrouch M, Gnouma M, Ejbali R et al (2018) Deep learning architecture for recognition of abnormal activities. In: Tenth International Conference on Machine Vision (ICMV 2017). International Society for Optics and Photonics, p 106960F
34. Ladjailia A, BOUCHRIKA I, Harrati N et al (2018) Encoding human motion for automated activity recognition in surveillance applications. In: Computer vision: Concepts, Methodologies, Tools, and Applications. IGI Global, pp 2042-2064
35. Ladjailia A, Bouchrika AL, Merouani H (2016) On the use of local motion information for human action recognition via feature selection. In: 4th International Conference on Electrical Engineering (ICEE). https://doi.org/10.1109/INTEE.2015.7416792
36. Li ZZW, Liu Z (2008) Expandable data-driven graphical modeling of human 320 actions based on salient postures. In: IEEE Transactions on Circuits and Systems for Video Technology, pp 1499–1510
37. Liu H, Ju Z, Ji X et al (2017) Study of human action recognition based on improved spatio-temporal features. In: Human Motion Sensing and Recognition. springer, Berlin, pp 233–250
38. Lucas BD, Kanade T et al (1981) An iterative image registration technique with an application to stereo vision

39. Lv F, Nevatia R (2007) Single view human action recognition using key pose matching and viterbi path searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8
40. Maity S, Bhattacharjee D, Amlan C (2016) A novel approach for human action recognition from silhouette images. Elsevier IETE Journal of Research
41. Mariem G, Ridha E, Mourad Z (2016) Detection of abnormal movements of a crowd in a video scene. In: International Journal of Computer Theory and Engineering, pp 398–402
42. Meng B, Liu XJ, et Wang X (2018) Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos. Multimedia Tools and Applications, 1–18
43. Qi J, Yang Z Learning dictionaries of sparse codes of 3d movements of body joints for real-time human activity understanding, Journals PloS One. https://doi.org/10.1371/journal.pone.0114147
44. 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
45. Seo J-J, Kim H-I, De Neve W et al (2017) Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection. Image Vis Comput 58:76–85
46. Singh V, Nevatia R (2011) Action recognition in cluttered dynamic scenes using pose-speciffic part models. In: Proceedings of IEEE International Conference on Computer Vision, pp 113–120
47. Sivagami M, Revathi T, et Jeganathan L (2017) An optimised background modelling for efficient foreground extraction. Int J High Performance Comput Netw 10(1-2):44–53
48. Teyeb I, Jemai O, Zaied M et al (2014) A novel approach for drowsy driver detection using head posture estimation and eyes recognition system based on wavelet network. In: The 5th International Conference on Information, Intelligence, Systems and Applications, IISA 2014. IEEE, pp 379–384
49. The data is available on the perception website http://perception.inrialpes
50. Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identiffication. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 1505–1518
51. Wang ZL, Wu Y (2014) Learning actionlet ensemble for 3d human action recognition. Part Ser Springer Briefs Comput Sci 260:11–40
52. Willems TTG, Gool LV An eficient dense and scale-invariant spatio-temporal interest point detector, Proceeding of the 10th European Conference on Computer Vision
53. Willems TTG, Gool LV An eficient dense and scale-invariant spatio-temporal interest point detector, Proceeding of the 10th European Conference on Computer Vision
54. Yu S, Cheng Y, Su S et al (2017) Stratified pooling based deep convolutional neural networks for human action recognition. Multimed Tool Appl 76(11):13367–13382
55. Yu ZL, Yuan J (2014) Iscriminative orderlet mining for real-time recognition of human-object interaction. In: Proceedings of the Asian Conference on Computer Vision
56. Zaied M, Mohamed R, et Amar CB (2012) A power tool for content-based image retrieval using multiresolution wavelet network modeling and dynamic histograms. In: International Review on Computers and Software (IRECOS), vol 7
57. Zhen X, Shao X (2014) Action recognition by spatio-temporal oriented energies, Information Sciences, Elsevier



Mariem Gnouma received the license degree in Computer Science from Faculty of Sciences of Gabes (FSG), and the master degree of computer science and multimedia of the higher Institute of Computer and Multimedia of Gabes (ISIMG) respectively in 2012,2015. she is pursuing the Ph.D. degree in Computer Engineering in the Research Team on Intelligent Machines (RTIM_Gabes), University of Gabes.

Ammar Ladjailia was born in Souk Ahras, Algeria, on November the 29nd 1977. He received the computer science Engineering degree from the University of Annaba in 2000. He obtained a Magister Degree in Computer Science from the University of Annaba in 2003. Ladjailia is now working towards her PhD degree at the Image Processing Research Group at the University of Annaba, Algeria. Whilst he is working as assistant lecturer of Computer Science from the university of Souk Ahras. Her research includes Human activities recognition, Smart Automated Visual Surveillance, image and video processing.

Ridha Ejbali received the Ph.D. degree in computer engineering, the Master degree and computer engineer degree from the National Engineering School of Sfax Tunisia (ENIS) respectively in 2012, 2006 and 2004. He was an assistant technologist at the Higher Institute of Technological Studies, Kebili Tunisia since 2005. He joined the Faculty of Sciences of Gabes Tunisia (FSG) where he is an assistant in the Department computer sciences since 2012. His research area is now in pattern recognition and machine learning using Wavelets and Wavelet networks theories.

Mourad Zaied He received the HDR, the Ph.D. degrees in computer engineering and the master degree of science from the National Engineering School of Sfax respectively in 2013, 2008 and in 2003. He obtained the degree of computer engineer from the National Engineering School of Monastir in 1995. Since 1997 he served in several institutes and faculties in university of Gabes as teaching assistant. He joined in 2007 the National Engineering School of Gabes (ENIG) as where he is currently an associate professor in the Department of Electrical Engineering. His research interests include computer vision and image and video analysis. These research activities are centered around wavelets and wavelet networks and their applications to data classification and approximation, pattern recognition and image, audio and video coding and indexing. He was the chair of the Workshop on Intelligent Machines: Theories & Applications (WIMTA II 2009) and he organized two winter schools on wavelet and its applications (2005) and on Matlab toolkits (2004).