Vision Based Human Activity Recognition: A Review

Allah Bux, Plamen Angelov and Zulfiqar Habib

Abstract Human activity recognition (HAR) is an important research area in computer vision due to its vast range of applications. Specifically, the past decade has witnessed enormous growth in its applications, such as Human Computer Interaction, intelligent video surveillance, ambient assisted living, entertainment, human-robot interaction, and intelligent transportation systems. This review paper provides a comprehensive state-of-the-art survey of different phases of HAR. Techniques related to segmentation of the image into physical objects, feature extraction, and activity classification are thoroughly reviewed and compared. Finally, the paper is concluded with research challenges and future directions.

Keywords Computer vision • Human activity recognition • Objects segmentation • Feature extraction • Action recognition • Review

1 Introduction

In recent years, automatic human activity recognition (HAR) based on computer vision has drawn much attention of researchers around the Globe due to its promising results. The major applications of HAR include; Human Computer Interaction (HCI), intelligent video surveillance, ambient assisted living, human-robot interaction, entertainment, and video indexing, etc. In HCI, activity

A. Bux (✉) · P. Angelov
School of Computing and Communications Infolab21,
Lancaster University, Lancester LA1 4WA, UK
e-mail: a.bux@lancaster.ac.uk

P. Angelov
e-mail: p.angelov@lancaster.ac.uk

A. Bux · Z. Habib
Department of Computer Science, COMSATS Institute
of Information Technology, Lahore, Pakistan
e-mail: drzhabib@ciitlahore.edu.pk

P. Angelov et al. (eds.), *Advances in Computational Intelligence Systems*,
Advances in Intelligent Systems and Computing 513,
DOI 10.1007/978-3-319-46562-3_23

recognition systems observe the task carried out by the user and guide him/her to complete it by providing a feedback. In video surveillance, activity recognition system can automatically detect a suspicious activity and report to the authorities for immediate action. Similarly, in entertainment, these systems can recognize the activities of different players playing a game. Activity can be a single person action to multiple people activities and behaviour recognition which may consists of sequence of actions and their context.

A number of surveys have been published on the processes of activity recognition during the last decade. A survey on conventional and recent methods of background modelling is presented in [1]. It discussed how to handle critical situation such as occlusion, viewing invariance, and illumination. It also covered the available resources, libraries and datasets for foreground detection and highlighted future research directions in this area.

Another survey on recognition of human activities is presented in [2]. In this survey, activities were categorized based on the complexity and recognition methodologies. Different challenges in HAR were discussed in [3]. In this study, authors also discussed the progress and limitations of the state-of-the-art techniques and identified the future research directions. An important study was presented in [4] where three levels of HAR, including core technology, HAR systems, and their applications were presented and discussed. They also discussed the abnormal activity and crowd behaviour recognition.

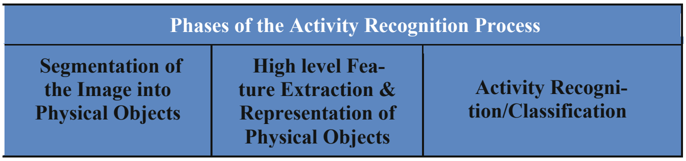
According to the level of the complexity, human activities can be categorized as "action" and "activity". Usually, an action is performed by a single person and activity is performed by multiple people. In [5], authors discussed HAR methods for four groups of activities (atomic action, human interactions, group activities, and human-object interaction). They classified HAR techniques into single-layered approaches and hierarchical approaches. Single-layered approaches recognize the simple activities directly from the video data while hierarchical approaches recognize more complex activities by decomposing them into simple activities (sub-events).

Vision-based human recognition systems are significantly affected by challenges such as occlusion, anthropometry, execution rate, background clutter, and camera motion. Research reported in [6] presents existing methods and their abilities for handling the above mentioned challenges. Moreover, it identifies the publicly available datasets and challenges that the activity recognition faces. Based on these challenges, potential research areas were also identified. Another study was conducted in [7] on HAR methods using 3-D data that include depth information. It identifies the advantages and disadvantages of these methods and indicates future research directions in this area.

A review article on semantic-based human recognition methods is presented in [8]. It presents state-of-the-art methods for activity recognition that use semantic-based features. In this research semantic space, and semantic-based features such as pose, poselet, related objects, attributes, and scene context were defined. A review on pixel-wise background subtraction techniques was proposed in [9]. It compares and contrasts the attributes and capabilities of most prominent pixel-wise techniques. However, all these surveys discussed above do not cover the comparison and critical analysis of segmentation, feature extraction and representation, and activity classification techniques, which is very essential for freshmen as well as experienced researchers to identify the research problems. We present a deeper analysis of the basic three phases (foreground detection, feature extraction and representation, and classification) of HAR in a single article by covering more recent articles. We also present the current status of the research and challenges that are still unaddressed.

In vision based HAR, a video object is segmented from its background using different object segmentation techniques. After the segmentation, important characteristics of the silhouettes are extracted and presented as a set of features. Then, these features are used for classification using any classifier. Figure 1 shows the framework of the present study. Video object segmentation methods have been categorized into background construction-based methods, and foreground extraction-based methods. In background construction-based methods, the camera is static, hence the background information is obtained in advance and the model is build up for object segmentation. In the latter case, the videos are captured by the moving camera; consequently, both object and the background are also moving. Hence, the background model cannot be built in advance. Therefore, the model is obtained online.

Feature extraction and representation methods have been categorized into global, local and semantic-based methods. Global methods use global features which consider image as a whole, while local methods use local features which operate on a pixel level. Semantic-based methods use semantic-based features which represent high level action of the human body such as pose, poselet, attributes, etc. In addition to this, classification models have been surveyed and discussed. These models have been categorized based on the classifier used for HAR. The rest of the paper is organized as follows. Segmentation of image into physical objects is presented in Sect. 2, feature extraction and representation methods are presented in Sect. 3, Sect. 4 presents the state-of-the-art methods for activity classification, discussion and the conclusion are presented in Sect. 5.



Phases of the Activity Recognition Process		
Segmentation of the Image into Physical Objects	High level Feature Extraction & Representation of Physical Objects	Activity Recognition/Classification

Fig. 1 Overview of HAR methods based on computer vision

2 Segmentation of the Image into Physical Objects

Object Segmentation is considered as a basic phase of the HAR process. The purpose of this phase is to extract the required objects from the sequence of images. The regions of interest in the foreground of the image are usually extracted. Based on the background information obtained in advance or detected at a later stage, object segmentation can be categorized as background construction-based segmentation and foreground extraction-based segmentation as shown Fig. 2, and detailed comparison of these methods can be found in Table 1.

2.1 Challenges and Issues in Segmentation

Noisy image: Image may be of poor quality due to the image source or image compression.

Camera jitter: In situations, when camera sways back and forth due to wind, it causes nominal motion in the sequence. In this case, without robust maintenance scheme false detection cannot be avoided.

Camera automatic adjustments: Due to automatic adjustment features of modern cameras such as auto focus, auto brightness, auto white balancing, etc., this may affect the colour level adjustment between different frames.

Illumination changes: Due to sudden changes of illumination in an indoor or outdoor scene, the false detection of foreground mask can occur in several parts of the image.

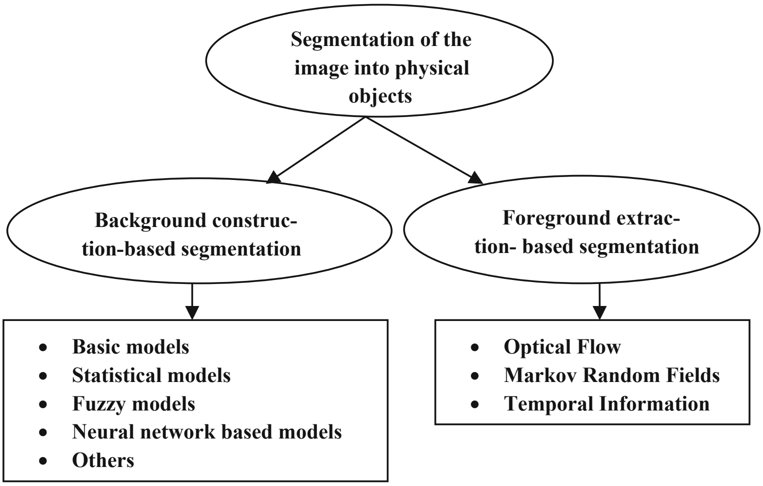

Segmentation of the image into physical objects
Background construc-tion-based segmentation
Foreground extrac-tion- based segmentation
Basic models
Statistical models
Fuzzy models
Neural network based models
Others
Optical Flow
Markov Random Fields
Temporal Information

Fig. 2 Segmentation of the image into physical objects

Table 1 Comparison of segmentation models

Segmentation type	Model and references	Advantages	Limitations
Background subtraction based methods	Basic [14, 16, 17]	• Simple and computationally less expensive	• Cannot handle multimodal and complex backgrounds
-do-	Statistical [19–24]	• Can handle multimodal backgrounds • Adaptive in nature, and different threshold selected for each pixel are adapted with time • Parameters are updated adaptively without keeping large buffer of frames in the memory	• Cannot effectively handle sudden changes in the scene and drastic lighting changes • Require many parameters some are selected manually • Appropriate initialization of the Gaussian is important
-do-	Fuzzy [27–30]	• Useful for modelling uncertainties in dynamic backgrounds • Robust to shadow detection and illumination changes • Helpful in background maintenance in case of shadow and illumination changes	• The threshold value is estimated by trial and error, which increases the detection time • Moving objects with same gray level as their backgrounds are not detected accurately
-do-	Neural Network [33–35]	• Can handle light changes, moving backgrounds, camouflage, and bootstrapping problems	• Over-fitting problem may affect the accuracy of results
	Others [36–42]	• Eigen space-based models have low computational complexity	• Cannot handle dynamic backgrounds effectively
Foreground extraction based methods	Optical flow [44, 45]	• Useful for segmentation of videos captured by a moving camera • Can handle occlusion and distortion	• Needs to calculate dense optical flow over long time frames, which is time consuming
-do-	Temporal information [52]	• Useful for segmentation of videos captured by moving camera	• Requires camera motion compensation, which may be sensitive to noise due to consecutive image differencing

(continued)

Table 1 (continued)

Segmentation type	Model and references	Advantages	Limitations
		• Simple and computationally less expensive	
-do-	Markov Random Fields (MRFs) [47, 50]	• Preserves boundaries for segmented objects • Effective for complex backgrounds	• Computationally expensive

Bootstrapping: If representative background is not available during training, then representative background image cannot be computed.

Camouflage: If foreground pixels are incorporated in the background model, then background and foreground cannot be distinguished.

Foreground aperture: Changes may not be detected in moving objects with uniformly coloured regions. Due to this, complete object may not appear as a foreground.

Moving background objects: Background objects can be moving but these should not be considered as foreground.

Inserted background objects: If new background object is inserted, this object should not be treated as a part of the foreground. In some cases this background object is detected as foreground if a robust maintenance scheme is not available.

Dynamic backgrounds: Background can be "shilly-shally", which results in false detection.

Walking person: Any movement of a background object will result in a change in the background.

Sleeping foreground object: A foreground sleeping object cannot be distinguished from its background.

Shadows: Shadows can come from the background of a moving object but will appear as a foreground object.

In order to handle the above mentioned challenges, different background models have been proposed [10]. These models have been discussed in the subsequent sub-sections in more detail.

2.2 Background Construction-Based Segmentation

In background construction-based segmentation, the background model is first constructed, and then the object is identified from the successive video frames using background subtraction method. In fact, background subtraction is achieved by differencing the current frame and the background. This type of segmentation is effective for tacking fast moving objects from the videos captured by static cameras. These methods have low computational cost, and are easy to implement. Object segmentation methods have been studied since 1990s, and many methods have been proposed for segmenting objects from their backgrounds. One of the background subtraction libraries provides implementation of 29 background subtraction algorithms [11]. Generally, background subtraction methods consist of the following three steps:

Background initialization: The purpose of this step is to initialize the background model. A model can be designed in different ways; using statistical, fuzzy, or neural networks techniques. It is often assumed that the background model can be initialized by some blank frames at the start of the video sequence. But, this approach might not be effective in the presence of cluttered backgrounds. Often, the model is initialized by the first frame of the video but it is quite challenging to get first background model when foreground objects are present in half of the training frames.

Background maintenance: This step is aimed at updating the background model with respect to the changes which occur in the scene. This has to be incremental. There are different maintenance and update schemes i.e., blind, selective, and adaptive fuzzy schemes [1]. The blind scheme has a common rule for updating all pixels. The main disadvantage of this scheme is that pixels classified as foreground are used in the computation of new background, which results in polluted background image [12]. To overcome this problem, selective maintenance scheme was introduced which updates the background using different learning rates based on the experience of the previous classification of pixels. In this case, erroneous classification may result in a permanently incorrect background model. This issue is resolved by the adaptive fuzzy scheme which considers the uncertainty factor during the classification.

Foreground detection: This step is aimed at labelling pixels during classification as background or foreground pixels. Generally, background subtraction methods show better results when the camera is fixed, the background is static, and illumination is constant. However, in real world applications these conditions cannot be maintained. For example in [13] ten important challenges were highlighted for background subtraction techniques in video surveillance. Three more challenges were introduced in [10]. These challenging situations disturb the ideal conditions in which background subtraction models produce best results. These challenges are described in the following sub-section.

2.2.1 Basic Models

In case of stationary camera, usually, the background is also stationary then any change in the scene is considered as moving foreground objects. Different basic models were proposed for background subtraction of simple backgrounds. One of the simplest ways to develop a background model is to build an average model of the scene, subtract each video frame from it, and threshold the results.

One important technique was proposed in [14]. It is using discriminative texture features for background modelling, and a modified version of the Local Binary Pattern (LBP) operator for feature extraction. Foreground detection was performed by comparing the LBP with the background histograms using the same proximity measure.

A multi-layer background subtraction method was proposed in [15] which is an extension of the proposed method in [14]. Authors used LBP, and photometric invariant colour measurement in RGB colour space, combining the advantages of both texture and colour features. LBP operator performs well with respect to light variations on rich texture regions, but it is not efficient on uniform texture regions. They also introduced an update strategy in background models for handling moving backgrounds such as waving trees, and shadows cast. Finally, bilateral filter was used to remove noise and enhance foreground objects as a post-processing step.

A method based on a combination of local colour, intensity and texture features was proposed in [16]. It used Double Local Binary Pattern (DLBP) operator for feature extraction, which is a modified version of the classical LBP. Generally, there are three classes in gray scale images for representing change i.e., ascending > 0, homogenous $= 0$, and descending < 0, respectively. But, a classical LBP method divides the gray differences into two classes. It cannot differentiate between ascending and homogenous gray differences, and does not represent the texture efficiently. In [17], all these cases were handled using the DLBP operator.

A score board algorithm for estimating stationary background was proposed in [18]. This algorithm was used to record the intensity variations of pixels between previously estimated background and the current image. The small variations were assigned positive scores, and large variations were assigned negative scores. In this way, an accumulative score was calculated. This score was used for estimation of current background. A running weighted average was calculated for the positive score and running modulus was calculated for the negative score. Since, the positive score indicates small variation; a weighted running average was used for background estimation. In case of a negative score, the running modulus method was used for the background estimation.

A background modelling method based on a subpixel edge map was proposed in [17]. The edge position and orientations are modelled using a Gaussian mixture model (GMM). This method is suitable for detecting foreground objects with cluttered background and handling illumination changes. Basic background subtraction models are simple and computationally less expensive. However, these methods cannot handle more realistic multimodal and complex backgrounds.

2.2.2 Statistical Models

A single Gaussian per pixel model was used for real time tracking of people and interpreting their behaviour [19]. This system reports good results in an indoor scene but there have been no reports for its success in outdoor scenes. However, this model has limitations to complex distribution of pixel values and more elaborate models are required. A GMM was used in [20]. This probabilistic model classifies each pixel as a shadow, moving object, or a background with an unsupervised learning algorithm. One of the most common methods for updating GMM was proposed in [21]. This is an adaptive method in which background pixels are determined based on the persistence and variance of the mixture of Gaussian. This system reports good results with lighting changes, cluttered regions, slow-moving objects, and inserting or removing objects from the scene. The GMM model proposed in [21] was further extended in [22]. This method presents a framework consisting of two algorithms for pan-tilt camera of a mobile robot.

This is the first algorithm that handled motion blur, inaccurate motion estimations, geometric calibration errors, mixed pixels and motion boundaries, and used Bayesian approach for uncovering the background online. Adaptive GMM was also used in [23, 24]. This method used recursive equations for updating parameters and selecting appropriate number of components for each pixel. In [25], each pixel was defined as a 3D multivariate Gaussian rather than as a mixture of Gaussian distributions. The mean and variance were estimated by recursive Bayesian learning. In this way, the multimodality of the background is preserved and the necessary layers for representing each pixel are estimated. A hierarchical background model was proposed in [26]. This model combined the pixel-based and block-based approaches into a single framework. In addition to this, a novel descriptor for block-based modelling was introduced for achieving a coarse background model. This method achieved good results in case of IP-based and other intelligent cameras.

Statistical methods often use GMM. These models offer several advantages. (1) They can effectively handle multimodal backgrounds. (2) They are adaptive, different thresholds selected for each pixel is adapted in time. (3) These are parametric models and parameters are updated adaptively without keeping large buffer of frames in the memory. However, these models have several limitations as well. (1) These models cannot effectively handle sudden changes in the scene and drastic lighting changes. (2) These methods require many parameters that should be selected appropriately and are user and problem specific. (3) Appropriate initialization of the Gaussian is also an issue that needs to be handled.

2.2.3 Fuzzy Models

Different models have been proposed for background subtraction and foreground detection using fuzzy techniques. For example, in [27] for background subtraction using Sugeno fuzzy integral for aggregating the colour and texture features was proposed. The objects were detected by thresholding the results. Authors handled small motion of background objects such as bushes and swaying tree branches. In [12], colour and textures similarity measures were integrated using Choquet integral for foreground detection. A Choquet integral is more suitable for cardinal aggregation and showed better results in case of illuminations changes, shadows and background changes as compared to Sugeno integral [27]. A background modelling algorithm for infrared videos was proposed in [28] based on type-2 fuzzy mixture of Gaussian models.

It has to be stressed that colour information alone is not sufficient for handling dynamic environments, while edge and texture features alone are not sufficient for handling uniform texture regions. Thus, [29] proposed a background modelling technique based on texture, colour, and edge features. Authors integrated all these features using Choquet fuzzy integral to avoid uncertainties in classification. They also introduced an edge gray scale confidence map and texture confidence map. Then, a median filter and connected component algorithm were used for labelling disconnected regions and for noise removal. A background subtraction technique based on the fuzzy logic inference rules for dynamic environments was developed in [30].

Fuzzy logic based methods are used in different steps of the background subtraction process. They offer several advantages as compared to statistical models discussed earlier. Firstly, these methods are useful for modelling uncertainties in dynamic backgrounds. Secondly, these methods are more robust to shadow detection and illumination changes. Thirdly, these methods are also helpful in background maintenance in case of shadow and illumination changes [31]. However, there are some drawbacks of fuzzy logic based methods; firstly, the threshold value is estimated by trial and error, which increases the detection time. Secondly, the moving objects with same gray level as their backgrounds are not detected accurately [32].

2.2.4 Neural Network Based Models

In this class of models each pixel of the sequence is classified either as a foreground or background using artificial neural networks. A background subtraction approach based on self-organization neural networks was proposed in [33]. The background is automatically generated by self-organization method with prior knowledge regarding involved patterns. Then, weight vectors of the network are initialized using pixel values. The HSV colour space is used for representing each weight vector. Finally, the weights are updated if the best match is found with the current pixel using selective weighted running average.

A multi-layer feed-forward probabilistic neural networks (PNN) for background subtraction with 124 neurons was proposed in [34]. The background model is trained and learned by the PNN, and Bayesian classifier is used for differentiating between the background and foreground pixels. An improvement to the previous work done in [33] is reported in [35]. Authors used the previous method based on self-organization using neural networks. They used fuzzy function to improve the robustness of the method against false detection rates, and to deal with decision problems in case of crisp settings.

Neural network based methods offer several advantages; these methods can handle light changes, moving backgrounds, camouflage, and bootstrapping problems [33]. However, neural network based methods have some limitations as well, such as over-fitting problem, which may affect the accuracy of results. Therefore, generalization and regularization are very important to be taken into account.

2.2.5 Other Models

An Eigen space model for background subtraction was proposed in [36]. The Eigen space model is built by computing mean background and covariance of N sample images. Eigen space-based background method for segmentation was employed due to its lower computational complexity. In [37] a non-parametric background subtraction model was presented. It is based on estimation of probability distribution, which is a probabilistic way to define a background model. Another foreground segmentation algorithm was proposed in [38] by combining the statistical background estimation with per-pixel Bayesian inference. More complex models have also been reported in the literature. For example in [39, 40] a background model using single mixture of Gaussians for each pixel was proposed. In [41, 42] a segmentation method using tracking was proposed. Unlike, point-based segmentation methods where the background model is built in advance, these methods used dynamic time wrapping algorithm for segmentation purpose.

2.3 Foreground Extraction-Based Segmentation

In foreground extraction-based segmentation, videos are captured by the moving camera. Hence, the background, foreground, and camera all move. In this case, camera may be installed on cars, moving robots, unmanned aerial vehicles (UAVs), or it can be a pan-tilt-zoom camera (PTZ). Object segmentation in these scenarios is very challenging as compared to the static camera because the object motion and camera motion are mixed. Therefore, techniques used for background construction-based videos are not effective for foreground extraction-based object segmentation.

In the foreground extraction-based segmentation, temporal information, spatial information, or spatio-temporal information is utilized for getting the initial object from video and then objects in the following frames are determined using change information, motion information, or any other feature-based information. Different techniques for foreground-extraction-based segmentation are discussed in the following sub-sections.

2.3.1 Optical Flow Based Methods

This is a classic approach for foreground extraction-based segmentation; it depends upon the distribution of movements of bright patterns in an image. It gives the information about the spatial arrangements of the objects [43]. A method for detecting moving objects from a video sequence captured by the moving camera was proposed in [44]. In this method, feature points were extracted using Harris corner detector [44], and then optical flow method is employed for feature matching in two consecutive frames. These features are classified as foreground or background using multiple view geometry. Then, foreground regions are obtained based on the foreground feature points, and image difference is calculated based on the background feature points using affine transform.

By merging the foreground regions and image difference, moving object regions are obtained. Ultimately, moving objects are detected based on the motion history. Some methods have been proposed for moving objects segmentation using point trajectories. The basic idea behind these types of method is that camera motion is restricted by some geometrical constraints while motion caused by the moving object is free of such kind of constraints. Thus, moving objects can be detected and segmented by analysing the trajectories of some key points. However, there are limitations of this type of methods; these methods need to calculate dense optical flow over long time frames, which is time consuming. An optical flow method using long term motion cues for motion estimation was proposed in [45]. This approach relies on point trajectories computed through optical flow with focus on those areas of the image where optical flow woks best. Authors of [45] also claim handling occlusions and distortion without any extra efforts. Optical flow based methods are computationally complex and are used for complex dynamic image analysis.

2.3.2 Markov Random Fields (MRFs) Based Methods

MRFs are used to model the foreground field and enhance the spatial and temporal continuity of the moving objects [46]. A method for real time object segmentation from videos captured by a moving camera was proposed in [47]. This method is based on the colour information and region merging through Markov Random Fields (MRFs). The same was used in [48]. It was applied to videos captured by both static and moving cameras. A method for moving vehicles detection from videos using joint random fields (JRFs) was proposed in [49], which is an extension to the MRF model. A region-matching-based method was proposed in [50] for object detection and tracking by moving cameras. It incorporates fuzzy edge information of each pixel in the MRF model. This method also preserves the boundaries of objects to be segmented. The maximum a posterior probability (MAP) principle is employed to address the issue of spatial segmentation, and moving objects in the subsequent frames using region-based estimation. However, this method has higher computational cost and might not be effective for real time applications.

2.3.3 Temporal Information Based Methods

Temporal differencing does not use background extraction process. It uses pixel wise difference between three consecutive frames in an image for extracting moving objects. Generally, differencing based techniques are poor in extracting all relevant pixels and leave some cavities inside the moving entities. This happens when the moving object has uniform textures or it is moving slowly. To overcome this issue, additional filtering algorithms might be needed [51]. A segmentation method for videos captured by a freely moving camera was proposed in [52]. This method uses two successive frames as input to extract the moving object. A two layer affine transformation model was introduced in this paper. The outer layer is used for filtering the background and foreground feature set based on the camera motion compensation model parameters. Therefore, these methods are employed in many human action recognition techniques. Some methods also use the combinations of spatial and temporal information for this purpose. Temporal information based methods are simple and computationally less expensive. There are some drawbacks of these methods as well. These methods require estimating camera motion, which is quite difficult, and may be sensitive to noise due to consecutive image differencing. Moreover, these approaches leave some cavities inside the moving objects white extracting moving pixels.

3 High Level Feature Extraction and Representation of Physical Objects

Feature extraction and representation is an important phase in activity recognition. Once the objects are segmented from the background, these are represented in the form of features such as shape, silhouette, colour, motion features, etc. Feature extraction and representation methods have been categorized into global, local, and semantic-based features. Global features consider image as a whole, while local features operate on a pixel level. Semantic-based features represent high level action of the human body such as pose, poselet, attributes, etc. and are more human intelligible. The overall performance of the activity recognition systems mainly depends on the proper feature extraction and representation mechanism. Different kinds of features and representation mechanisms have been proposed, these are discussed in the following section. The categorization of these features is shown in Fig. 3 and comparisons of these features have been presented in Table 2.

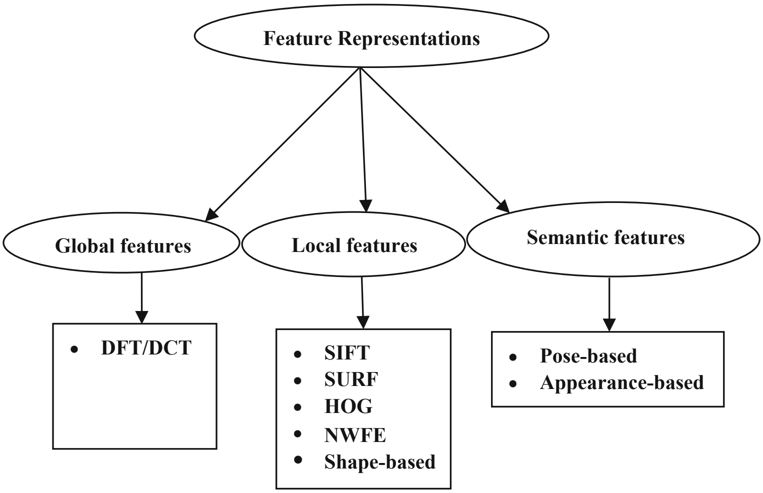


Fig. 3 Different types of feature descriptors

Table 2 Comparison of feature descriptors

Category	Descriptors and references	Advantages	Limitations
Global features	Space-Time Volumes (STVs) [100–104]	• Do not require background subtraction methods	• Not suitable for recognition of multiple people in a scene • Sensitive to noise and occlusion
-do-	Discrete Fourier Transform (DFT) [53, 54]	• Simplified processing and conversion to frequency domain • Do not require much computational resources	• It is not truly attainable in practice (we cannot sample a function for every x ∈ R)
Local features	Scale Invariant Feature Transform (SIFT) [55, 56]	• These features are invariant to scale, rotation and translation, and partially invariant to 3D projection, and illumination changes	• High dimensionality of data • It does not include colour information
	SURF descriptor [57–59]	• The SURF is several time faster and more robust than SIFT for different image transformations.	• Patented software • Low performance with high dimensionality of data
-do-	Histograms of Oriented Gradients (HOG) [60, 61]	• The tacking and recognition is handled into a single framework	• Extraction of local descriptors at fixed scale causes performance to be

(continued)

Table 2 (continued)

Category	Descriptors and references	Advantages	Limitations
		• This is invariant to viewpoints, poses, and illumination changes	influenced by the object size
-do-	Nonparametric Weighted [62–64]	• Reduces the effect of the singularity problem • Shows good performance even with non- normal datasets and reduces the effect of outliers	• No significant limitations noticed
-do-	Shape-based features [65–68, 105]	• Robustness to noise and rationality with the human perceptions	• Accurate silhouette segmentation is required
Semantic-based features	Pose estimation [70–75]	• Pose based methods do not suffer from inter-class variations	• Difficulty in pose extraction under realistic conditions
-do-	Appearance based [77, 78]	• Take into account the contextual information	• Suffers from intra-class variations

3.1 Global Features

Global features consider the whole image for feature extraction. Different techniques have been proposed using global features for activity recognition. These techniques are presented in the subsequent subsections.

3.1.1 Discrete Fourier Transform Features

Discrete Fourier Transform (DFT) has been employed in many image processing and computer vision applications. Generally, DFT represents the intensity variation of an image. It transforms an image from the spatial domain to frequency domain by dividing it into different spectral sub bands. A method using DFT features for action recognition was proposed in [53]. After normalization, the image frames were divided into small blocks and DFT features of each block were calculated. Then, the average of DFT features was calculated. These average DFT features were used for further classification. For classification purpose the K-nearest neighbour algorithm was used. In [54] an activity recognition method was proposed based on Discreet Cosine Transform (DCT). In fact, DCT is a special case of DFT when only real coefficients are considered. In [54], the DCT features are calculated from a tri-axial accelerometer data. Then, PCA is used to extract the most discriminative features from the set of original DCT features. These features are then passed to an SVM [54] for classification of activities. The major advantage of DCT is the simplified processing, and the fact that its conversion to frequency domain is computationally light.

3.2 Local Features

Unlike, global features, these features consider the specific points of interests in the image rather than the whole image. Generally, these features are robust to noise and occlusion and invariant to scale, rotation, and transformation. These features are discussed in the subsequent sections.

3.2.1 Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) is an algorithm for describing local features of an image [55]. These features are invariant to scale, rotation, translation, and partially invariant to 3D projection, and illumination changes. These features are generated by four steps of computation. The first step is scale-space extrema detection; this step is used to search scale and locations in the image. This is implemented by difference-of-Gaussians function to detect invariant points of interest. The second step is localization of key points; key points are localized, and selected on the basis of a stability measure. In the third step, an orientation is assigned to each key point based on gradient directions of the image. The last step is key point descriptor when a local image gradient of selected regions is measured around each key point. These image gradients are transformed for illumination changes and local shape distortion and used for object recognition and classification.

In [56], 3D scale invariant feature transform (SIFT) descriptor for encoding the local space and time information in video was used. In a way, this allows robustness to noise and orientations. Thus, extended "bag of words" paradigm from 2D to 3D, where the third dimension is time. The work in [56] described videos as "bags of spatio-temporal words" using a 3D SIFT descriptor. In [56] they discovered relationship between words for forming spatio-temporal word grouping on the basis of co-occurrences of words. These word groupings were further used for action recognition.

The SIFT descriptor is well known for invariance to scale, rotation, and robustness to noise and illumination changes. However, there are some drawbacks of the SIFT descriptor such as the high dimensionality. Image gradients are represented by highly dimensional vectors, and features are not sufficiently discriminative in some cases such as features found on the background and on humans object. Moreover, the SIFT descriptor is based on grayscale information.

3.2.2 SURF Descriptor

The Speeded-Up-Robust-Feature is a feature detector and descriptor suitable for the tasks such as object recognition, human action recognition etc. It was introduce by [57] in 2006 at European Conference on Computer Vision. Based on moving SURF interest points, a novel spatio-temporal feature descriptor was proposed for human action recognition [58]. Authors reported that the proposed method is equally useful for controlled videos such as KTH dataset and uncontrolled videos such as YouTube. In [59], a SURF descriptor and dense optical flow were used to find the matches between frames and homography was estimated with RANSAC for human action recognition. Author claimed significantly improved results as compared to Histograms of Optical Flow (HOF) and Motion Boundary Histograms (MBH) descriptors. The SURF descriptor is several time faster and more robust than SIFT for different image transformations.

3.2.3 Histograms of Oriented Gradients Features

Histogram of Oriented Gradients (HOG) descriptor was used in [60] for detecting humans in videos. This descriptor employs normalized HOG in a grid. It considers scaled gradients, orientation binning, contrast normalization of overlapping blocks, and coarse spatial binning. A template-based algorithm based on PCA-HOG descriptor for tracking and recognition of athlete's actions was proposed in [61]. First, a HOG descriptor is used for feature extraction and then PCA is used on the extracted features. The HOG descriptor offers several advantages as compared to previous template-based approaches. The tacking and recognition is handled within a single framework. It is invariant to viewpoints, poses, and illumination changes. However, the drawback of the HOG descriptor is the extraction of local features at fixed scale, which causes the performance to be influenced by the object size.

3.2.4 Nonparametric Weighted Feature Extraction (NWFE)

A framework for HAR, based on subspace learning and motion information was proposed in [62]. By applying background subtraction, body silhouettes are extracted, and their boundaries are obtained by contour tracking. Then, the width and distance signal features are computed from the contour of the human pose. These two features are combined to create a pose-discriminative representation. Thus, the human activity is represented as a sequence of symbols, obtained by quantization of features into code-words. The K-means clustering algorithm is used to build the codebook. Then, NWFE approach is applied to measure the distance between these code-words.

In [63], a HAR method was proposed based on the curvature estimation of the human posture. These feature sequences are presented in the form of sets of strings. The NWFE method was employed for string matching. NWFE method offers several advantages: first, unlike parametric discrimination analysis, it allows to specify the number of desired features, thus, reducing the effect of singularity problem. Second, it shows good performance even with non-normal datasets and reduces the effect of outliers. Third, it assigns greater weights to samples near the expected decision boundaries, which increases the accuracy of the classification [64].

3.2.5 Shape-Based Features

The use of shape-based features in object recognition, and activity classification is reported in the literature. A method for comparing two stationary shapes for activity recognition and gait recognition was proposed in [65]. The shape deformations of the human silhouette during walking are exploited. Parametric models such as autoregressive (AR), autoregressive and moving average (ARMA) are employed as pattern classifiers for shape sequences, and applied for gait and activity recognition. An activity recognition method from short snippets of video was proposed in [66]. The motion and shape features are extracted in a biologically inspired manner. First, these two types of features are considered separately as high level features, and then combined at the classification stage. Moreover, authors also confirmed that, very short snippets (1–7 frames) are sufficient to recognize an action. In [67], a method was proposed using optical flow features combined with random sample consensus (RANSAC). The optical flow features are used to detect the direction and presence of motions, while RANSAC is used to localize and identify the most prominent motion within a frame. Finally, the Euclidean distance and SVM classifiers are used for classification.

Shape-based methods are successfully applied for action recognition as discussed above. These methods offer several advantages such as robustness to noise and rationality with the human perceptions [68]. However, this approach has some drawbacks as well. For shape based features extraction, accurate silhouette segmentation is required, which is difficult to achieve even with state of the art background subtraction models. However, this drawback can be minimized with optical flow based features which require no background segmentation. In [69] authors presented a region-based method for human action recognition. They have used surrounding space of a human silhouette known as negative space. These regions are naturally formed surrounding the human body inside the bounding box. They have reported 100 % accuracy on Weizmann dataset.

3.3 Semantic-Based Features

In linguistics, semantics is the study of meaning. For instance, when two people communicate through message passing, they infer the context of the message based on the semantic of the message. In action recognition, user applies prior knowledge for recognition process, which is based on the semantic understanding. Semantic understanding also plays a very important role in human visual perception. People analyse the body posture along with its physical and social settings for activity recognition. The use of semantic knowledge has been reported for vision-based HAR. Semantic-based features are more human intelligible if compared to other local and global features discussed above. In the semantic based methods, the first task is to detect the person in a video frame for feature extraction. This is done with the help of a bounding box or human contours to show the person's location in the frame and locate the region for feature extraction. Since, the human body is in different poses while performing different actions, poses can be extracted from the whole body or some body parts (poselets). Different techniques based on the semantic features are discussed in the following sections.

3.3.1 Pose Estimation Based Methods

Pose estimation is a challenging task in real world situations because poses are varied at different degrees while performing actions. In [70] a general categorization of pose estimation methods was proposed. Some of these methods use prior human model while other work without using any prior human model. Based on this information, these models are categorised into three classes described below.

Model-free: This class of models does not use any prior human model; it rather uses direct mapping from 2D image sequence to 3D pose. Different methods have been proposed under this category. A method for human torso tracking was proposed in [71]. In this method, a blob detection module consists of foreground detection; blob tracking was used for estimating the size and location of the torso. Then, other body parts such as head and hands are located with respect to the torso location.

Another method for human tracking, consisting of four steps was proposed in [72]. In the first step, the regions with human are extracted from the image while, in the second step, the simulated image is generated based on the information from the previous step. In the third step, the actual direction of the motion is found by comparing the newly captured image and a simulated image from the previous step. In the fourth step, the position of the human is updated based on the calculation of all similarity values.

Indirect model: This class of models use a prior human model for pose estimation. Different methods have been proposed under this class for the HAR and tracking. For example, a human body labelling system based on the motion information was proposed in [73]. This method consists of two processes; the first one extracts the outline of the moving human from an image sequence using edges, and the second process interprets the outlines and generates the human body stick figure for each frame of the image sequence.

Direct model: This class of models uses a 3D geometric depiction of the human shape and kinematics structure for reconstructing the pose. Different methods have been proposed under this category. In [74] a feature extraction method from a scanned body was proposed. Authors used the ISO 8559 semantic definition of the body parts and identified 21 features, and 35 feature lines on the human torso. The algorithm was tested on adult Asian females aging from 18 to 60. In [75] a view invariant method for shape and pose features extraction for pedestrian environment was proposed.

Pose-based methods have several advantages: First, these methods do not suffer from inter-class variances. Particularly, 3D poses are invariant to viewpoint and appearance. Second, pose representation simplifies the learning for action recognition due to extraction of high level information [76]. However, the major limitation of these methods is the difficulty in pose extraction under realistic conditions.

3.3.2 Appearance-Based Methods

A pose estimation method for 3D human pose estimation based on the HOG, and using PCA for dimensionality reduction was proposed in [77]. This method was named Local Appearance Context (LAC) descriptor. Relevance Vector Machine (RVM) was used for mapping between HLAC and the 3D pose space. The HLAC descriptor demonstrated better performance than SIFT descriptor for the HAR. In [78], an appearance-based method was proposed for people tracking. This method works in two stages; first, the model for each person was built from the video sequence and then, these models were tracked in each video frame. For building models, two algorithms were developed; first, the bottom-up approach was used and detected the human body parts in a sequence of a video. The second algorithm used top-down approach for building models and detecting the key poses from a sequence of another video. Building this type of discriminative models is helpful, since it exploits the structure of the background without background subtraction. However, these models are sensitive to clothing and illumination changes.

Appearance-based methods offer several advantages; first, unlike pose-estimation-based methods these methods do not require much high level processing. Second, these methods are not restricted to the human body and can take into account the contextual information as well. They are applicable in situations where pose estimation is difficult. However, these methods suffer from intra-class variations [76].

4 Activity Recognition and Classification

After feature extraction from a video, the next step is to select the suitable classification algorithm for activity recognition. The following sections discuss the different classifiers used for activity recognition. A detailed comparison of classification models is shown in Table 3.

Table 3 Comparison of classification models

Classification model	References	Advantages	Limitations
Hidden Markov Models (HMMs)	[83–85]	• Useful for recognition of more complex activities	• Large number of training videos are required for recognition of complex activities
Support Vector Machine (SVM)	[87–89, 101, 106, 107]	• Widely used due to its simplicity and good performance	• Generally, suitable for binary classification
Kalman Filter (KF)	[90, 91, 108]	• Produces good results with effective foreground segmentation	• Not effective for handling occlusion
Artificial Neural Network (ANN)	[92], [93]	• Better performance than its predecessor statistical models	• In complex model over fitting is likely to occur which affects the prediction ability of the classifier
K-Nearest Neighbour (KNN)	[53, 79]	• Model is simple and needs few parameters for tuning • Test time is not dependent on the number of classes • Robust with respect to the search space	• Performance is dependent on the selection of K • Different values need to be tried for selecting the best value of K
Multidimensional indexing	[94]	• Robust way of activity recognition and retrieval • Computationally less expensive	• Not fully view invariant
Deep Neural Networks (DNNs)	[96–99, 109]	• Better performance • Can act on raw input image	• Requires more training time • Much data required for training

4.1 K-Nearest Neighbour

The K-nearest neighbour (KNN) algorithm is a classification method based on a pre-defined constant K. A point is classified to a class which is most frequent among K nearest training points. In [79] KNN and accelerometer were used for activity recognition using mobile device. Another method based on KNN was proposed in [53] for human activity recognition using DFT features extracted from small image blocks.

The KNN algorithm is suitable for multi-modal activities and the decision of the classification is based on the neighbourhood of objects. It has several advantages in comparison to other classification algorithms: (1) the model is simple; (2) it needs few parameters for tuning; (3) test time is not dependent on the number of classes; (4) it is robust with respect to the search space, i.e., classes do not have to be linearly separable. However, the main drawback of this algorithm is that its performance is dependent on the selection of K. In order to select the best value of K, different values are tried during the training phase before classification.

4.2 Dynamic Time Wrapping (DTW)

The DTW algorithm, initially developed for speech processing, has been successfully employed for matching two video sequences. A method based on DTW was proposed in [80] for action recognition by utilizing 3D (XYZ) model-based body parts tracking. First, the 3D skeleton model of each frame is created, and movement is analysed. For this purpose, a 3D model of the human body parts composed of different segments and joint angles was obtained using multiple cameras. This model is also known as “stick figure with some degree of freedom”. The angle values were used as features representing human movement at each frame. These features were analysed using the DTW algorithm comparing them with a reference sequence of each action. The waving and twisting gestures were recognized using this method.

An extension to the DTW algorithm was proposed in [81] by taking time wrapping function into account for matching two sequences. Authors explicitly modelled inter and intra-personal variations of the speed of executing an activity. The action execution was modelled with the help of two functions; (1) the function that represents feature changes with respect to time; (2) function that represents the space of the time wrapping. This method was used for recognition of different actions such as throwing, pushing, waving and picking objects. Authors reported high accuracy for recognition of these actions. In [82], DTW was used for recognition of simple actions such as waving, clapping and punching. A depth camera was used to acquire 3D information of the human body parts. After acquiring this information, a feature vector was built using joint orientation of each body part. Then, dynamic time wrapping algorithm was used for classification.

The limitations of DTW-based methods are: firstly, these methods take into consideration the whole action and do not consider how the action is being performed. For this reason, these methods are not suitable for action localization and segmentation. Secondly, DTW-based algorithms take polynomial amount of computation for finding optimal non-linear match between two sequences. Moreover, as described above, these methods are suitable for simple actions and might not be effective for recognition of complex activities.

4.3 Hidden Markov Models

Hidden Markov Models (HMMs) have been widely used for action recognition. These models are suitable for activity recognition methods that represent an activity as the model consisting of a set of states. At each time frame, a human is considered to be in one state, which generates an observation, known as feature vector. In the next frame, a system transits into another state and the transition probabilities between the states are calculated. Once, the HMM is being trained then a certain activity can be recognized by evaluation of the sequence of actions performed. A method for human behaviour recognition was proposed in [83]. A set of local motion descriptors and trajectory features were used for action representation. Hidden Markov models were used for smoothing the sequence of actions. The behaviour recognition was achieved by computing the likelihood of the transition between two actions.

The simple HMM model is a sequential model, where one state is activated at a time. Due to its sequential nature, it has some limitations e.g. it cannot represent activities performed by multiple agents. In [36] a coupled hidden Markov model (CHMM) was introduced for modelling an interaction between two persons. A CHMM is a result of coupling multiple HMMs, where each HMM represents the motion of one agent. The Hierarchical Context-Hidden Markov Model (HC-HMM) was proposed in [84] for human behaviour understanding of elderly people at a nursing centre. This method infers human activities using three contexts which are activities, spatial, and temporal context. By considering hierarchical structure, HC-HMM builds different module for each context which is helpful for behaviour recognition.

A modification to the structure of HMM was proposed in [85]. In this method a quasi-periodic algorithm was used with HMM for HAR. This is a cyclic HMM, which is considered as a left-to-right model and having transitions from the ending to the start state. HMM-based methods have been used for recognition of more complex activities. However, large numbers of training videos are required for recognition of complex activities using this method.

4.4 Support Vector Machine

The Support Vector Machine (SVM) is a prominent supervised classifier for pattern recognition problems [86]. It is a binary classifier aiming to find an optimal hyper plane for maximizing the margin of separation between the two classes. This classifier has been widely used for human activity recognition due to its simplicity and good performance. In [87] SVM was used as a classifier for recognition of 50 human actions from web-based videos. Scene context descriptors and motion descriptors features were used for classification. These features were extracted from standard dataset UCF50, which is the largest dataset publicly available for action recognition. A multi-class SVM classifier with binary tree architecture was proposed in [88] for home-care surveillance systems. In this method, multiple SVMs were combined for recognition of actions but each SVM was trained separately for a better performance. The system was tested with a dataset of six activities, including, jogging, stand-to-sit, walk, stand-to-squat, fall, and in place action like standing, sitting and squat. Another method for action recognition was proposed in [89] under view changes.

4.5 Kalman Filter

In [90], Kalman filter (KF) was used for tracking and classification of pedestrians for videos captured by the camera positioned at a particular point. The tracking of pedestrians was accomplished by considering the position and velocity of each of them. KF was also used in [91] for human tracking. However, KF-based methods need effective foreground segmentation, and these methods are less effective for handling occlusion.

4.6 Artificial Neural Network

Artificial Neural Network (ANN) is one of the popular types of classifiers used for classification. In [92] a method was proposed for fall detection. In this method, activities were classified using motion capture system and back propagation neural networks learning method. A four layer network was proposed in [93] for activity monitoring in a healthcare environment where the data was received from accelerometer sensors. ANN has several advantages over statistical models. However, it has several limitations as well, such as when the model is complex then the problem of over-fitting is likely to occur which affects the prediction ability of the classifier.

4.7 Multidimensional Indexing

Multidimensional indexing has been used for human activity classification. In [94], the activity was represented by major human poses and velocity of major body parts such as hands, legs, and torso. This information was stored as a set of multidimensional hash tables. A separate hash table was used to store the information of each body part. Activity recognition was achieved by indexing and sequencing of few pose vectors in the hash tables. A sequence-based voting approach was employed to make it invariant to the speed of the activity. It is claimed that the proposed approach is robust in varying view angles in the range of ±30° and partial occlusion. The major advantage of multi indexing is that it is computationally less expensive.

4.8 Deep Learning

Conventional machine learning algorithms have limited ability to process the data in their raw forms. Therefore, a carefully engineered feature extractors are required that could transform the raw data into a feature vector suitable for classification. Deep learning uses computational models with multiple processing layers based on representation learning with multiple levels of abstraction. Representation learning encompasses a set of methods that enable the machine to process the data in raw form and automatically transform it into a suitable representation needed for classification. This transformation process is handled at different layers, for example, an image consists of an array of pixels then first layer represents the edges at particular location and orientation. The second layer represents the motifs by recognising the particular arrangement of edges in an image. The third layer may combine the motifs into larger objects and following layers would detect and recognize the objects as combination of these parts. These layers are learned from the data using a general purpose learning procedure and do not need to be designed manually by human engineers [95].

A video classification method was proposed in [96] using CNN for a dataset of 1 million YouTube videos comprised of 487 classes. The results indicate a significant improvement in performance compared to feature-based baselines. Another method based on fuzzy CNN was proposed in [97] for HAR. In this method CNN was used for action recognition from local patterns. In [98] CNN was used for pose estimation. Deep Neural Networks model has shown better performance than the classic ANN in the domains of speech recognition, image recognition and HAR. Once the network is trained a DNN model requires short test time. This model can act on the raw input directly for feature construction and classification.

5 Conclusion and Future Research Directions

Human activity recognition has drawn much attention of the research community around the Globe during the last decade. This is due to its promising applications and research challenges, which are not yet addressed adequately. Human activity recognition process consists of three phases, which are surveyed in this article including (a) Object segmentation, (b) feature extraction and representation, and (c) activity classification. This paper presents state-of-the-art methods for these three phases of human activity recognition and analytical analysis, advantages, and limitation of each of these methods.

Object segmentation methods are further divided into two categories i.e., (a) background construction based methods, which are used in case of static camera and (b) foreground extraction based methods, which are used in the case of a moving camera. Methods that can handle camera motion are few and the view invariance is one of the major challenges for real world applications. Most of the methods for handling view invariance have not been able to address this issue adequately. Moreover, segmentation of cluttered and an occluded background is also a major challenge that has yet to be addressed. However, fuzzy logic based models have shown good performance in handling complex backgrounds.

In the feature extraction and representation phase, different kinds of features such as global, local, and semantic-based have been discussed. The proper feature extraction and representation plays a major role in the overall results of the activity recognition. As reported in the literature, research challenges such as anthropometric variations and execution rate are resolved to a certain extent by combining different features. However, still there is a space for improvement and more robust methods are desirable for handling these issues.

In the activity classification phase, methods based on different classifiers have been surveyed and compared. The performance of action recognition methods is quite encouraging; however, there is still room for improvement. Specifically, the introduction of more robust classifiers is desirable that can work with less training samples and can handle inter and intra-class variations adequately. Moreover, the performance of action recognition methods has to be improved in the context of real-time applications. One way to boost the performance of the activity recognition methods is the use of GPU as reported in the literature, which is almost 50 times faster than its CPU counterpart and use of deep neural networks can also be useful for a better performance.

Deep Neural networks model has become very popular due to promising results in different domains including human activity recognition. Recently, few articles have been published for HAR using CNN [96, 97, 99]. There are some other architectures of DNN such as deep belief networks and deep auto-encoders. It would be interesting to use these architectures for HAR in future research which have great potential to overcome many HAR related challenges. Specifically, unsupervised training model of these architectures should be explored for promising results.

References

1. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: an overview. Comput. Sci. Rev. **11**, 31–66 (2014)
2. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. Circuits Syst. Video Technol. IEEE Trans. **18**, 1473–1488 (2008)

3. Poppe, R.: A survey on vision-based human action recognition. Image Vision Comput. **28**, 976–990 (2010)
4. Ke, S.-R., Uyen, H.L., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., Choi, K.-H.: A review on video-based human activity recognition. Computers. **2**, 88–131 (2013)
5. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. ACM Comput. Surv. (CSUR) **43**(3), 16 (2011)
6. Ramanathan, M., Yau, W.-Y., Teoh, E.K.: Human action recognition with video data: research and evaluation challenges. Human-Mach. Syst. IEEE Trans. **44**(5), 650–663 (2014)
7. Aggarwal, J., Xia, L.: Human activity recognition from 3d data: a review. Pattern Recogn. Lett. **48**, 70–80 (2014)
8. Ziaeefard, M., Bergevin, R.: Semantic human activity recognition: a literature review. Pattern Recogn. **48**(8), 2329–2345 (2015)
9. Morris, G., Angelov, P.: Real-time novelty detection in video using background subtraction techniques: State of the art a practical review. In: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE (2014)
10. Sobral, A., Vacavant, A.: A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Comput. Vision Image Underst. **122**, 4–21 (2014)
11. Sobral, A.: BGSLibrary: An opencv c ++ background subtraction library. In: IX Workshop de Visao Computational (WVC'2013), Rio de Janeiro, Brazil (2013)
12. El Baf, F., Bouwmans, T., Vachon, B.: Foreground detection using the Choquet integral. In: WIAMIS'08. Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 2008. IEEE (2008)
13. Toyama, K., et al.: Wallflower: principles and practice of background maintenance. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999. IEEE (1999)
14. Heikkila, M., Pietikainen, M.: A texture-based method for modeling the background and detecting moving objects. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 657–662 (2006)
15. Yao, J., Odobez, J.-M.: Multi-layer background subtraction based on color and texture. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE (2007)
16. Jian, X., et al.: Background subtraction based on a combination of texture, color and intensity. In: 9th International Conference on Signal Processing, 2008. ICSP 2008. IEEE (2008)
17. Jain, V., Kimia, B.B., Mundy, J.L.: Background modeling based on subpixel edges. In: IEEE International Conference on Image Processing, 2007. ICIP 2007. IEEE (2007)
18. Lai, A.H., Yung, N.H.: A fast and accurate scoreboard algorithm for estimating stationary backgrounds in an image sequence. In: Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, 1998. ISCAS'98. IEEE (1998)
19. Wren, C.R., et al.: Pfinder: Real-time tracking of the human body. Pattern Anal. Mach. Intell. IEEE Trans. **19**(7), 780–785 (1997)
20. Friedman, N., Russell, S.: Image segmentation in video sequences: a probabilistic approach. In: Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc (1997)
21. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE (1999)
22. Hayman, E., Eklundh, J.-O.: Statistical background subtraction for a mobile observer. In: Proceedings of Ninth IEEE International Conference on Computer Vision, 2003. IEEE (2003)
23. Zivkovic, Z. *Improved adaptive Gaussian mixture model for background subtraction*. in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. 2004. IEEE
24. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27**(7), 773–780 (2006)

25. Tuzel, O., Porikli, F., Meer, P.: A bayesian approach to background modeling. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005, CVPR Workshops. IEEE (2005)
26. Chen, Y.-T., et al.: Efficient hierarchical method for background subtraction. Pattern Recogn. **40**(10), 2706–2715 (2007)
27. Zhang, H., Xu, D.: Fusing color and texture features for background model. In: Third International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2006, Xi'an, China, 24–28 Sept 2006. Springer (2006)
28. El Baf, F., Bouwmans, T., Vachon, B.: Fuzzy statistical modeling of dynamic backgrounds for moving object detection in infrared videos. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE (2009)
29. Azab, M.M., Shedeed, H.A., Hussein, A.S.: A new technique for background modeling and subtraction for motion detection in real-time videos. In: ICIP (2010)
30. Sivabalakrishnan, M., Manjula, D.: Adaptive background subtraction in dynamic environments using fuzzy logic. Int. J.Video Image Process. Netw. Secur. **10**(1) (2010)
31. Bouwmans, T.: Background subtraction for visual surveillance: a fuzzy approach. In: Handbook on Soft Computing for Video Surveillance, pp. 103–134 (2012)
32. Shakeri, M., et al.: A novel fuzzy background subtraction method based on cellular automata for urban traffic applications. In: 9th International Conference on Signal Processing, ICSP 2008. IEEE (2008)
33. Maddalena, L., Petrosino, A.: A self-organizing approach to background subtraction for visual surveillance applications. Image Process. IEEE Trans. **17**(7), 1168–1177 (2008)
34. Culibrk, D., et al.: Neural network approach to background modeling for video object segmentation. Neural Netw. IEEE Trans. **18**(6), 1614–1627 (2007)
35. Maddalena, L., Petrosino, A.: A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection. Neural Comput. Appl. **19** (2), 179–186 (2010)
36. Oliver, N.M., Rosario, B., Pentland, A.P.: A Bayesian computer vision system for modeling human interactions. Pattern Anal. Mach. Intell. IEEE Trans. **22**(8), 831–843 (2000)
37. Goyat, Y., et al.: Vehicle trajectories evaluation by static video sensors. In: Intelligent Transportation Systems Conference, ITSC'06. IEEE (2006)
38. Godbehere, A.B., Matsukawa, A., Goldberg, K.: Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In: 2012 American Control Conference (ACC). IEEE (2012)
39. Permuter, H., Francos, J., Jermyn, I.: A study of Gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recogn. **39**(4), 695–706 (2006)
40. Yoon, S., et al.: Image classification using GMM with context information and with a solution of singular covariance problem. In: Proceedings of Data Compression Conference, DCC 2003. IEEE (2003)
41. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. In: IEEE 12th International Conference on Computer Vision, 2009. IEEE (2009)
42. Yu, T., et al.: Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models. In: IEEE Workshop on Motion and Video Computing, 2007. WMVC'07. IEEE (2007)
43. Gowsikhaa, D., Abirami, S., Baskaran, R.: Automated human behavior analysis from surveillance videos: a survey. Artif. Intell. Rev. **42**(4), 747–765 (2014)
44. Hu, W.-C., et al.: Moving object detection and tracking from video captured by moving camera. J. Visual Commun. Image Represent. (2015)
45. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. Pattern Anal. Mach. Intell. IEEE Trans. **36**(6), 1187–1200 (2014)
46. Mak, C.-M., Cham, W.-K.: Fast video object segmentation using Markov random field. In: 2008 IEEE 10th Workshop on Multimedia Signal Processing. IEEE (2008)

47. Cucchiara, R., Prati, A., Vezzani, R.: Real-time motion segmentation from moving cameras. Real-Time Imaging **10**(3), 127–143 (2004)
48. Jodoin, P., Mignotte, M., Rosenberger, C.: Segmentation framework based on label field fusion. Image Process. IEEE Trans. **16**(10), 2535–2550 (2007)
49. Wang, Y.: Joint random field model for all-weather moving vehicle detection. Image Process. IEEE Trans. **19**(9), 2491–2501 (2010)
50. Ghosh, A., Subudhi, B.N., Ghosh, S.: Object detection from videos captured by moving camera by fuzzy edge incorporated Markov random field and local histogram matching. Circuits Syst. Video Technol. IEEE Trans. **22**(8), 1127–1135 (2012)
51. Murali, S., Girisha, R.: Segmentation of motion objects from surveillance video sequences using temporal differencing combined with multiple correlation. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009. AVSS'09. IEEE (2009)
52. Wan, Y., Wang, X., Hu, H.: Automatic moving object segmentation for freely moving cameras. Math. Probl. Eng. **2014** (2014)
53. Kumari, S., Mitra, S.K.: Human action recognition using DFT. In: 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). IEEE (2011)
54. He, Z., Jin, L.: Activity recognition from acceleration data based on discrete consine transform and svm. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2009. IEEE (2009)
55. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
56. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia. ACM (2007)
57. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European Conference on Computer Vision. Springer (2006)
58. Noguchi, A., Yanai, K.: A surf-based spatio-temporal feature for feature-fusion-based action recognition. In: European Conference on Computer Vision. Springer (2010)
59. Wang, H., et al.: A robust and efficient video representation for action recognition. Int. J. Comput. Vision 1–20 (2-15)
60. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE (2005)
61. Lu, W.-L., Little, J.J.: Simultaneous tracking and action recognition using the pca-hog descriptor. In: The 3rd Canadian Conference on Computer and Robot Vision, 2006. IEEE (2006)
62. Lin, C.-H., Hsu, F.-S., Lin, W.-Y.: Recognizing human actions using NWFE-based histogram vectors. EURASIP J. Adv. Signal Process. **2010**, 9 (2010)
63. Hsu, F.-S., Lin, C.-H., Lin, W.-Y:. Recognizing human actions using curvature estimation and NWFE-based histogram vectors. In: Visual Communications and Image Processing (VCIP). IEEE (2011)
64. Kuo, B.-C., Landgrebe, D.A.: Nonparametric weighted feature extraction for classification. Geosci. Remote Sensing, IEEE Trans. **42**(5), 1096–1105 (2004)
65. Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R.: Matching shape sequences in video with applications in human movement analysis. Pattern Anal. Mach. Intell. IEEE Trans. **27**(12), 1896–1909 (2005)
66. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008. IEEE (2008)
67. Mahbub, U., Imtiaz, H., Ahad, A.: An optical flow-based action recognition algorithm. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)

68. Yang, M., Kpalma, K., Ronsin, J.: A survey of shape feature extraction techniques. Pattern Recogn. 43–90 (2008)
69. Rahman, S.A., Cho, S.-Y., Leung, M.K.: Recognising human actions by analysing negative spaces. IET Comput. Vision **6**(3), 197–213 (2012)
70. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2), 90–126 (2006)
71. Dargazany, A., Nicolescu, M.: Human body parts tracking using torso tracking: applications to activity recognition. In: 2012 Ninth International Conference on Information Technology: New Generations (ITNG). IEEE (2012)
72. Nakazawa, A., Kato, H., Inokuchi, S.: Human tracking using distributed vision systems. In: Proceedings of Fourteenth International Conference on Pattern Recognition, 1998. IEEE (1998)
73. Leung, M.K., Yang, Y.-H.: First sight: A human body outline labeling system. Pattern Anal. Mach. Intell. IEEE Trans. **17**(4), 359–377 (1995)
74. Leong, I.-F., Fang, J.-J., Tsai, M.-J.: Automatic body feature extraction from a marker-less scanned human body. Comput. Aided Des. **39**(7), 568–582 (2007)
75. Rogez, G., Guerrero, J.J., Orrite, C.: View-invariant human feature extraction for video-surveillance applications. In: IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007. IEEE (2007)
76. Yao, A., et al.: Does human action recognition benefit from pose estimation? In: BMVC (2011)
77. Sedai, S., Bennamoun, M., Huynh, D.: Context-based appearance descriptor for 3D human pose estimation from monocular images. In: Digital Image Computing: Techniques and Applications, DICTA'09. IEEE (2009)
78. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking people by learning their appearance. Pattern Anal. Mach. Intell. IEEE Trans. **29**(1), 65–81 (2007)
79. Kaghyan, S., Sarukhanyan, H.: Activity recognition using K-nearest neighbor algorithm on smartphone with tri-axial accelerometer. In: International Journal of Informatics Models and Analysis (IJIMA), vol. 1, pp. 146–156. ITHEA International Scientific Society, Bulgaria (2012)
80. Gavrila, D., Davis, L.: Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In: International workshop on automatic face-and gesture-recognition. Citeseer (1995)
81. Veeraraghavan, A., Chellappa, R., Roy-Chowdhury, A.K.: The function space of an activity. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE (2006)
82. Sempena, S., Maulidevi, N.U., Aryan, P.R.: Human action recognition using dynamic time warping. In: 2011 International Conference on Electrical Engineering and Informatics (ICEEI). IEEE (2011)
83. Robertson, N., Reid, I.: A general method for human activity recognition in video. Comput. Vis. Image Underst. **104**(2), 232–248 (2006)
84. Chung, P.-C., Liu, C.-D.: A daily behavior enabled hidden Markov model for human behavior understanding. Pattern Recogn. **41**(5), 1572–1580 (2008)
85. Thuc, H.L.U., et al.: Quasi-periodic action recognition from monocular videos via 3D human models and cyclic HMMs. In:), 2012 International Conference on Advanced Technologies for Communications (ATC). IEEE (2012)
86. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
87. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
88. Qian, H., et al.: Recognition of human activities using SVM multi-class classifier. Pattern Recogn. Lett. **31**(2), 100–111 (2010)
89. Junejo, I.N., et al.: View-independent action recognition from temporal self-similarities. Pattern Anal. Mach. Intell. IEEE Trans. **33**(1), 172–185 (2011)

90. Bodor, R., Jackson, B., Papanikolopoulos, N.: Vision-based human tracking and activity recognition. In: Proceedings of the 11th Mediterranean Conference on Control and Automation. Citeseer (2003)
91. Chu, C.-T., et al.: Human tracking by adaptive Kalman filtering and multiple kernels tracking with projected gradients. In: 2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC). IEEE (2011)
92. Sengto, A., Leauhatong, T.: Human falling detection algorithm using back propagation neural network. In: Biomedical Engineering International Conference (BMEiCON), 2012. IEEE (2012)
93. Sharma, A., Lee, Y.-D., Chung, W.-Y.: High accuracy human activity monitoring using neural network. In: Third International Conference on Convergence and Hybrid Information Technology, ICCIT'08. IEEE (2008)
94. Ben-Arie, J., et al.: Human activity recognition using multidimensional indexing. Pattern Anal. Mach. Intell. IEEE Trans. **24**(8), 1091–1104 (2002)
95. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
96. Karpathy, A., et al.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2014)
97. Ijjina, E.P., Mohan, C.K.: Human action recognition based on motion capture information using fuzzy convolution neural networks. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR). IEEE (2015)
98. Toshev, A., Szegedy, C.: Deep pose: human pose estimation via deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2014)
99. Ji, S., et al.: 3D convolutional neural networks for human action recognition. Pattern Anal. Mach. Intell. IEEE Trans. **35**(1), 221–231 (2013)
100. Gorelick, L., et al.: Actions as space-time shapes. Pattern Anal. Mach. Intell. IEEE Trans. **29** (12), 2247–2253 (2007)
101. Ke, Y., Sukthankar, R., Hebert, M.: Spatio-temporal shape and flow correlation for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR'07. IEEE (2007)
102. Dollár, P., et al.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. IEEE (2005)
103. Lu, X., Liu, Q., Oe, S.: Recognizing non-rigid human actions using joints tracking in space-time. In: Proceedings of International Conference on Information Technology: Coding and Computing, ITCC 2004. IEEE (2004)
104. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. IEEE (2005)
105. Danafar, S., Gheissari, N.: Action recognition for surveillance applications using optic flow and SVM. In: Computer Vision–ACCV 2007, pp. 457–466. Springer (2007)
106. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004. IEEE (2004)
107. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. IEEE (2009)
108. Sorenson, H.W.: Kalman Filtering: Theory and Application. IEEE (1960)
109. Deng, L.: Three classes of deep learning architectures and their applications: a tutorial survey. APSIPA Trans. Signal Inf. Process. (2012)