Forensic Science International: Digital Investigation 32 (2020) 200901

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Vision-based human action recognition: An overview and real world challenges

Imen Jegham [a], Anouar Ben Khalifa [b, *], Ihsen Alouani [c], Mohamed Ali Mahjoub [b]

[a] Université de Sousse, Institut Supérieur d'Informatique et des Techniques de Communication de H. Sousse, LATIS- Laboratory of Advanced Technology and Intelligent Systems, 4011, Sousse, Tunisia
[b] Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS- Laboratory of Advanced Technology and Intelligent Systems, 4023, Sousse, Tunisia
[c] IEMN-DOAE, Université polytechnique Hauts-de-France, Valenciennes, France

ARTICLE INFO

Article history:
Received 31 July 2019
Received in revised form
14 October 2019
Accepted 20 December 2019
Available online 21 January 2020

Keywords:
Vision-based
Action recognition
Activity recognition
Real world challenges
Datasets

ABSTRACT

Within a large range of applications in computer vision, Human Action Recognition has become one of the most attractive research fields. Ambiguities in recognizing actions does not only come from the difficulty to define the motion of body parts, but also from many other challenges related to real world problems such as camera motion, dynamic background, and bad weather conditions. There has been little research work in the real world conditions of human action recognition systems, which encourages us to seriously search in this application domain. Although a plethora of robust approaches have been introduced in the literature, they are still insufficient to fully cover the challenges. To quantitatively and qualitatively compare the performance of these methods, public datasets that present various actions under several conditions and constraints are recorded. In this paper, we investigate an overview of the existing methods according to the kind of issue they address. Moreover, we present a comparison of the existing datasets introduced for the human action recognition field.

1. Introduction

The motions of the different human body parts are most often part of functional movements without showing intentions and thoughts. Human activities are grouped into four categories depending on body parts engaged in the action and its complexity (Aggarwal and Ryoo, 2011):

- Gesture: It is a visible bodily action representing a specific message. It is not a matter of verbal or vocal communication but rather a movement made with the hands, face or other parts of the body such as Okay gestures and thumbs up.
- Action: It is a set of physical movements conducted by only one person like walking and running.
- Interactions: It is a set of actions executed by at most two actors. At least one subject is a person and the other one can be a human or an object (hand shaking, chatting, etc).
- Group activities: It is a mixture of gestures, actions, or interactions. The number of performers is at least two plus one or more interactive objects (playing volleyball, obstacle racing, etc).

Human Action Recognition (HAR) aims to automatically examine and recognize the nature of an action from unknown video sequences. Due to the growing demand for automatic interpretation of human behavior, HAR has caught the attention in both academia and industry. In fact, analyzing and understanding a person's behavior is fundamentally required for a wide range of applications such as video indexing, biometrics, surveillance and security.

Circuit miniaturization in new submicron technologies has allowed embedded applications to support a variety of sensors that inform about human actions (Lara and Labrador, 2013; Pfeifer and Voelker, 2015; Chen et al., 2017). The mentioned sensors differ in terms of price, ease of installation and output data type. Sensors such as accelerometers, GPS, camera and Leap Motion can be used to identify the activity of a person (Ameur et al., 2016; Mimouna et al., 2018). Depending on the type of collected information, each sensor is used according to the application requirements (Valentin,

* Corresponding author.
E-mail addresses: imen.jegham@isitc.u-sousse.tn (I. Jegham), anouar.benkhalifa@eniso.rnu.tn (A. Ben Khalifa), ihsen.alouani@uphf.fr (I. Alouani), mohamedali.mahjoub@eniso.rnu.tn (M.A. Mahjoub).

https://doi.org/10.1016/j.fsidi.2019.200901
1742-2876/

2010; Debes et al., 2016). The use of vision sensors, for instance, is particularly interesting because of the big amount of information a camera can provide (Berrached, 2014). In this paper, we focus principally on vision-based action recognition systems that use a camera as the first sensor.

A variety of issues are present in the HAR field, thereby making it a challenging topic: anthropometric variation (Rahmani et al., 2018; Baradel et al., 1703), multiview variation (Liu et al., 2017a; Wang et al., 2014), cluttered and dynamic background (Afsar et al., 2015; Duckworth et al., 2016), inter-class similarity and intra-class variability (Zhao and Ji, 2018; Lei et al., 2018), occlusion (Piyathilaka and Kodagoda, 2015; Xiao and Song, 2018), illumination variation, shadow and scale variation (Kumar and Bhavani, 2016; Vasconez and Cheein, 2018), camera motion (Jegham and Ben Khalifa, 2017; Yang et al., 2013a), low quality videos (Rahman and See, 2016), insufficient data (Zhang et al., 2017a; Ishan Misra et al., 2016) and poor weather conditions (Afsar et al., 2015; Chebli and Ben Khalifa, 2018). The purpose of HAR systems is to analyze a scene in the real world and to recognize human actions effectively. Fig. 1 shows the general layout of any HAR system. A feature extraction mechanism computes numeric or symbolic information from images or video frames. Then, labels are accordingly affected to these extracted features by a classifier. The process consists of many procedures that ensure the efficient description of actions.

Recently, multiple public datasets dedicated to human activity and action recognition have been published. These datasets were captured by different types of sensors, such as Kinect, accelerometers, Motion Capture (MoCap), and infrared and thermal cameras. However, RGB (Red, Green, Blue) data have attracted huge attention for a large range of applications (Zhang et al., 2017). The order of appearance of the various human action datasets operates parallel to the HAR problems the scientific community has to face. Currently, available datasets are very close to reality and describe multiple action variations, which makes it interesting to assess the performance of frame representation and classification methods.

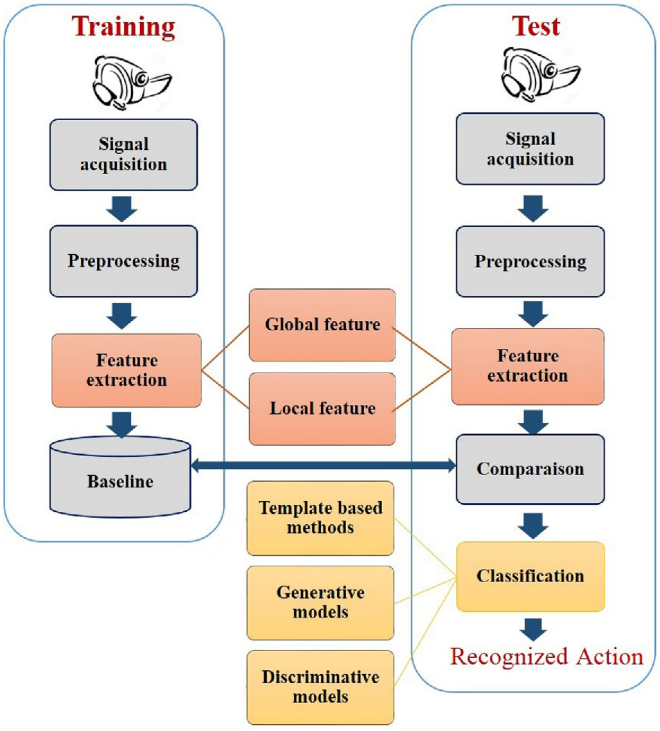

Fig. 1. General layout of HAR system.

We exhaustively went through surveys concerning HAR. None of them has presented an overview of the methods facing all real world challenges within public datasets that describe these issues. However, there are some detailed surveys that introduce only the public datasets present in the literature and describing human actions (Ahad et al., 2011; Chaquet et al., 2013; Zhang et al., 2016; Liu et al., 2017b, 2019) and others which introduce a detailed review on the different approaches proposed to recognize human actions. To detect physical actions, many sensors have been used. The proposed surveys can be then classified according to the type of sensors employed. For example, Aggrawal et al. (Aggarwal and Xia, 2014) summarized the major approaches in HAR from 3D data with particular attention to approaches that employed depth data. Sunny et al. (2015) presented an overview on applications and challenges of HAR using smart phone sensors. Nevertheless, vision sensors, in particular cameras, have attracted the attention of the majority of researchers due to the richness and usefulness of images and videos in HAR. For this reason, a plethora of surveys have reviewed the vision-based HAR methods (Moeslund et al., 2006; Turaga et al., 2008; Weinland et al., 2011; Aggarwal and Ryoo, 2011; Guo and Lai, 2014; Subetha and Chitrakala, 2016; Dhamsania and Ratanpara, 2016; Dhulekar et al., 2017; Zhang et al., 2017, 2019; Herath et al., 2017; Wang et al., 2018; Singh and Vishwakarma, 2019; Ji et al., 2019). A previous survey was introduced by Poppe (2010) in 2010, in which a detailed overview on vision-based HAR was provided. Ramanathan et al. (2014) provided a survey of the existing approaches based on their ability to handle the majority of HAR challenges and how these approaches could be generalized. However, these surveys are still insufficient to cover all the recently presented research and all the datasets in the literature given the increasing number of HAR related publications.

The goal of this survey is to investigate the suggested methods and the public datasets according to their ability to handle HAR challenges. The main contributions of this paper are as follows:

- We extensively study all the challenges facing HAR systems as well as the characterization methods proposed to handle these challenges.
- We review and divide the classification approaches based on their category.
- We introduce and categorize the existing datasets that describe real world issues for evaluating the performance of different suggested HAR methods.

The remaining of the paper is organized as follows: Section II introduces the challenges related to HAR and image representation. In Section III, action classification approaches and training strategies are discussed. Then, an overview on the main datasets used for testing HAR techniques are shown in Section IV. Finally, we conclude the paper in Section V.

2. Challenges and issues

In this section, we enumerate some of the difficulties in HAR and outline the different methods to deal with them. At least one of the challenges presented below can dramatically degrade the system performance. Ideally, extracted features have to generalize several variations including human appearances, points of view and backgrounds to overcome the challenges encountered in action recognition. Furthermore, these characteristics have to guarantee a relevant classification. Image representation is mainly categorized into two kinds: local and global.

Local representation schemes extract features from local specific patches. For action recognition, local representation emerges

following the work of Laptev (Laptev et al., 2008) on spatio-temporal interest points. This representation does not depend on people detection or human segmentation. Local representation is invariant to partial occlusion, background clutter, viewpoint variations, human appearances and, possibly, scales and rotations (Poppe, 2010).

Global representation extracts global features directly from original images or videos. The image is globally processed so that all pixels can be potentially used to compute the descriptor. A step of human detection and background subtraction is then required before descriptor computation. This step generates a Region Of Interest (ROI) generally obtained through background removal or human tracking. These ROIs include information about human appearances, motion, geometric structures, etc. In general, the residual noise remains due to imperfect extraction, which makes it very sensitive to the cluttered background. Glibal approaches focus on computing handcrafted features. They are usually sensitive to variation in viewpoints, noise and occlusion (Poppe, 2010; Zhang et al., 2016).

Table 1 introduces an overview of the studied real world challenges and related work.

2.1. Anthropometric variation

While performing an action, each person has their own comfort zone and body size proportion. In fact, the person can be in different postures and different angles of view. Depending on the age, the body flexibility or the position in an observed scene, humans exhibit size variabilities. As human movements are quite complex and present infinite variability, the slightest wink and movement can give meaning which is context-dependent (Rahmani et al., 2018; Baradel et al., 1703). At the end, a person can appear under various looks including the dress level according to the season or tastes. The person body is very flexible and is composed of multiple parts that can make different disjointed deformations. These deformations are restrained due to the human body skeletal structure.

The actors present a huge variation in poses and appearances, in a way that features extracted from shapes and positions are not efficient. For this reason, researchers have sought to extract motion

Table 1
Challenges for real world HAR and related reference work.

Challenge	Representative work	
Anthropometric variation	Pose and shape	Bobick et al. (Bobick and Davis, 1996) (1996) Weinland et al. (Weinland et al., 2006) (2006) Xu et al. (Xuet al., 2007) (2007) Kumari et al. (Kumari and Mitra, 2011) (2011)
	Motion	Dollar et al. (Dollar et al., 2005) (2005) Laptev et al. (Laptev et al., 2008) (2008) Messing et al. (Messing et al., 2009) (2009) Jiang et al. (Jiang et al., 2012) (2012) Bilinski et al. (Bilinski and Bremond, 2012) (2012) Peng et al. (Peng et al., 2013) (2013) Bilinski et al. (Bilinski et al., 2013) (2013) Wang et al. (Wang et al., 2013a) (2013) Zaidenberg et al. (Zaidenberg et al., 2014) (2014) Bilinski et al. (Bilinski and Bremond, 2015) (2015) Wang et al. (Wang et al., 2015) (2015) Rahmani et al. (Rahmani et al., 2018) (2017)
	Fusion of motion and appearance	Jiang et al. (Jiang et al., 2012) (2012) Cheron et al. (Chron et al., 2015) (2015) Baradel et al. (Baradel et al., 1703) (2017) Duta et al. (Duta et al., 2017) (2017)
Multiview variation		Yilmaz et al. (Yilmaz and Shah, 2005b) (2005) Yanet al. (Yan et al., 2008) (2008) Weinlandet al. (Weinland et al., 2010) (2010) Iosifidiset al. (Iosifidis et al., 2012) (2012) Luet al. (Lu et al., 2012) (2012) Liet al. (Li and Zickler, 2012) (2012) Zhanget al. (Zhang et al., 2013a) (2013) Wanget al. (Wang et al., 2014) (2014) Liuet al. (Liu et al., 2017a) (2017) Rahmaniet al. (Rahmani et al., 2018) (2017)
Cluttered and dynamic background		Schuldt et al. (Schuldt et al., 2004) (2004) Dollar et al. (Dollar et al., 2005) (2005) Laptev et al. (Laptev, 2005) (2005) Niebles et al. (Niebles et al., 2008) (2008) Wu et al. (Wu et al., 2011) (2011) Guha et al. (Guha and Ward, 2012) (2012) Ikizler et al. (Ikizler-Cinbis and Sclaroff, 2012) (2012) Shao et al. (Shao et al., 2012) (2012) Wu et al. (Wu et al., 2013) (2013) Ramirez et al. (Ramirez-Amaro et al., 2013) (2013) Chaaraoui et al. (Chaaraoui et al., 2013) (2013) Wu et al. (Wu and Shao, 2013) (2013) Wu et al. (Wu et al., 2013) (2013) Wang et al. (Wang and Schmid, 2013) (2013) Rahmani et al. (Rahmani et al., 1408) (2014) Afsar et al. (Afsar et al., 2015) (2015) Xu et al. (Xu et al., 2016) (2016) Duckworth et al. (Duckworth et al., 2016) (2016)
Intra-class variability and inter-class similarity		Park et al. (Park and Aggarwal, 2004) (2004) Wang et al. (Wang and Mori, 2009) (2009) Minhas et al. (Minhas et al., 2010) (2010) Junjo et al. (Junejo et al., 2011) (2011) Maji et al. (Maji et al., 2011) (2011) Zhou et al. (Zhou and De la Torre, 2012) (2012) Ang et al. (Ang et al., 2012) (2012) Rahman et al. (Rahman et al., 2012) (2012) Jiang et al. (Jiang et al., 2012) (2012) Bilinski et al. (Bilinski and Bremond, 2012) (2012) Ikizler et al. (Ikizler-Cinbis and Sclaroff, 2012) (2012) Subramanian et al. (Subramanian and Suresh, 2012) (2012) Ji et al. (Ji et al., 2013a) (2013) Lavee et al. (Lavee et al., 2013) (2013) Zhang et al. (Zhang et al., 2013c) (2013) Zhang et al. (Zhang et al., 2013b) (2013) Wu et al. (Wu and Shao, 2013) (2013) Raptis et al. (Raptis and Sigal, 2013) (2013) Song et al. (Song et al., 2013) (2013) Barnachon et al. (Barnachon et al., 2014) (2014) Liu et al. (Liu et al., 2015b) (2015) Cumin et al. (Cumin and Lefebvre, 2016) (2016) Chang et al. (Chang, 2016) (2016) Liu et al. (Liu et al., 2016b) (2016) Liu et al. (Liu et al., 2016a) (2016) He et al. (He et al., 2017) (2017) Nasiri et al. (Nasiri et al., 2017) (2017) Pan et al. (Pan et al., 2017) (2017) Zhao et al. (Zhao and Ji, 2018) (2018) Hong et al. (Hong et al., 2018) (2018) Lei et al. (Lei et al., 2018) (2018)
Low quality videos		Efros et al. (Efros et al., 2003) (2003) Chen et al. (Chen and Aggarwal, 2011) (2011) Reddy et al. (Reddy et al., 2012) (2012) Rahman et al. (Rahman and See, 2016) (2016)
Occlusion		Li et al. (Li et al., 2010) (2010) Piyathilaka et al. (Piyathilaka and Kodagoda, 2015) (2015) Afsar et al. (Afsar et al., 2015) (2015) Xiao et al. (Xiao and Song, 2018) (2018)
Illumination variation, shadow and scale variation		Shan et al. (Shan et al., 2003) (2003) Xie and Lam (Xie and Lam, 2005) (2005) Willems et al. (Willems et al., 2008) (2008) Zhang et al. (Zhang and Parker, 2011) (2014) Soumya et al. (Soumya and Thampi, 2015) (2015) Kumar et al. (Kumar and Bhavani, 2016) (2016) Vasconez et al. (Vasconez and Cheein, 2018) (2018)
Camera motion		Yang et al. (Yang et al., 2013b) (2013) Jegham et al. (Jegham and Ben Khalifa, 2017) (2017) Hadfield et al. (Hadfield et al., 2017) (2017) Chebli et al. (Chebli and Ben Khalifa, 2018) (2018)
Insufficient data		Laptev et al. (Laptev et al., 2008) (2008) Niebles et al. (Niebles et al., 2008) (2008) Ikizler et al. (Ikizler-Cinbis et al., 2009) (2009) Duchenne et al. (Duchenne et al., 2009) (2009) Li et al. (Li and Zickler, 2012) (2012) Wang et al. (Wang et al., 2013b) (2013) Sun et al. (Sun et al., 2014) (2014) Misra et al. (Ishan Misra et al., 2016) (2016) Zhang et al. (Zhang et al., 2017a) (2017)
Poor weather conditions		Grushin et al. (Grushin et al., 2013) (2013) Afsar et al. (Afsar et al., 2015) (2015) Jegham et al. (Jegham and Ben Khalifa, 2017) (2017) Chebli et al. (Chebli and Ben Khalifa, 2018) (2018)

related features such as optical flow and motion. Other researchers have thought of hybrid features where both motion and shape information are combined.

The simplest representation to solve the problem of anthropometric variation is to extract actor's appearance and shape features. There are several ways to capture the appearance information such as pose-based constraints, appearance models and silhouettes (Ramanathan et al., 2014). Bobick et al. (Bobick and Davis, 1996) proposed a basic technique to stack the silhouettes into two components: Motion Energy Image (MEI) and the Motion History Image (MHI). This representation could be expressive and robust against many variations in illumination and clothing, but it was not efficient for some other variations. In fact, it was strongly dependent on the model of the background, the position and the motion of the camera. Silhouettes and shape representations were based on global feature representation that required a clear background model and a stationary camera or a good camera motion compensated model. In view of this invariant property, the extraction of silhouettes was a hard task from a single view. Therefore, a lot of solutions have been suggested, such as Envelop Shape representation extraction using cameras placed orthogonally (Xuet al., 2007), motion history volumes resulted from silhouettes extraction from multiple cameras (Weinland et al., 2006), etc. Nevertheless, silhouette and shape representations are helpless to represent the internal motion, e.g. mouth or eye movements in the human body contour (Chakraborty et al., 2017). Other representations such as the discrete Fourier transform (Kumari and Mitra, 2011) can support these variations. However, they are very sensitive to some issues such as dynamic backgrounds and camera egomotion.

The fact that silhouette and pose features are not robust enough to get through many challenges has led to the emergence of motion-based features. These features tend to ignore human appearances and silhouettes and are very useful for recognizing human action (Moeslund et al., 2006). However, they require background removal to avoid the background motion effect (Jiang et al., 2012). To reduce the effects of background variation, researchers have used local feature representation. However, there is no guarantee that the represented motion contains the desired humans motion exclusively. To characterize motion, after partitioning space time volumes into cuboids, Laptev et al. (2008) computed Histogram of Gradients (HOG) (Dalal and Triggs, 2005) and Histogram of Optical Flow (HOF) (Wang et al., 2011) features for each cuboid. Dollar et al. (2005) used different local descriptors based on optical flow, brightness and gradient. They studied various methods of combined descriptors: simple concatenation of pixel values, a single global histogram and a grid of local histograms. They concluded that concatenated gradient information acheives the best performance. Entity trajectories have been also a relevant motion representation approach. In fact, they have received much attention since they have shown significantly good results. The used video representation was based on the trajectory shape and the descriptors calculated according to the volume around the trajectory points. Messing et al. (2009) extracted the trajectories using KLT Tracker (Lucas and Kanade, 1981) to follow Harris3D interest points (Sipiran and Bustos, 2011). To improve a dense trajectory approach, peng et al. (Peng et al., 2013) reduced the number of valid trajectories using a motion boundary strategy. These motion trajectories were then represented using a set of descriptors such as HOG (Dalal and Triggs, 2005), HOF (Wang et al., 2011) and motion boundary histogram (Dalal et al., 2006). Other dense trajectory work has been proposed in this context (Bilinski and Bremond, 2012, 2015; Wang et al., 2013a, 2015; Bilinski et al., 2013; Zaidenberg et al., 2014; Rahmani et al., 2018).

Many inconveniences and advantages are related to the use of either pose or motion features. Hence, hybrid features that fuse silhouette and motion features offer a good trade-off. The combination of shape and motion features has been used extensively in recent years (Jiang et al., 2012; Chron et al., 2015; Baradel et al., 1703; Duta et al., 2017). They improve robustness because they take advantage of different methods to boost performance (Chron et al., 2015).

Anthropometric variation is the common issue of HAR systems. Therefore, many researchers have tended to extract anthropometric features in order to handle this issue. Hybrid features show their robustness since they capture more data about an activity than a single feature. However, the majority of used datasets are recorded in controlled settings.

2.2. Multi view variation

The HAR problem becomes more and more complex because of the view invariance. Most introduced methods have addressed the HAR problem (Wang et al., 2009; Wang and Schmid, 2013; Rahmani et al., 2014a; Simonyan and Zisserman, 2014; Rahmani et al., 1408; Rahmani et al., 2014b; Rahmani et al., 2016a; Shahroudy et al., 2016a; Liu et al., 2017b; Dai et al., 2017; Zhao et al., 1712; Nazir et al., 2018) from a fixed viewpoint. While these methods are very successful using a single viewpoint, their performance drops significantly under many viewpoints. The view angle has a great effect on the recognition of activities. Certainly, from different perspectives, several appearances of the same action can be formed. Moreover, different actions can be considered the same (Liu et al., 2017a; Wang et al., 2014). Fig. 2 and Fig. 3 show the influence of viewpoint changes on HAR, which makes this task not trivial. Fig. 2 illustrates two views of an actor performing two different actions. Fig. 2. a describes the action “waving one hand from front view” while Fig. 2. b represents the action “waving two hands from right view”. On the other hand, Fig. 3 depicts an actor performing the same action “throwing” from various views. The motion of patterns, locations and appearances of human varies notably in each view.

The most popular solution to solve the viewpoint variation problem is to introduce multiple synchronized cameras. This solution can overcame effectively the self-occlusion issues, and train the classifier using collected data from different views. However, it is a laborious task for many applications (Aggarwal and Xia, 2014). In order to handle the viewpoint variation, each capture of a camera is discretized into spaced divisions. Then, there are two main solutions: The first consists of a single classifier that is trained using the fusion of features extracted from each view. In the second solution, every set of classifiers is trained by features extracted from a specific view point (Forsyth et al., 2006; Lejmi et al., 2017a).

Nevertheless, this approach is not practical because it extends the information extracted from a single view to a number of views, which generates a huge amount of information that increases computation time and complexity.

Several methods have been proposed for view-invariant action recognition. For instance, Yilmaz et al. (Yilmaz and Shah, 2005a) employed epipolar geometry to impose fundamental matrix constraints for point correspondences between actions. Lu et al. (2012) used MEI and MHI to represent the human motion. According to the camera rotation viewpoint, the view space is partitioned into multiple sub-view spaces. After that, Liu et al. (2017a) used multi-view space Hidden Markov Model (HMM) algorithm in each sub-view space. However, these methods could be interrupted by the background motion or the presence of another person or object in the scene. To overcome this issue, some methods used a precise silhouette extraction. These approaches tried to remove the background and only leave silhouettes (Singh et al., 2008; Gorelick et al.,

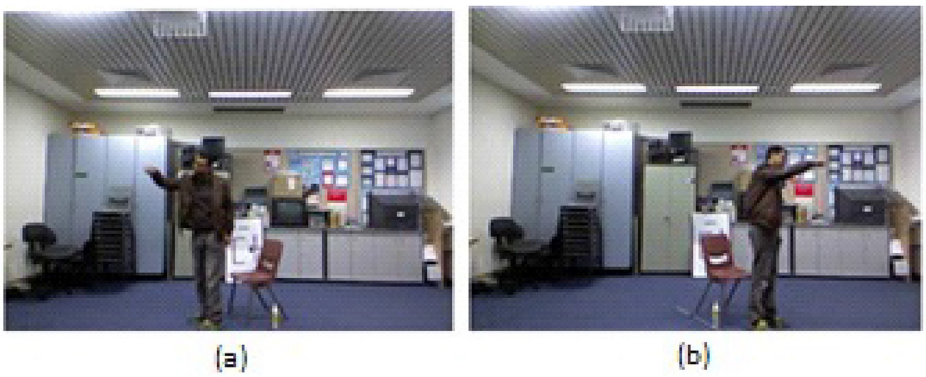

Fig. 2. A person performing two different actions from two various views: (a) The action is "waving one hand from front view". (b) The action is "waving two hands from right view". Images from UWA3D multiview dataset (Rahmani et al., 2016b).

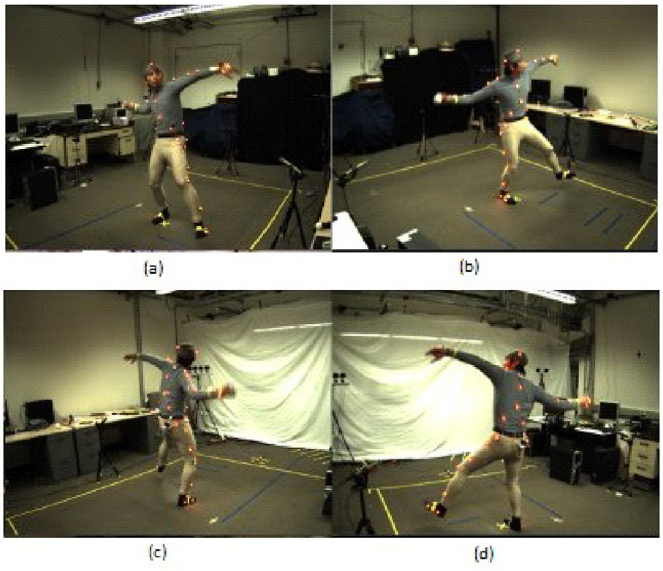

Fig. 3. A person performing the same action "throwing" from multiple views. Images from Berkley MHAD dataset (Ofli et al., 2013).

2007a).

Generally, datasets are acquired in controlled environments (indoors, no shadows or varying illumination, etc) and from synchronized multiple cameras. Nevertheless, during the implementation, many limitations are caused by quantizing the viewpoint space, which conducts to several view-dependent descriptions of a single body pose (Ji and Liu, 2010). This depends on a high number of labeled samples for each view.

A good HAR system has to recognize any action from unseen and unknown viewpoints. 3D or 2D body postures in an action are modeled depending on the representation used, such as silhouettes (Iosifidis et al., 2012) or visual hulls (Yan et al., 2008). Detecting and combining the techniques of separate body parts techniques are also proposed to create a body model (Weinland et al., 2010). However, the track of changes in silhouettes and their reflection on the created representations is one of the main problem in these techniques.

Furthermore, some methods have created models for cross-view human actions (Li and Zickler, 2012; Zhang et al., 2013a; Wang et al., 2014; Rahmani et al., 2018) to achieve view invariance. The methods based on knowledge transfer have become increasingly popular. They find an independent latent space in which features extracted from each view are immediately compared. Between source and target views, virtual views are constructed, so that the action descriptor is regularly transformed between the two viewpoints. However, during training, this method is quite dependent on samples from both source and target views.

From 2D images or video frames, some techniques have inferred 3D modeling of human body postures and used geometric transformations (Yan et al., 2008; Weinland et al., 2010; Iosifidis et al., 2012). These methods show more robustness than all the other approaches previously cited. All human activities are represented by human body features as a frame of reference. This allows the representation of actions in relative terms rather than in absolute ones. However, they are often dependent on the robust joint estimation, need a high number of parameters, and can fall in ambiguous poses because of the perspective projection.

2.3. Cluttered and dynamic background

The environments in which human activities are recorded are very important to reliably recognize actions. Most methods assume that HAR algorithms achieve high performances in an indoor environment where the background is uniform and static (Iosifidis et al., 2012; Rahman et al., 2015; Liu et al., 2015a). However, this performance degrades significantly in an outdoor environment. A cluttered or dynamic background is a form of interruption caused by the background noise. In real world situations, the extraction of global features will encode the background noise as ambiguous information, which leads to performance degradation. For example, in methods based on the optical flow the calculated motion is very sensitive to cluttered and moving backgrounds. They calculate unwanted background motion and combine it with human motion.

One of the main proposed solutions is to extract features that are robust against noisy backgrounds. The most popular used features are those based on key poses (Wu et al., 2011; Chaaraoui et al., 2013; Wu and Shao, 2013) due to their robustness towards the cluttered background. There are many other challenges like occlusion and anthropometric variations. Other features have proved their robustness towards this challenge, such as pbHOG features (Ikizler-Cinbis and Sclaroff, 2012) and pruning of motion features (Wang and Schmid, 2013). The use of local features like space time interest points (Schuldt et al., 2004; Dollar et al., 2005; Laptev, 2005; Niebles et al., 2008; Guha and Ward, 2012; Wu et al., 2013; Rahmani et al., 1408; Jegham et al., 2018) and edge maps (Jiang et al., 2006) in volumetric analysis can effectively deal with the background clutter issue. While these features achieve acceptable results, they remain insufficient to solve the problem globally.

Removing clutter can be acheived using a Kinect or infrared

camera for image acquisition. However, for an ordinary camera, the most common solution is to isolate the background from the foreground. For this reason, researchers have proposed color-based and region-based segmentation techniques that depend on a non-varying background for the tracking and segmentation of the foreground (Moeslund et al., 2006). Other methods such as segmentation and prefiltering have been suggested. They assume a non-varying distribution in the background, which makes it difficult to deal with complex backgrounds. For this reason, spatio-temporal features based on volumetric analysis are extracted (Shao et al., 2012; Ramirez-Amaro et al., 2013; Lejmi et al., 2017b). These features are immune to dynamic backgrounds, noise, camera-jitter, illumination and size variations (Ryoo and Aggarwal, 2009). Modeling the background is another solution that seems to be more robust than the other methods where a graphical model of the background is created. Using a static camera, this task is less difficult because of the static background. A simple subtraction of the static cluttered background results only in a moving object. Nevertheless, this condition is not always verified especially in real world situations. The Gaussian mixture models (Wu et al., 2011; Afsar et al., 2015) and the latent semantic analysis (Niebles et al., 2008; Ikizler-Cinbis and Sclaroff, 2012; Xu et al., 2016; Duckworth et al., 2016) are some examples of graphical models. These methods are dynamically changed as regards the background variations, achieve good background removing and are very robust to shadows. However, they fail to model complex backgrounds during a long period. The fusion of modeling background based methods and local features can be a good alternative to solve dynamic and noisy background issues.

2.4. Inter-class similarity and intra-class variability

It is very rare for the same person to repeat the same action with the same exact execution. Moreover, each individual behaves differently when performing the same action. For example, it is very rare for a young person and a senior to perform the action “running” in the same way. This issue comes from anthropometric variations between individuals, people habits, execution rate, etc. For complex, diverse and long actions, it is difficult to develop one simple model of the same action.

In order to minimize inter-class similarity and intra-class variability, several efforts have been made to obtain discriminative features. A robust and efficient HAR approach should cover variations and similitudes between classes. For a large number of activity classes, this task will be more difficult because of the high overlap between classes (Poppe, 2010). Research in the temporal domain, which needs a preprocessing step, is considered. However, it is not always realistic.

The most popular and common solutions to this issue is the probabilistic methods such as HMMs (Yang et al., 2013a; Ji et al., 2013a; Chakraborty et al., 2017; Zhao and Ji, 2018), fuzzy based systems (Subramanian and Suresh, 2012; Cumin and Lefebvre, 2016; He et al., 2017; Nasiri et al., 2017), dynamic Bayesian methods (Park and Aggarwal, 2004; Ikizler-Cinbis and Sclaroff, 2012; Zhang et al., 2013b, 2013c; Lavee et al., 2013), finite state machine (Baek and Yun, 2008; Trinh et al., 2011) and Conditional Random Fields (CRF) (Wang and Mori, 2009; Song et al., 2013; Liu et al., 2015b, 2016a, 2016b; Chang, 2016).

The mentioned methods use the temporal domain. The spatial domain is also considered by seeing the full length of videos. Many approaches have been proposed such as time warping techniques, where the template and the input data are converted into a common scale of time to ensure a comparison between frames (Junejo et al., 2011; Jiang et al., 2012; Ang et al., 2012; Rahman et al., 2012; Zhou and De la Torre, 2012; Barnachon et al., 2014) and Histogram based techniques (Minhas et al., 2010; Maji et al., 2011; Guha and Ward, 2012; Bilinski and Bremond, 2012; Wu and Shao, 2013; Raptis and Sigal, 2013; Wang and Schmid, 2013; Pan et al., 2017; Hong et al., 2018; Lei et al., 2018), which use the same fundamentals to find one description for the same action regardless how the action is performed. In this way, several methods were proposed such as codebooks, bag of features and dictionary based approaches. However, these methods are subject to quantization errors.

Usually, the inter-class similarity and the intra-class variability are present in datasets. A lot of similar actions have been shown and each action is executed by different subjects many times.

2.5. Low quality videos

HAR from poor quality videos including IP cameras and closed-circuit television cameras is still a challenging topic due to different complex problems such as slow frame rates, motion blurring, low resolution and compression artifacts in addition to classic activity recognition problems. While there are few approaches available in the literature that addressed the problem of video quality, they have little focus on poor video resolution. Other quality issues such as compression and blurring are almost unexplored.

Efros et al. (2003) came up with the idea of recognizing human actions from a distance. They introduced a motion-based descriptor able to encode features from human figures, which had an approximate height of 30 pixels. Chen and Aggarwal (2011) put forward speech -like mid-level modeling of human actions in order to recognize them on poor resolution videos. They achieved a relatively high performance across divers low quality datasets. Moreover, Reddy et al. (2012) proposed and investigated the performance of 3D spatiotemporal gradients under various quality conditions such as rotation, scale and sampling rates. They suggested to use a descriptor that has viewpoint manipulation capability to have a better performance. Furthermore, Rahman et al. (Rahman and See, 2016) introduced a spatio-temporal mid-level feature bank that integrated the advantages of local explicit patterns from interest points and global salient statistical patches.

2.6. Occlusion

Occlusion is defined as the temporary disappearance of human body parts by being behind another object or person of a greater apparent diameter. There are three main types of occlusion:

- Self-occlusion: From one point of view, some body parts are obscured by another part. For example, from the front view, the action “talking” cannot be recognized when a person puts their hand in front of their mouth (Fig. 4a).
- Crowd issue: When two or more people are hiding each other (Fig. 4b).
- Occlusion created by an object: This is when from one point of view, some body parts are occluded with an object. For example, recognizing the action “texting” or any other action performed with feet, from the front view, is a complex task when the subject is well installed in an office (Fig. 4c).

Human common sense reasoning that can recognize complex human actions based on few clues can be hardly modeled. A lot of solutions have been proposed to overcome this problem. Local features are used to represent and analyze human actions. The space time volume and the special volumetric analysis have proved their robustness to avoid occlusion issues since it is hard to find the body parts occluded in the whole video (Gorelick et al., 2007b; Poppe, 2010). Hence, features are extracted from the body parts

Fig. 4. Three main types of occlusion.

when they are not occluded by analyzing spatio-temporal volumes. Local features generate pose models, which can characterize the human activity in short periods and present a big weakness against real time applications.

Probabilistic methods and particularly HMMs (Poppe, 2010; Li et al., 2010; Afsar et al., 2015) and Bayesian networks (Piyathilaka and Kodagoda, 2015; Xiao and Song, 2018) have been used to establish appearance representations to model body parts. Each part of the human body is represented in the HMM model as a state that can be changed according to a probabilistic model. These approaches insure good results and inherit the advantages of local features because they consider each body parts as a separate entity (Weinland et al., 2011).

As the feature extraction from the occluded human body parts is rarely possible, it is important to find robust classifiers that can handle the occlusion problem. Until now, pose-based methods and probabilistic approaches offer the most admissible performance when the human performing the action is partially occluded.

2.7. Illumination variation, shadow and scale variation

The light sources can make a considerable difference in quality of the human action representation. In different lighting conditions, the same person as well as the human action may appear differently. This dramatically affect HAR system performances. Moreover, when the light source is blocked, a dark area appears. This area is known as a shadow. Like a silhouette, the shadow shape is a two-dimensional projection of the person who is blocking the light. As shown in Fig. 5, a human action will be partially duplicated at different scales according to the position of the light source relative to the person. Scale variance issues have been also exposed. Actually, this problem depends mainly on the distance between the subject and the camera. According to this distance, multiple scales and representations of the same subject may appear.

Three main solutions to this issues were proposed in (Chen, 2006): The construction of a generative shape model was one of these solutions. The model provided to recognize human actions was independent of lighting conditions and scale variances. Extracting features at multiple scales and in varying light conditions may be a solution but it requires a huge training dataset. For this reason, local invariant features are extracted. This approach tends to extract features invariant to illumination changes such as SURF points (Willems et al., 2008). Performing a preprocessing and normalization step can also be a solution. In this approach, human action frames are preprocessed using image processing methods to normalize the images. For illumination normalization, many widely used methods are proposed such as Gamma correction (Kumar and Bhavani, 2016), histogram equalization (Zhang and Parker, 2011) and logarithm transform (Vasconez and Cheein, 2018). However, this solution is still insufficient because of the non uniform

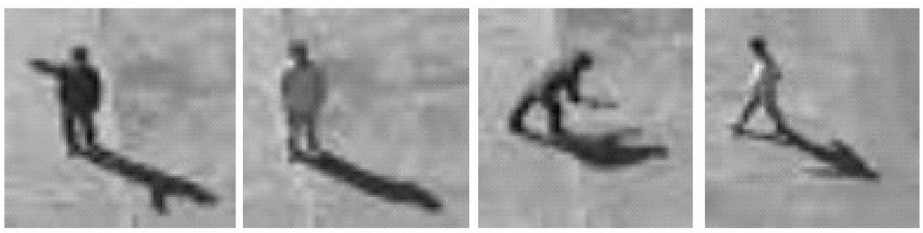

Fig. 5. Actions performed by humans and their shadows. Images from WEIZMANN dataset (Gorelick et al., 2007a).

illumination variation. Therefore, an adaptive preprocessing is needed using local preprocessing techniques like block-based histogram equalization (Xie and Lam, 2005), region-based histogram equalization (Shan et al., 2003) and adaptive histogram equalization (Soumya and Thampi, 2015).

Using local features to solve the issues of shadows and illumination and scale variations shows better results than utilizing global features. They present better robustness against shape, pose, illumination and scale changes (Jun et al., 2013).

2.8. Camera motion

In many HAR systems, researchers mostly use fixed cameras. However, some real world applications require dynamic cameras, which will certainly cause a drastic performance loss. In fact, camera motion significantly affects motion features since the action features are mixed with erroneous and misleading motion patterns.

The most used solution is to model and compensate camera motion (Yang et al., 2013a; Jegham and Ben Khalifa, 2017; Chebli and Ben Khalifa, 2018). In our previous work, we focused on this topic in (Jegham and Ben Khalifa, 2017). We assume that the differences between frames defined all motions caused by camera egomotion and moving objects. Using SURF keypoints, the motion of the background was measured to be compensated later. After that, the foreground was detected by a simple subtraction between frames. Yang et al. (2013a) compensated also camera motion before the feature extraction process by undergoing an homography from SIFT correspondences, called the feature based image alignment. Chebli et al. (Chebli and Ben Khalifa, 2018) compensated camera motion, presented as global motion, which was measured after dividing each frame into blocks. Then, a search of similar blocks in many frames was performed. A motion vector for each block was obtained, and using this vector a global translation was applied. In fact, camera egomotion compensation methods are simple to implement, but they can delete some motion features belonging to the foreground.

When using a non-stationary camera, methods based on epipolar geometry seem to be a good solution (Hadfield et al., 2017). The tracking matrix is divided into two matrices: one representing the shape itself and the other one representing camera poses and the foreground object. This approach is mainly used to solve multi view problems. For using multi view dynamic cameras, these techniques are extended by deriving a temporal matrix and also analyzing geometric transformations. However, this approach depends significantly on the cameras synchronization and calibration (Yilmaz and Shah, 2005b). Seen that, the camera system and its geometry are taken into account while modeling human activities, and the epipolar geometry based approach acheives acceptable performance. Nevertheless, the major drawback is the unavailability of public datasets that contain camera motion. Many existing datasets taken from movies or web videos are realistic and contain labeled sequences, but a number of training and test sequences is still limited. Some researchers just assume that zooming can be considered as a camera motion (Liu and Yuen, 2010).

2.9. Insufficient data

Computer vision has become a trending field of research due to the capacity to store and process huge amount of data. To design realistic techniques, datasets should contain good scale ranges, occlusion, intra and inter class variations, etc. However, most of human activity datasets contain a limited number of labeled videos. For this reason, the use of realistic datasets recorded in an unconstrained environment is introduced. These datasets usually consist of labeled videos collected from web videos or movies (Hollywood (Laptev et al., 2008), UCF Sports (Soomro and Zamir, 2014), etc). The major problem of these datasets is the limited number of labeled training and test sequences. The annotation of a large dataset is challenging and time consuming. Many other solutions have been proposed to handle the lack of appropriate datasets. For example, researchers have used web video search results (Ikizler-Cinbis et al., 2009), video subtitles (Wang et al., 2013b) and movie script matching (Laptev et al., 2008; Duchenne et al., 2009). However, in large datasets, this is practically infeasible.

2.10. Poor weather conditions

Bad weather conditions create a natural threat for action recognition systems. Darkness, rain, blowing snow and fog, for example, affect the visibility and represent a challenge to action recognition. In poor weather conditions, many parameters are changed: colors are extremely affected, distances can hardly be evaluated, etc.

A lot of researchers have used infrared imagery since they are not affected by poor lighting and robust to bad weather conditions (Jegham and Ben Khalifa, 2017; Chebli and Ben Khalifa, 2018). Afsar et al. (2015) proposed an action recognition technique based on HMM with tuned parameters to achieve high results. Grushin et al. (2013) showed the robustness of Long Short Term Memory (LSTM) techniques in recognizing and classifying human actions under bad weather conditions performed on the modified KTH dataset (Schuldt et al., 2004). This dataset was edited by injecting noise to simulate poor weather conditions.

3. Action classification

Classification of actions is the main stage in HAR systems. Action classification approaches can be grouped into three main models: template-based methods, discriminative models and generative models (Zhang et al., 2017). Hybrid models are also proposed by combining different categories (Ye et al., 2019). In this section, we focus on training and classification approaches as well as issues to solve for action classification.

3.1. Template based methods

Template based methods use a typical model that englobes the common characteristics of one action. For example, the template can be static images or a sequence of view models. Researchers extract static templates and compare them with unknown models extracted from test images or videos by computing the similarity between them. These methods are relatively simple and give acceptable results. Shechtman et al. (Shechtman and Irani, 2007) measured a behavior-based similarity by comparing a constructed 3D space-time video template from a short video sequence to every test video sequence. This method shows its robustness in detecting complex attitudes and activities, especially those performed in a controlled environment, even in the presence of simultaneous complex actions. However, these methods are unable to generate a unique template for each human action. Thus, the computational complexity of these methods is very high.

3.2. Generative model

Let us consider a sequence of observations X and a particular label Y. Generative classifiers learn the joint probability model P(X,Y). Then using the Bayes rules, they try to find a prediction by calculating P(Y|X). Label Y that has the highest probability is picked (Ng and Jordan, 2002). There are mainly two kind of generative models: HMM and Dynamic Bayesian network (DBN).

The recognition task is one of the HMM problems and can be solved by the forward algorithm. For each frame, features that indicate the number of pixels in each divided mesh are extracted as observations. After that, HMMs are trained using the observation feature vector sequences for each class including the initial probability of hidden states, the transition matrix and the confusion matrix. HMMs are widely used in HAR. They show high robustness against many kinds of issues such as intra-class variability, inter-class similarity, variations in lighting and viewpoints and occlusion (Natarajan and Nevatia, 2008). They are particularly effective at estimating missing data such as occluded body parts and cluttered backgrounds. However, HMMs also show some imperfections in particular cases. For example, a single variable state representation is not enough to model multiple interacting parts (Natarajan and Nevatia, 2008). For this reason, many HMM variants have been proposed in the literature, such as Coupled-HMM (Brand et al., 1997), hierarchical variable transition-HMM (Natarajan and Nevatia, 2008), etc.

DBNs are rarely used in HAR (Piyathilaka and Kodagoda, 2015; Anitha and Baghavathi Priya, 2019; Xiao and Song, 2018), but it is widely applied in human interaction recognition. In contrast to HMM, the hidden state of DBN is represented as a set of random variables (Murphy). Several DBN variants have been proposed in the literature. However, only hierarchical dynamic Bayesian network (HDBN) (Xiao and Song, 2018) is used to recognize actions.

3.3. Discriminative model

Given a sequence of observations X and a particular label Y, discriminative classifiers model the posterior probability P(Y|X) or learn a direct mapping that links inputs X to their correspondence class labels (Ng and Jordan, 2002).

CRFs are used to label human action sequences. They are an undirected graphical models that design the conditional probability P(Y—X) where Y represent a particular label sequence and X represent a sequence of observations. They allow the incorporation of complex features of the sequence without violating the independence assumptions of the model. These models use the entire observation sequence (Vail et al., 2007). Compared with the generative model, Vail et al. (2007) showed that CRFs achieve comparable or better results than HMMs. Many variants of CRF were proposed in the literature, such as Hidden CRF (Quattoni et al., 2007) and Max-margin hidden CRF (Wang and Mori, 2009).

Due to their high performance and simplicity, the Support Vector Machine (SVM) (Suykens and Vandewalle, 1999) and the Nearest Neighbor (NN) (Cover and Hart, 1967) are frequently used. These supervised learning methods focus on finding a hyperplane which maximizes the margin of many classes. They generally use the Euclidian distance between features which can be inefficient for cases that require high dimensional features such as HAR. For this reason, a lot of other distances are used in HAR such as the Chamfer distance (Oikonomopoulos et al., 2006), the Riemannian metric (Guo et al., 2013) and the Mahalanobis distance (Bobick and Davis, 2001). These methods assume that the data distribution in training and testing stages are the same. However, this is not the case in many cases, thereby requiring a large amount of training data to recognize all possible variations.

Dynamic Time Warping (DTW) is a distance calculated between two sequences that may have different lengths (Müller, 2007). In order to detect similar shapes with different phases, Semprena et al. (Sempena, 2011) applied DTW in HAR due to its efficiency in measuring time-series similarity, which minimized the effects of time distortion and shifting through the creation of a warping path. This method did not depend on a huge amount of training data. However, the computational complexity would increase when dealing with a high intra-class variation due to the need of extensive templates to store those invariances.

To train as much data as possible, several researchers have used the web as a source of information. Others have utilized an incremental procedure, such as transform learning model, which uses features from one dataset to learn and other features from a target dataset to classify actions. These approaches employ simple classifiers (e.g. adaboost, SVM) and can succesfully proceed with a limited amount of training data. However, they cannot differentiate between negative and positive training data (Zhang et al., 2017). To make the right decision, a classifier must learn the most discriminative features with weightage. Accordingly, a fuzzy rule based classification method can solve most of the HAR challenges. Several methods have been introduced in the literature. The meta-Cognitive learning algorithm for a neuro-fuzzy inference system is one of the fuzzy rule based classification algorithms which is generally employed in an incremental manner to decide automatically when, what exactly and how to learn the available knowledge in the classifier and the new training samples (Subramanian and Suresh, 2012).

When dealing with high volume datasets, deep learning is the most common used method. It is a branch of machine learning that intends to design high level abstractions in data using complex structured architectures. Deep learning architectures can be mainly grouped into three categories: Convolutional Neural Network (CNN or convnet), Recurrent Neural Network (RNN) and other emergent architectures. Convnet is used to automatically extract features that will be then classified in a multilayer perceptron such as resNet (He et al., 2016) and mobileNet (Howard et al., 1704). CNN is the most widely used deep learning technique due to its impressive results in different pattern recognition application domains. However, datasets used in HAR are generally insuffuicient for efficient CNN training. Common solutions have been proposed to generate or create more training instances. In HAR, many extensions of convnet have been introduced such as Pose based CNN (Chron et al., 2015), 3D CNN (Ji et al., 2013b) and Long-term temporal convolutions CNN (Varol et al., 2018).

RNNs are utilized to learn complex temporal dynamics. They have been explored successfully on many tasks especially on text generation and speech recognition. As human actions are made of complex sequences of motor movements that can be seen as temporal dynamics, RNNs represent a suitable solution. LSTM is the most popular specific RNN architecture that can store information for an extended period. It shows high performance in recognizing actions in real world scenarios (Grushin et al., 2013). Many extensions have appeared such as spatio-temporal LSTM (Liu et al., 2016c), global context-aware attention LSTM (Liu et al., 2017c), etc. To further improve HAR performances, researchers proposed to combine CNN and LSTM (Ordóñez and Roggen, 2016; Nez et al., 2018). This combination acheives better results rather than using a single architecture. While discriminative models achieve promising results, they are limited by their dependency on large datasets to satisfy HAR requirements.

4. Datasets

Testing HAR algorithms is necessary to analyze qualitative and quantitative performance. However, an effective analysis requires datasets that describe actions under various conditions. Unfortunately, such datasets that present all scenarios are not available. Consequently, many researchers tend to the creation of their own datasets (Afsar et al., 2015). Most of the work introduced in the literature uses public datasets. Even if these datasets do not describe the desired challenges, they try to insert some changes in them (Grushin et al., 2013). For example, they add some noise to

simulate poor weather conditions or neglect annotation files.

Introducing new public datasets runs parallel to the appearance of HAR real world issues that the scientific community try to overcome. This is important for two main reasons:

- Saving resources and time when researchers focus on implementing algorithms rather than wasting time, resources and energy on recording new videos.
- Ensuring an efficient comparison of different approaches and the overview given on the capabilities of the numerous available methods by using the same datasets.

Many public datasets are unrealistic and recorded in controlled environments such as fixed and uniform backgrounds and static cameras. Lately, realistic datasets that contain labeled sequences from web videos or movies introduced increasing complexity, which makes this research field a more challenging field. Since the appearance of the first public dataset, the most common faced challenge has been the anthropometric variations. The most commonly used public datasets for HAR are described in this section. They are classified into four main categories: constrained datasets, multiview datasets, realistic datasets and synthetic datasets.

4.1. Constrained dataset

KTH (Schuldt et al., 2004) and Weizmann (Gorelick et al., 2007a) datasets are two classic datasets that are used as benchmarks for HAR techniques. The important factor that makes them two important datasets is the high anthropometric and intraclass variations. The recognition of actions using these datasets is considered an easy task (100% accuracy (Natarajan and Nevatia, 2008; Tran and Sorokin, 2008)) as only one subject is acting in each frame and the background is relatively simple. To include further challenges, they have been modified to present some degraded conditions (Grushin et al., 2013). Table 2 summarizes the main public constrained datasets introduced in the literature.

4.1.1. KTH dataset

KTH (Schuldt et al., 2004) is one of the most popular and most used HAR datasets. It was introduced in 2004 and recorded six actions (walking, jogging, running, hand clapping, hand waving, and boxing) executed by 25 actors in four different controlled environments: indoors, outdoors, outdoors with different clothes and outdoors with a scale variation. This dataset was recorded using a relatively stationary camera, and the zoom was considered as the camera motion.

4.1.2. Weizmann datasets

The Weizmann institute of science provides two datasets:

- Weizmann Event-Based Analysis dataset (Zelnik-Manor and Irani, 2001): It was introduced in 2001 and was one of the first public datasets. This dataset is composed of a sequence of 6000 frames, showing various actors, wearing different clothes, and performing four divers actions: walking, running, running in place and waving.
- Weizmann Actions as Space-Time Shapes dataset (Gorelick et al., 2007a): It was introduced in 2005 and describes 10 actions: walking, running, skipping, bending, jumping jack, jumping forward on two legs, jumping in place on two legs, galloping sideways, waving one hand and waving two hands. Each action was performed by 10 subjects and recorded using a static camera. The dataset provides relatively simple sequences of the background making the background subtraction an obvious task to perform. The first goal of this dataset was to apply algorithms based on space-time shape volumes.

4.2. Multiview dataset

For the view invariant HAR issue, a lot of multiview datasets introduced various actions in different environments. However, they lack realistic human actions and are consequently not adapted to complex activity recognition issues. Table 3 summarizes the main multiview datasets introduced in the literature.

4.2.1. IXMAS dataset

The INRIA Xmas Motion Acquisition Sequences dataset (INRIA, 2008) was acquired in 2006 at the scientific research center INRIA, Grenoble, France. The multi view dataset was collected using five cameras and presented 11 actors performing 14 daily actions (nothing, crossing arms, checking watch, scratching head, getting up, turning around, throwing, sitting down, walking, waving, kicking, punching, pointing and picking up). Each action was recorded three times. This dataset was introduced to study how to create spatio-temporal action models that could support the recognition of simple actions independently of many factors such as the point of view, the actor body size and the gender.

4.2.2. UTD multiview action dataset

The University of Texas, Dallas collected the UTD multiview Action dataset (Utdallas, 2018) in 2015 using a Kinect depth camera v2 and an inertial sensor placed on the right wrist of subjects at the same time. The dataset includes six actions (catch, draw tick, draw circle, draw triangle, throw and knock). Five actors were asked to perform each action with five different views. For each view, a subject repeated an action six times to generate 900 action samples. Contrary to other multiview datasets, only one camera was used to capture each action with five different subject orientations, as depicted in Fig. 6.

4.2.3. NTU-RGB dataset

The Nanyang Technological University created NTU RGB + D Action Recognition dataset (Shahroudy et al., 2016b) in 2016. This large-scale dataset for HAR presents more than 56,000 video samples and 4,000,000 frames. It contains 60 different actions including daily, health-related and mutual activities executed by 40 different performers. This dataset was simultaneously recorded by three Microsoft Kinect v.2 cameras. It was acquired from a huge number of views (80), which explains its unicity. An extended version of this dataset is currently available (Liu et al., 2019). The extension contains 120 distinct actions executed by 106 various subjects.

4.3. Realistic dataset

Facing the real world issues, the disposed situations are more and more complex. To effectively handle these problems, introducing new datasets obtained from real world situations is required. The main purpose of a realistic dataset is to recognize human activities in realistic conditions and under diverse video settings. Action recognition in realistic environments is advanced by the demand of real applications on natural HAR solving the variations in camera views, scenes surroundings, postures, motion, illumination, clothing, occlusion, etc. A lot of realistic datasets that contain labeled sequences gathered from movies or web videos are exposed in Table 4.

4.3.1. Hollywood dataset

The IRISA institute created two datasets: Hollywood 1 and

Table 2
Main constrained datasets in the literature.

Category	Dataset	Input type	Subject	Classes	Repetition	Year	Degraded conditions									
							A[a]	MV[b]	CB[c]	IC[d]	LQ[e]	O[f]	ISS[g]	CM[h]	ID[i]	PW[j]
Constrained dataset	KTH (Schuldt et al., 2004)	RGB	25	6	–	2004	*			*	*		*			
	Weizmann Actions as Space-Time Shape (Gorelick et al., 2007a)	RGB	9	10	–	2005	*			*	*		*			
	HumanEva (Sigal et al., 2010)	RGB,MC[k],S[l]	4	6	–	2006	*			*						
	CAD-60 (Sung et al., 2011 Saxena)	RGB,D[m],S	4	12	–	2011	*		*	*		*				
	RGB-D HuDaActivity (Ni, Wang, Moulin p rgbd, 2011)	RGB,D	30	12	2–4	2011	*		*	*		*				
	MSRDaily Activity 3D (Wang et al., 2012)	RGB,D,S	10	16	–	2012	*		*			*				
	UTKinect (Xia et al., 2012)	RGB,D,S	10	10	2	2012	*	*	*			*				
	G3D (Bloom et al., 2012)	RGB,D,S	10	20	3	2012	*						*			
	DHA (Lin et al., 2012)	RGB,D,HM[n]	21	17	–	2012	*			*						
	CAD-120 (Koppula et al., 2013)	RGB,D,S	4	10	3	2013	*		*	*		*				
	Workout SU-10 (Negin et al., 2013)	RGB,D,S	12	10	–	2013	*			*						
	MSR Action Pair (Oreifej and Liu, 2013)	RGB,D,S	10	12	2	2013	*			*						
	IAS-lab (Munaro et al., 2014)	RGB,D,S	11	15	3	2013	*			*						
	Osaka (Mansur et al., 2013)	RGB,D,S	8	10	–	2013	*									
	Mivia (Carletti et al., 2013)	RGB,D	14	7	5	2013	*									
	RGBD-SAR (Yang et al., 2013b)	RGB,D	30	9	3	2013	*		*	*		*				
	TJU (Liu et al., 2015b)	RGB,D	20	15	4	2015	*			*						
	3D online action (Yu et al., 2014)	RGB,D	36	7	–	2015	*		*			*				
	SYSU (Hu et al., 2017)	RGB,D,S	40	12	–	2017	*			*						
	RGBD Activity (Wu et al., 2015)	RGB,D,S	7	21	–	2015	*			*						
	UTD MHAD (Chen et al., 2015)	RGB,D,S	8	27	–	2015	*			*						

[a] Anthropometric variation.
[b] Multiview variation.
[c] Cluttered or dynamic background.
[d] Intra-class variability and Inter-class similarity.
[e] Low quality.
[f] Occlusion.
[g] Illumination variation, shadow and scale variation.
[h] Camera motion.
[i] Insufficient data.
[j] Poor weather conditions.
[k] Motion capture.
[l] Skeleton.
[m] Depth.
[n] Human masks.

Table 3
Main multiview datasets in the literature.

Category	Dataset	Input type	Subject	Classes	Repetition	Year	Degraded conditions									
							A	MV	CB	IC	LQ	O	ISS	CM	ID	PW
multiview dataset	Inria XMAS (Weinland et al., 2007)	RGB	11	12	–	2007	*	5								
	I3DPost Multiview (Gkalelis et al., 2009)	RGB	8	12	–	2009	*	8		*						
	MuHAVi (Singh et al., 2010)	RGB	14	17	–	2010	*	8		*			*			
	ATC4[2] Cheng et al. (2012)	RGB,D	24	14	–	2012	*	4	*	*			*			
	Berkeley (Ofli et al., 2013)	RGB,D,MC, A[a], Au[b]	12	11	5	2013	*	4		*						
	DMLSmart Actions (Amiri et al., 2013)	RGB,D, HDRGB[c]	16	12	–	2013	*	3	*			*				
	Multiview3D event (Wei et al., 2013)	RGB,D,S	8	8	20	2013	*	3	*			*				
	NJUST (Song et al., 2014)	RGB,D,S,HM	10	19	2	2014	*	2	*	*						
	Northwestern UCLA (Wang et al., 2014)	RGB,D,S	10	10	–	2014	*	3	*	*						
	UWA3D multiview (Rahmani et al., 2014b)	RGB,D,S	10	30	2–3	2014	*	2		*						
	UTD multiview Action Dataset (Chen et al., 2015)	RGB,D,I[d]	5	6	6	2015	*	5		*						
	Multiview TJU (Liu et al., 2015c)	RGB,D,S	22	20	4	2015	*	2					*			
	UWA3D multiview II (Utdallas, 2018)	RGB,D,S	10	30	–	2015	*	5	*	*						
	NTU RGBD (Shahroudy et al., 2016b)	RGB,D, IR, 3DJoints	40	60	–	2016	*	80				*				
	M^2I (Liu et al., 2017b)	RGB,D,S	22	22	–	2017	*	2	*	*			*			
	NTU RGBD 120 (Liu et al., 2019)	RGB,D, IR, 3DJoints	106	120	–	2019	*	155				*				
	Toyota Smarthomes (Koperski, 2017)	RGB	20	28	7	–	*	7	*			*				
	MDAD (Jegham et al., 2019)	RGB,D	50	16	–	2019	*	2	*	*		*	*			*

[a] Acclerometer signal.
[b] Audio.
[c] High definition RGB.
[d] Inertial sensor data.

Hollywood 2, to handle real world issues. The datasets provide a benchmark for HAR in realistic and challenging contexts. They include variations in human posture and motion, illumination, perspective effects, camera motion, occlusion, and recording environments.

- HOLLYWOOD dataset (Laptev et al., 2008): It was introduced in 2008 and contains diverse video samples. Each sample is labeled

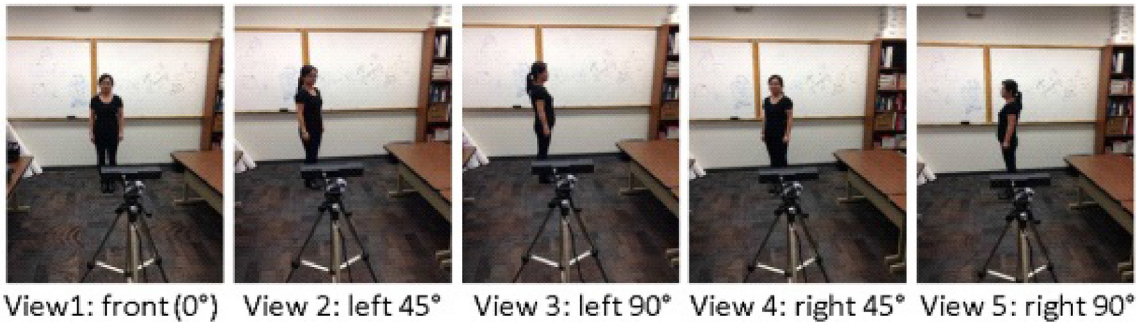

Fig. 6. Five different positions of an actor. UTD multiview Action Dataset (Utdallas, 2018).

Table 4
Main realistic datasets in the literature.

Category	Dataset	Input type	Subject	Classes	Repetition	Year	Degraded conditions									
							A	MV	CB	IC	LQ	O	ISS	CM	ID	PW
Realistic dataset	Hollywood (Laptev et al., 2008)	RGB	–	8	–	2008	*		*	*		*	*	*	*	
	UCF sports (Soomro and Zamir, 2014)	RGB	–	10	–	2008	*						*	*	*	
	Hollywwod2 (Marszalek et al., 2009)	RGB	–	12	–	2009	*		*	*		*	*	*	*	
	UCF Youtube Action (Liu et al., 2009)	RGB	–	11	–	2009	*		*				*		*	
	Olympic (Niebles et al., 2010)	RGB	–	16	–	2010	*		*	*			*	*	*	
	UT Tower (Ryoo et al., 2010)	RGB	6	9	2	2010	*		*	*			*	*		
	ASLAN (Kliper-Gross et al., 2012)	RGB	–	432	–	2012	*		*	*			*	*	*	
	UCF50 (Reddy and Shah, 2013)	RGB	–	50	–	2012	*		*	*			*	*	*	
	Penn (Zhang et al., 2013d)	RGB		15	–	2013	*		*	*		*		*		
	HMDB51 (Kuehne et al., 2013)	RGB	–	51	–	2013	*		*	*	*			*	*	
	UCF101 (Soomro et al., 1212)	RGB	–	101	–	2013	*		*	*			*	*	*	
	MPii (Andriluka et al., 2014)	RGB	40,000	410	–	2014	*		*	*		*				
	UMinho (Afsar et al., 2015)	RGB,IR[a]	–	7	–	2015	*						*			*
	Distracted driver detection (farm, 2016)	RGB	26	10	–	2016	*		*	*		*	*	*		
	Distracted driver (Eraqi et al., 2019)	RGB	26	10	–	2017	*		*	*		*	*			
	SLAC (Zhao et al., 1712)	RGB		200	–	2017	*		*	*		*	*	*		
	MultiTHUMOS (Yeung et al., 2018)	RGB		45	–	2018	*		*	*		*	*	*		

[a] Infra red

according to at least one of eight actions (get out car, answer phone, hand shake, hug person, sit down, sit up, stand up and kiss). The dataset was obtained from 32 movies: a test set extracted from 20 movies and two training sets obtained from 12 other movies.

- HOLLYWOOD2 dataset (Marszalek et al., 2009): It was introduced in 2009 to extend the Hollywood dataset. It contains 12 classes of actions (the same classes of actions presented in Hollywood and four other actions: run, driving car, eat and fight) and 10 categories of scenes distributed over 3669 video obtained from 69 movies and about 20 h of video sequences in total.

4.3.2. UCF datasets

The laboratory of computer vision at the University of Central Florida created several human action datasets (CRCV, 2011). The datasets present many human action classes. Action recognition using these datasets is a challenging task due to their unconstrained environment, abound intraclass variability and the considerable variation in human appearance, camera movement and viewpoint. For instance: the UCF YouTube Action dataset was constructed to recognize actions from video sequences recoded by an amateur using a handheld camera under uncontrolled conditions (in the wild). Moreover, the UCF50 is an extension of UCF YouTube Action Dataset and provides 50 human activity classes. Besides, the UCF 101 that is an extension of UCF50 and contains 101 action classes. In addition, the UCF Sports Action dataset is mainly used for benchmarking HAR algorithms based on temporal template matching.

4.3.3. HMDB51 dataset

In 2011, the Serre research laboratory at Brown University published the HMDB dataset (Kuehne et al., 2011), which contains around 7000 videos manually labeled and extracted from different sources, such as Google videos, YouTube and the Prelinger archive. The large video sequences dataset for human motion recognition is divided into 51 action classes, grouped in five types: body motion for human interaction, general body motion, facial actions with object manipulation, general facial actions and body motion with object interaction. The major challenges associated with the use of video sequences extracted from real world videos are the presence of cluttered background and camera motion.

4.4. Synthetic dataset

4.4.1. PHAV

To train deep action recognition networks, the computer vision research center of "Universitat Autnoma de Barcelona", constructed the Procedural Human Action Videos dataset (Souza et al., 2017) in 2017. It contains 39,982 videos in total, with more than 1000 examples for each human action of 35 categories. It presents 35 different action categories, including 21 simple categories present in HMDB51. In addition to these human actions, 10 action classes involving a single actor and four action classes involving two people interaction were defined. Actions were performed at a specific

Table 5
Overview of random variables of parametric generative model of action videos (Souza et al., 2017).

Parameter	Number of parameters	Values
Human model	20	Models designed by artists
Environment	7	Simple, house interior, lake, Urban, stadium, green, middle
Weather	4	Clear, rain, overcast, fog
Period of day	4	Day, night, dusk, dawn
Variation	5	None, objects, action blending, muscle perturbation and weakening

period of the day, in various environments under certain weather conditions. Table 5 explores the divers parameter used.

5. Conclusion

Automatically understanding and analyzing the actions of a human is an interesting, yet very complex task. Moreover, real world applications in uncontrolled environment makes it more challenging and interesting. Hence, proposing solutions to overcome HAR related issues is a promising research field. This overview brings rich information about HAR real world challenges that have not been fully solved yet, which can help and encourage researchers to work in this trending topic. Anthropometric variations have been solved since they can be effectively considered as intra-class variations that can be overcome by combining different features. Image-quality and frame-rate issues have been partially overcome with the introduction of vision sensors that can provide high-quality images with low noise. Multiview variations represent a motivating challenge for real-world applications in an uncontrolled setting. Introducing new methods to achieve an impressive view invariance is still an open field of research. Methods that address poor weather conditions, insufficient data and camera motion showed promising results but are still insufficient. There have been great efforts to overcome illumination variations, dynamic and cluttered backgrounds and occlusion. However, there is still area for improvement, particularly in real-world settings. Proposing new classifiers to handle the various challenges of HAR remains a promising research direction. The introduction of public datasets that involve all these issues to ensure a good comparison of the performances of the introduced methods is required. During the introduction of new HAR systems, the generalization of proposed approaches have to be considered. Up to our knowledge, none of the proposed methods is able to overcome all these challenges. This overview is a first step in identifying challenges that are not yet fully resolved. Much efforts are still needed to establish a robust system that can address HAR in a global manner.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Afsar, P., Cortez, P., Santos, H., 2015. Automatic human action recognition from video using hidden markov model. In: 2015 IEEE 18th International Conference on Computational Science and Engineering, pp. 105–109. 10.1109/CSE.2015.41.

Aggarwal, J.K., Ryoo, M.S., 2011. Human activity analysis: a review. ACM Comput. Surv. 43 (3), 16.

Aggarwal, J., Xia, L., 2014. Human activity recognition from 3d data: a review. Pattern Recognit. Lett. 48, 70–80. https://doi.org/10.1016/j.patrec.2014.04.011 celebrating the life and work of Maria Petrou. http://www.sciencedirect.com/science/article/pii/S0167865514001299. name="Line_manu_179".

Ahad, M.A.R., Tan, J., Kim, H., Ishikawa, S., 2011. Action dataset a survey. In: SICE Annual Conference 2011, pp. 1650–1655.

Ameur, S., Ben Khalifa, A., Bouhlel, M.S., 2016. A comprehensive leap motion database for hand gesture recognition. In: 2016 7th International Conference on Sciences of Electronics. Technologies of Information and Telecommunications (SETIT), pp. 514–519. 10.1109/SETIT.2016.7939924.

Amiri, S.M., Pourazad, M.T., Nasiopoulos, P., Leung, V.C.M., 2013. Non-intrusive human activity monitoring in a smart home environment. In: 2013 IEEE 15th International Conference on E-Health Networking, Applications and Services (Healthcom 2013), pp. 606–610. 10.1109/HealthCom.2013.6720748.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2d human pose estimation: new benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693. 10.1109/CVPR.2014.471.

Ang, T., Tan, A.W., Loo, C., Wong, W., 2012. Wavelet mach filter for omnidirectional human activity recognition. Intl. Journal of Innovative Computing, Information and Control 8, 1–20.

Anitha, G., Baghavathi Priya, S., 2019. Posture based health monitoring and unusual behavior recognition system for elderly using dynamic bayesian network. Cluster Computing 13583–13590 https://doi.org/10.1007/s10586-018-2010-9.

Baek, J., Yun, B., 2008. A sequence-action recognition applying state machine for user interface. IEEE Trans. Consum. Electron. 54 (2), 719–726. 10.1109/TCE.2008.4560153.

F. Baradel, C. Wolf, J. Mille, Pose-conditioned Spatio-Temporal Attention for Human Action Recognition, arXiv preprint arXiv:1703.10106.

Barnachon, M., Bouakaz, S., Boufama, B., Guillou, E., 2014. Ongoing human action recognition with motion capture. Pattern Recognit. 47 (1), 238–247. https://doi.org/10.1016/j.patcog.2013.06.020. http://www.sciencedirect.com/science/article/pii/S0031320313002720. name="Line_manu_282".

Berrached, C., 2014. Systeme de reconnaissance de gestes de la main. Ph.D. thesis. Université Abou Bekr Belkaid Tlemcen.

Bilinski, P., Bremond, F., 2012. Contextual statistics of space-time ordered features for human action recognition. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 228–233. 10.1109/AVSS.2012.29.

Bilinski, P.T., Bremond, F., 2015. Video covariance matrix logarithm for human action recognition in videos. In: IJCAI, pp. 2140–2147.

Bilinski, P., Corvee, E., Bak, S., Bremond, F., 2013. Relative dense tracklets for human action recognition. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7. https://doi.org/10.1109/FG.2013.6553699.

Bloom, V., Makris, D., Argyriou, V., 2012. G3d: a gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 7–12. https://doi.org/10.1109/CVPRW.2012.6239175.

Bobick, A., Davis, J., 1996. An appearance-based representation of action. In: Proceedings of 13th International Conference on Pattern Recognition, vol. 1, pp. 307–312. https://doi.org/10.1109/ICPR.1996.546039 vol. 1.

Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23 (3), 257–267. 10.1109/34.910878.

Brand, M., Oliver, N., Pentland, A., 1997. Coupled hidden markov models for complex action recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 994–999. 10.1109/CVPR.1997.609450.

Carletti, V., Foggia, P., Percannella, G., Saggese, A., Vento, M., 2013. Recognition of human actions from rgb-d videos using a reject option. In: International Conference on Image Analysis and Processing. Springer, pp. 436–445.

Chaaraoui, A.A., Climent-Prez, P., Flrez-Revuelta, F., 2013. Silhouette-based human action recognition using sequences of key poses. Pattern Recognit. Lett. 34 (15), 1799–1807. https://doi.org/10.1016/j.patrec.2013.01.021 smart Approaches for Human Action Recognition. http://www.sciencedirect.com/science/article/pii/S0167865513000342. name="Line_manu_246".

Chakraborty, B.K., Sarma, D., Bhuyan, M.K., MacDorman, K.F., 2017. Review of constraints on vision-based gesture recognition for human–computer interaction. IET Comput. Vis. 12 (1), 3–15.

Chang, J.Y., 2016. Nonparametric feature matching based conditional random fields for gesture recognition from multi-modal video. IEEE Trans. Pattern Anal. Mach. Intell. 38 (8), 1612–1625. 10.1109/TPAMI.2016.2519021.

Chaquet, J.M., Carmona, E.J., Fernndez-Caballero, A., 2013. A survey of video datasets for human action and activity recognition. Comput. Vis. Image Understand. 117 (6), 633–659. https://doi.org/10.1016/j.cviu.2013.01.013. http://www.sciencedirect.com/science/article/pii/S1077314213000295. name="Line_manu_175".

Chebli, K., Ben Khalifa, A., 2018. Pedestrian detection based on background compensation with block-matching algorithm. In: 15th International Multi-Conference on Systems, Signals Devices (SSD). IEEE, pp. 497–501. https://doi.org/10.1109/SSD.2018.8570499.

Chen, W., 2006. Meng Joo Er, Shiqian Wu, Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36 (2), 458–466. 10.1109/TSMCB.2005.857353.

Chen, C.-C., Aggarwal, J., 2011. Modeling human activities as speech. In: CVPR 2011. IEEE, pp. 3425–3432.

Chen, C., Jafari, R., Kehtarnavaz, N., 2015. Utd-mhad: a multimodal dataset for

human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 168–172. 10.1109/ICIP.2015.7350781.
Chen, C., Jafari, R., Kehtarnavaz, N., 2017. A survey of depth and inertial sensor fusion for human action recognition. Multimed. Tools Appl. 76 (3), 4405–4425.
Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q., 2012. Human daily action analysis with multi-view and color-depth data. In: European Conference on Computer Vision. Springer, pp. 52–61.
Chron, G., Laptev, I., Schmid, C., 2015. P-cnn: pose-based cnn features for action recognition. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3218–3226. 10.1109/ICCV.2015.368.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27. 10.1109/TIT.1967.1053964.
CRCV, 2011. Ucf datasets. last accessed 03/01/2019. http://crcv.ucf.edu/data/. name="Line_manu_340".
Cumin, J., Lefebvre, G., 2016. A priori data and a posteriori decision fusions for human action recognition. In: 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP).
X. Dai, J. Liu, Z. Han, Z. Liu, Real-time single-view action recognition based on key pose analysis for sports videos, uS Patent 9, 600,717 (Mar. 21 2017).
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177.
Dalal, N., Triggs, B., Schmid, C., 2006. Human detection using oriented histograms of flow and appearance. In: European Conference on Computer Vision. Springer, pp. 428–441.
Debes, C., Merentitis, A., Sukhanov, S., Niessen, M., Frangiadakis, N., Bauer, A., 2016. Monitoring activities of daily living in smart homes: understanding human behavior. IEEE Signal Process. Mag. 33 (2), 81–94. 10.1109/MSP.2015.2503881.
Dhamsania, C.J., Ratanpara, T.V., 2016. A survey on human action recognition from videos. In: 2016 Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1–5. https://doi.org/10.1109/GET.2016.7916717.
Dhulekar, P., Gandhe, S., Chitte, H., Pardeshi, K., 2017. Human action recognition: an overview. In: Proceedings of the International Conference on Data Engineering and Communication Technology. Springer, pp. 481–488.
Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. 10.1109/VSPETS.2005.1570899.
Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J., 2009. Automatic annotation of human actions in video. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1491–1498. 10.1109/ICCV.2009.5459279.
Duckworth, P., Alomari, M., Gatsoulis, Y., Hogg, D.C., Cohn, A.G., 2016. Unsupervised activity recognition using latent semantic analysis on a mobile robot. In: IOS Press Proceedings, vol. 285, pp. 1062–1070.
Duta, I.C., Uijlings, J.R., Ionescu, B., Aizawa, K., Hauptmann, A.G., Sebe, N., 2017. Efficient human action recognition using histograms of motion gradients and vlad with descriptor shape information. Multimed. Tools Appl. 76 (21), 22445–22472.
Efros, A.A., Berg, A.C., Mori, G., Malik, J., 2003. Recognizing action at a distance. In: Null. IEEE, p. 726.
Eraqi, H.M., Abouelnaga, Y., Saad, M.H., Moustafa, M.N., 2019. Driver distraction identification with an ensemble of convolutional neural networks. Journal of Advanced Transportation 2019, 1–12. https://doi.org/10.1155/2019/4125865.
farm, S., 2016. Distracted driver detection dataset. last accessed 17/02/2019. https://www.kaggle.com/c/state-farm-distracted-driver-detection/. name="Line_manu_378".
Forsyth, D.A., Arikan, O., Ikemoto, L., O'Brien, J., Ramanan, D., et al., 2006. Computational studies of human motion: part 1, tracking and motion synthesis. Found. Trends® Comput. Graph. Vis. 1 (2–3), 77–254.
Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I., 2009. The i3dpost multi-view and 3d human action/interaction database. In: 2009 Conference for Visual Media Production, pp. 159–168. 10.1109/CVMP.2009.19.
Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R., 2007a. Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. 29 (12), 2247–2253. 10.1109/TPAMI.2007.70711.
Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R., 2007b. Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. 29 (12), 2247–2253. 10.1109/TPAMI.2007.70711.
Grushin, A., Monner, D.D., Reggia, J.A., Mishra, A., 2013. Robust human action recognition via long short-term memory. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. 10.1109/IJCNN.2013.6706797.
Guha, T., Ward, R.K., 2012. Learning sparse representations for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 34 (8), 1576–1588. 10.1109/TPAMI.2011.253.
Guo, G., Lai, A., 2014. A survey on still image based human action recognition. Pattern Recognit. 47 (10), 3343–3361. https://doi.org/10.1016/j.patcog.2014.04.018. http://www.sciencedirect.com/science/article/pii/S0031320314001642. name="Line_manu_184".
Guo, K., Ishwar, P., Konrad, J., 2013. Action recognition from video using feature covariance matrices. IEEE Trans. Image Process. 22 (6), 2479–2494. 10.1109/TIP.2013.2252622.
Hadfield, S., Lebeda, K., Bowden, R., 2017. Hollywood 3d: what are the best 3d features for action recognition? Int. J. Comput. Vis. 121 (1), 95–110.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
He, W., Liu, B., Xiao, Y., 2017. Multi-view action recognition method based on regularized extreme learning machine. In: 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 1, pp. 854–857. 10.1109/CSE-EUC.2017.171.
Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: a survey, Image and Vision Computing 60, 4 – 21, regularization Techniques for High-Dimensional Data Analysis. https://doi.org/10.1016/j.imavis.2017.01.010. http://www.sciencedirect.com/science/article/pii/S0262885617300343. name="Line_manu_188".
Hong, S., Ryu, J., Yang, H.S., 2018. Not all frames are equal: aggregating salient features for dynamic texture classification. Multidimens. Syst. Signal Process. 29 (1), 279–298.
A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv preprint arXiv:1704.04861.
Hu, J., Zheng, W., Lai, J., Zhang, J., 2017. Jointly learning heterogeneous features for rgb-d activity recognition. IEEE Trans. Pattern Anal. Mach. Intell. 39 (11), 2186–2200. 10.1109/TPAMI.2016.2640292.
Ikizler-Cinbis, N., Sclaroff, S., 2012. Web-based classifiers for human action recognition. IEEE Trans. Multimed. 14 (4), 1031–1045. 10.1109/TMM.2012.2187180.
Ikizler-Cinbis, N., Gokberk Cinbis, R., Sclaroff, S., 2009. Learning actions from the web. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 995–1002. 10.1109/ICCV.2009.5459368.
INRIA, 2008. Inria xmas motion acquisition sequences (ixmas). last accessed 25/01/2019. http://4drepository.inrialpes.fr/public/viewgroup/6. name="Line_manu_336".
Iosifidis, A., Tefas, A., Pitas, I., 2012. Neural representation and learning for multi-view human action recognition. In: IJCNN, pp. 1–6.
Ishan Misra, Lawrence Zitnick, C., Martial, H., 2016. Unsupervised learning using sequential verification for action recognition. Computer Science, Published in ArXiv. https://arxiv.org/pdf/1603.08561.pdf.
Jegham, I., Ben Khalifa, A., 2017. Pedestrian detection in poor weather conditions using moving camera. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 358–362. 10.1109/AICCSA.2017.35.
Jegham, I., Ben khalifa, A., Alouani, I., Mahjoub, M.A., 2018. Safe driving : driver action recognition using surf keypoints. In: 2018 30th International Conference on Microelectronics (ICM), pp. 60–63. 10.1109/ICM.2018.8704009.
Jegham, I., Ben Khalifa, A., Alouani, I., Mahjoub, M.A., 2019. Mdad: a multimodal and multiview in-vehicle driver action dataset. In: Computer Analysis of Images and Patterns. Springer International Publishing, Cham, pp. 518–529.
Ji, X., Liu, H., 2010. Advances in view-invariant human motion analysis: a review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40 (1), 13–24. 10.1109/TSMCC.2009.2027608.
Ji, X., Wang, C., Li, Y., Wu, Q., 2013a. Hidden markov model-based human action recognition using mixed features. J. Comput. Inf. Syst. 9, 3659–3666.
Ji, S., Xu, W., Yang, M., Yu, K., 2013b. 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 221–231. 10.1109/TPAMI.2012.59.
Ji, Y., Yang, Y., Shen, F., Shen, H.T., Li, X., 2019. A survey of human action analysis in hri applications. IEEE Trans. Circuits Syst. Video Technol. 1–1.
Jiang, Hao, Drew, M.S., Li, Ze-Nian, 2006. Successive convex matching for action detection. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1646–1653. 10.1109/CVPR.2006.297.
Jiang, Z., Lin, Z., Davis, L., 2012. Recognizing human actions by learning and matching shape-motion prototype trees. IEEE Trans. Pattern Anal. Mach. Intell. 34 (3), 533–547. 10.1109/TPAMI.2011.147.
Jun, B., Choi, I., Kim, D., 2013. Local transform features and hybridization for accurate face and human detection. IEEE Trans. Pattern Anal. Mach. Intell. 35 (6), 1423–1436. 10.1109/TPAMI.2012.219.
Junejo, I.N., Dexter, E., Laptev, I., Perez, P., 2011. View-independent action recognition from temporal self-similarities. IEEE Trans. Pattern Anal. Mach. Intell. 33 (1), 172–185. 10.1109/TPAMI.2010.68.
Kliper-Gross, O., Hassner, T., Wolf, L., 2012. The action similarity labeling challenge. IEEE Trans. Pattern Anal. Mach. Intell. 34 (3), 615–621. 10.1109/TPAMI.2011.209.
Koperski, M., 2017. Human Action Recognition in Videos with Local Representation. Ph.D. thesis. Université Côte d'Azur.
Koppula, H.S., Gupta, R., Saxena, A., 2013. Learning human activities and object affordances from rgb-d videos. Int. J. Robot. Res. 32 (8), 951–970.
Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. Hmdb: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision, pp. 2556–2563. 10.1109/ICCV.2011.6126543.
Kuehne, H., Jhuang, H., Stiefelhagen, R., Serre, T., 2013. Hmdb51: a large video database for human motion recognition. In: High Performance Computing in Science and Engineering, vol. 12. Springer, pp. 571–582.
Kumar, K., Bhavani, R., 2016. Analysis of svm and knn classifiers for egocentric activity recognition. In: Proceedings of the International Conference on Informatics and Analytics. ACM, p. 83.
Kumari, S., Mitra, S.K., 2011. Human action recognition using dft. In: 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 239–242. 10.1109/NCVPRIPG.2011.58.

Laptev, I., 2005. On space-time interest points. Int. J. Comput. Vis. 64 (2–3), 107–123.
Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. https://doi.org/10.1109/CVPR.2008.4587756.
Lara, O.D., Labrador, M.A., 2013. A survey on human activity recognition using wearable sensors. IEEE Communications Surveys Tutorials 15 (3), 1192–1209. 10.1109/SURV.2012.110112.00192.
Lavee, G., Rudzsky, M., Rivlin, E., 2013. Propagating certainty in petri nets for activity recognition. IEEE Trans. Circuits Syst. Video Technol. 23 (2), 326–337. 10.1109/TCSVT.2012.2203742.
Lei, Q., Zhang, H., Xin, M., Cai, Y., 2018. A hierarchical representation for human action recognition in realistic scenes. Multimed. Tools Appl. 77 (9), 11403–11423.
Lejmi, W., Ben Khalifa, A., Mahjoub, M.A., 2017a. Fusion strategies for recognition of violence actions. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 178–183. 10.1109/AICCSA.2017.193.
Lejmi, W., Mahjoub, M.A., Ben Khalifa, A., 2017b. Event detection in video sequences: challenges and perspectives. In: 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 682–690. 10.1109/FSKD.2017.8393354.
Li, R., Zickler, T., 2012. Discriminative virtual views for cross-view action recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2855–2862. 10.1109/CVPR.2012.6248011.
Li, W., Zhang, Z., Liu, Z., 2010. Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 9–14. 10.1109/CVPRW.2010.5543273.
Lin, S., Shie, C., Chen, S., Lee, M., Hung, Y., 2012. Human action recognition using action trait code. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 3456–3459.
Liu, C., Yuen, P.C., 2010. Human action recognition using boosted eigenactions. Image Vis Comput. 28 (5), 825–835. https://doi.org/10.1016/j.imavis.2009.07.009 best of Automatic Face and Gesture Recognition 2008. http://www.sciencedirect.com/science/article/pii/S0262885609001577. name="Line_manu_302".
Liu, J., Luo, Jiebo, Shah, M., 2009. Recognizing realistic actions from videos in the wild. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003. 10.1109/CVPR.2009.5206744.
Liu, A.-A., Xu, N., Su, Y.-T., Lin, H., Hao, T., Yang, Z.-X., 2015a. Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 151, 544–553. https://doi.org/10.1016/j.neucom.2014.04.090. http://www.sciencedirect.com/science/article/pii/S0925231214013885. name="Line_manu_244".
Liu, A.-A., Nie, W.-Z., Su, Y.-T., Ma, L., Hao, T., Yang, Z.-X., 2015b. Coupled hidden conditional random fields for rgb-d human action recognition. Signal Process. 112, 74–82. https://doi.org/10.1016/j.sigpro.2014.08.038 signal Processing and Learning Methods for 3D Semantic Analysis. http://www.sciencedirect.com/science/article/pii/S0165168414004022. name="Line_manu_274".
Liu, A., Su, Y., Jia, P., Gao, Z., Hao, T., Yang, Z., 2015c. Multiple/single-view human action recognition via part-induced multitask structural learning. IEEE Transactions on Cybernetics 45 (6), 1194–1208. 10.1109/TCYB.2014.2347057.
Liu, T., Wang, X., Dai, X., Luo, J., 2016a. Deep recursive and hierarchical conditional random fields for human action recognition. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. https://doi.org/10.1109/WACV.2016.7477694.
Liu, C., Liu, J., He, Z., Zhai, Y., Hu, Q., Huang, Y., 2016b. Convolutional neural random fields for action recognition. Pattern Recognit. 59, 213–224. https://doi.org/10.1016/j.patcog.2016.03.019 compositional Models and Structured Learning for Visual Recognition. http://www.sciencedirect.com/science/article/pii/S0031320316300048. name="Line_manu_277".
Liu, J., Shahroudy, A., Xu, D., Wang, G., 2016c. Spatio-temporal lstm with trust gates for 3d human action recognition. In: European Conference on Computer Vision. Springer, pp. 816–833.
Liu, H., Ju, Z., Ji, X., Chan, C.S., Khoury, M., 2017a. A view-invariant action recognition action recognition based on multi-view space hidden markov models. In: Human Motion Sensing and Recognition. Springer, pp. 251–267.
Liu, A., Xu, N., Nie, W., Su, Y., Wong, Y., Kankanhalli, M., 2017b. Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Transactions on Cybernetics 47 (7), 1781–1794. 10.1109/TCYB.2016.2582918.
Liu, J., Wang, G., Hu, P., Duan, L., Kot, A.C., 2017c. Global context-aware attention lstm networks for 3d action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3671–3680. 10.1109/CVPR.2017.391.
Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L., Kot Chichung, A., 2019. Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2019.2916873, 1–1.
Lu, Y., Li, Y., Shen, Y., Ding, F., Wang, X., Hu, J., Ding, S., 2012. A human action recognition method based on tchebichef moment invariants and temporal templates. In: 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, vol. 2, pp. 76–79. https://doi.org/10.1109/IHMSC.2012.114.
Lucas, B.D., Kanade, T., 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In: Proceedings of the 7th international joint conference on Artificial intelligence, vol. 2, pp. 674–679. https://dl.acm.org/doi/10.5555/1623264.1623280.
Maji, S., Bourdev, L., Malik, J., 2011. Action recognition from a distributed representation of pose and appearance. In: CVPR 2011, pp. 3177–3184. 10.1109/CVPR.2011.5995631.
Mansur, A., Makihara, Y., Yagi, Y., 2013. Inverse dynamics for action recognition. IEEE Transactions on Cybernetics 43 (4), 1226–1236. 10.1109/TSMCB.2012.2226879.
Marszalek, M., Laptev, I., Schmid, C., 2009. Actions in context. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936. 10.1109/CVPR.2009.5206557.
Messing, R., Pal, C., Kautz, H., 2009. Activity recognition using the velocity histories of tracked keypoints. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 104–111. 10.1109/ICCV.2009.5459154.
Mimouna, A., Ben Khalifa, A., Ben Amara, N.E., 2018. Human action recognition using triaxial accelerometer data: selective approach. In: 2018 15th International Multi-Conference on Systems, Signals Devices (SSD). IEEE, pp. 491–496. 10.1109/SSD.2018.8570429.
Minhas, R., Baradarani, A., Seifzadeh, S., Wu, Q.J., 2010. Human action recognition using extreme learning machine based on visual vocabularies. Neurocomputing 73 (10), 1906–1917. https://doi.org/10.1016/j.neucom.2010.01.020 subspace Learning/Selected papers from the European Symposium on Time Series Prediction. http://www.sciencedirect.com/science/article/pii/S0925231210001517. name="Line_manu_283".
Moeslund, T.B., Hilton, A., Krger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Understand. 104 (2), 90–126. https://doi.org/10.1016/j.cviu.2006.08.002 special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour. http://www.sciencedirect.com/science/article/pii/S1077314206001263. name="Line_manu_181".
Müller, M., 2007. Dynamic Time Warping, Information Retrieval for Music and Motion, pp. 69–84.
Munaro, M., Basso, A., Fossati, A., Van Gool, L., Menegatti, E., 2014. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 4512–4519. 10.1109/ICRA.2014.6907518.
K. P. Murphy, Dynamic bayesian networks, probabilistic graphical models, M. Jordan 7.
Nasiri, J.A., Charkari, N.M., Mozafari, K., 2017. Human action recognition by fuzzy hidden markov model. Int. J. Comput. Vis. Robot. 7 (5), 538–557.
Natarajan, P., Nevatia, R., 2008. Online, real-time tracking and recognition of human actions. In: 2008 IEEE Workshop on Motion and Video Computing, pp. 1–8. https://doi.org/10.1109/WMVC.2008.4544064.
Nazir, S., Yousaf, M.H., Velastin, S.A., 2018. Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. Comput. Electr. Eng. 72, 660–669. https://doi.org/10.1016/j.compeleceng.2018.01.037. http://www.sciencedirect.com/science/article/pii/S0045790616306863. name="Line_manu_228".
Negin, F., Özdemir, F., Akgül, C.B., Yüksel, K.A., Erçil, A., 2013. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In: International Conference Image Analysis and Recognition. Springer, pp. 648–657.
Nez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Vlez, J.F., 2018. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. 76, 80–94. https://doi.org/10.1016/j.patcog.2017.10.033. http://www.sciencedirect.com/science/article/pii/S0031320317304405. name="Line_manu_333".
Ng, A.Y., Jordan, M.I., 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. In: Advances in Neural Information Processing Systems, pp. 841–848.
Ni, B., Wang, G., Moulin p rgbd, H., 2011. A colour-depth video database for human daily activity recognition [c]. In: Proceedings of IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 6–13. November.
Niebles, J.C., Wang, H., Fei-Fei, L., 2008. Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vis. 79 (3), 299–318.
Niebles, J.C., Chen, C.-W., Fei-Fei, L., 2010. Modeling temporal structure of decomposable motion segments for activity classification. In: European Conference on Computer Vision. Springer, pp. 392–405.
Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R., 2013. Berkeley mhad: a comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 53–60. 10.1109/WACV.2013.6474999.
Oikonomopoulos, A., Patras, I., Pantic, M., 2006. Spatiotemporal salient points for visual recognition of human actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36 (3), 710–719. 10.1109/TSMCB.2005.861864.
Ordóñez, F.J., Roggen, D., 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16 (1), 115.
Oreifej, O., Liu, Z., 2013. Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723. 10.1109/CVPR.2013.98.
Pan, R., Ma, L., Zhan, Y., Cai, S., 2017. A novel orientation-context descriptor and locality-preserving Fisher discrimination dictionary learning for action recognition. In: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. 10.1109/DICTA.2017.8227395.
Park, S., Aggarwal, J.K., 2004. A hierarchical bayesian network for event recognition of human actions and interactions. Multimed. Syst. 10 (2), 164–179.
Peng, X., Qiao, Y., Peng, Q., Qi, X., 2013. Exploring motion boundary based sampling

and spatial-temporal context descriptors for action recognition. BMVC 20, 93–96.
Pfeifer, M., Voelker, B., 2015. Sensors in human activity recognition. In: Seminar Course.
Piyathilaka, L., Kodagoda, S., 2015. Human activity recognition for domestic robots. In: Field and Service Robotics. Springer, pp. 395–408.
Poppe, R., 2010. A survey on vision-based human action recognition. Image Vis Comput. 28 (6), 976–990. https://doi.org/10.1016/j.imavis.2009.11.014. http://www.sciencedirect.com/science/article/pii/S0262885609002704. name="Line_manu_193".
Quattoni, A., Wang, S., Morency, L., Collins, M., Darrell, T., 2007. Hidden conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell. 29 (10), 1848–1852. 10.1109/TPAMI.2007.1124.
Rahman, S., See, J., 2016. Spatio-temporal mid-level feature bank for action recognition in low quality video. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1846–1850. 10.1109/ICASSP.2016.7471996.
Rahman, S., Cho, S.-Y., Leung, M., 2012. Recognising human actions by analysing negative spaces. IET Comput. Vis. 6 (3), 197–213.
Rahman, S., See, J., Ho, C.C., 2015. Action recognition in low quality videos by jointly using shape, motion and texture features. In: 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 83–88. 10.1109/ICSIPA.2015.7412168.
H. Rahmani, A. Mahmood, D. Huynh, A. Mian, Action Classification with Locality-Constrained Linear Coding, arXiv preprint arXiv:1408.3810.
Rahmani, H., Mahmood, A., Huynh, D.Q., Mian, A., 2014a. Real time action recognition using histograms of depth gradients and random decision forests. In: IEEE Winter Conference on Applications of Computer Vision, pp. 626–633. 10.1109/WACV.2014.6836044.
Rahmani, H., Mahmood, A., Huynh, D.Q., Mian, A., 2014b. Hopc: histogram of oriented principal components of 3d pointclouds for action recognition. In: European Conference on Computer Vision. Springer, pp. 742–757.
Rahmani, H., Huynh, D.Q., Mahmood, A., Mian, A., 2016a. Discriminative human action classification using locality-constrained linear coding. Pattern Recognit. Lett. 72, 62–71. https://doi.org/10.1016/j.patrec.2015.07.015 special Issue on ICPR 2014 Awarded Papers. http://www.sciencedirect.com/science/article/pii/S016786551500224X. name="Line_manu_224".
Rahmani, H., Mahmood, A., Huynh, D., Mian, A., 2016b. Histogram of oriented principal components for cross-view action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 38 (12), 2430–2443. 10.1109/TPAMI.2016.2533389.
Rahmani, H., Mian, A., Shah, M., 2018. Learning a deep model for human action recognition from novel viewpoints. IEEE Trans. Pattern Anal. Mach. Intell. 40 (3), 667–681. 10.1109/TPAMI.2017.2691768.
Ramanathan, M., Yau, W., Teoh, E.K., 2014. Human action recognition with video data: research and evaluation challenges. IEEE Transactions on Human-Machine Systems 44 (5), 650–663. 10.1109/THMS.2014.2325871.
Ramirez-Amaro, K., Kim, Eun-Sol, Kim, Jiseob, Zhang, Byoung-Tak, Beetz, M., Cheng, G., 2013. Enhancing human action recognition through spatio-temporal feature learning and semantic rules. In: 2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 456–461. 10.1109/HUMANOIDS.2013.7030014.
Raptis, M., Sigal, L., 2013. Poselet key-framing: a model for human activity recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2650–2657. 10.1109/CVPR.2013.342.
Reddy, K.K., Shah, M., 2013. Recognizing 50 human action categories of web videos. Mach. Vis. Appl. 24 (5), 971–981.
Reddy, K.K., Cuntoor, N., Perera, A., Hoogs, A., 2012. Human action recognition in large-scale datasets using histogram of spatiotemporal gradients. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance. IEEE, pp. 106–111.
Ryoo, M.S., Aggarwal, J.K., 2009. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1593–1600. 10.1109/ICCV.2009.5459361.
Ryoo, M., Chen, C.-C., Aggarwal, J., Roy-Chowdhury, A., 2010. An overview of contest on semantic description of human activities (sdha) 2010. In: Recognizing Patterns in Signals, Speech, Images and Videos. Springer, pp. 270–285.
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, vol. 3, pp. 32–36. ICPR 2004. 10.1109/ICPR.2004.1334462.
Sempena, S., 2011. Nur ulfa maulidevi, peb ruswono aryan, human action recognition using dynamic time warping. In: Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, pp. 1–5. https://doi.org/10.1109/ICEEI.2011.6021605.
Shahroudy, A., Ng, T., Yang, Q., Wang, G., 2016a. Multimodal multipart learning for action recognition in depth videos. IEEE Trans. Pattern Anal. Mach. Intell. 38 (10), 2123–2129. 10.1109/TPAMI.2015.2505295.
Shahroudy, A., Liu, J., Ng, T., Wang, G., 2016b. Ntu rgb+d: a large scale dataset for 3d human activity analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1010–1019. 10.1109/CVPR.2016.115.
Shan, Shiguang, Gao, Wen, Cao, Bo, Zhao, Debin, 2003. Illumination normalization for robust face recognition against varying lighting conditions. In: 2003 IEEE International SOI Conference. Proceedings (Cat. No.03CH37443), pp. 157–164. 10.1109/AMFG.2003.1240838.
Shao, L., Ji, L., Liu, Y., Zhang, J., 2012. Human action segmentation and recognition via motion and shape analysis. Pattern Recognit. Lett. 33 (4), 438–445. https://doi.org/10.1016/j.patrec.2011.05.015 intelligent Multimedia Interactivity. http://www.sciencedirect.com/science/article/pii/S0167865511001590. name="Line_manu_256".
Shechtman, E., Irani, M., 2007. Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? IEEE Trans. Pattern Anal. Mach. Intell. 29 (11), 2045–2056. 10.1109/TPAMI.2007.1119.
Sigal, L., Balan, A.O., Black, M.J., Humaneva, 2010. Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. 87 (1–2), 4.
Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576.
Singh, T., Vishwakarma, D.K., 2019. Human activity recognition in video benchmarks: a survey. In: Advances in Signal Processing and Communication. Springer, pp. 247–259.
Singh, M., Basu, A., Mandal, M.K., 2008. Human activity recognition based on silhouette directionality. IEEE Trans. Circuits Syst. Video Technol. 18 (9), 1280–1292. 10.1109/TCSVT.2008.928888.
Singh, S., Velastin, S.A., Ragheb, H., 2010. Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods. In: 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 48–55. https://doi.org/10.1109/AVSS.2010.63.
Sipiran, I., Bustos, B., 2011. Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. Vis. Comput. 27 (11), 963.
Song, Y., Morency, L., Davis, R., 2013. Action recognition by hierarchical sequence summarization. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3562–3569. 10.1109/CVPR.2013.457.
Song, Y., Tang, J., Liu, F., Yan, S., 2014. Body surface context: a new robust feature for action recognition from depth videos. IEEE Trans. Circuits Syst. Video Technol. 24 (6), 952–964. 10.1109/TCSVT.2014.2302558.
Soomro, K., Zamir, A.R., 2014. Action recognition in realistic sports videos. In: Computer Vision in Sports. Springer, pp. 181–208.
K. Soomro, A. R. Zamir, M. Shah, Ucf101: A Dataset of 101 Human Actions Classes from Videos in the Wild, arXiv preprint arXiv:1212.0402.
Soumya, T., Thampi, S.M., 2015. Day color transfer based night video enhancement for surveillance system. In: 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), pp. 1–5. https://doi.org/10.1109/SPICES.2015.7091556.
Souza, C.R.d., Gaidon, A., Cabon, Y., Lpez, A.M., 2017. Procedural generation of videos to train deep action recognition networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2594–2604. 10.1109/CVPR.2017.278.
Subetha, T., Chitrakala, S., 2016. A survey on human activity recognition from videos. In: 2016 International Conference on Information Communication and Embedded Systems (ICICES), pp. 1–7. https://doi.org/10.1109/ICICES.2016.7518920.
Subramanian, K., Suresh, S., 2012. Human action recognition using meta-cognitive neuro-fuzzy inference system. Int. J. Neural Syst. 22, 1250028, 06.
Sun, L., Jia, K., Chan, T.-H., Fang, Y., Wang, G., Yan, S., 2014. Dl-sfa: deeply-learned slow feature analysis for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2632.
Sung, J., Ponce, C., Selman, B., Saxena, A., 2011. Human Activity Detection from Rgbd images. Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 47–55. https://dl.acm.org/doi/10.5555/2908772.2908779.
Sunny, J.T., George, S.M., Kizhakkethottam, J.J., Sunny, J.T., George, S.M., Kizhakkethottam, J.J., 2015. Applications and challenges of human activity recognition using sensors in a smart environment. IJIRST Int. J. Innov. Res. Sci. Technol 2, 50–57.
Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9 (3), 293–300.
Tran, D., Sorokin, A., 2008. Human activity recognition with metric learning. In: European Conference on Computer Vision. Springer, pp. 548–561.
Trinh, H., Fan, Q., Jiyan, P., Gabbur, P., Miyazawa, S., Pankanti, S., 2011. Detecting human activities in retail surveillance using hierarchical finite state machine. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1337–1340. 10.1109/ICASSP.2011.5946659.
Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O., 2008. Machine recognition of human activities: a survey. IEEE Trans. Circuits Syst. Video Technol. 18 (11), 1473–1488. 10.1109/TCSVT.2008.2005594.
Utdallas, 2018. Utd multi-view action dataset. last accessed 02/01/2019. http://www.utdallas.edu/~kehtar/MultiViewDataset.pdf. name="Line_manu_337".
Vail, D.L., Veloso, M.M., Lafferty, J.D., 2007. Conditional random fields for activity recognition. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. ACM, p. 235.
Valentin, N.Z.E., 2010. Multisensor Fusion for Monitoring Elderly Activities at Home. Ph.D. thesis. Université Nice Sophia Antipolis.
Varol, G., Laptev, I., Schmid, C., 2018. Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1510–1517. 10.1109/TPAMI.2017.2712608.
Vasconez, J.P., Cheein, F.A., 2018. Finding a proper approach to obtain cognitive parameters from human faces under illumination variations. In: 2018 5th International Conference on Control. Decision and Information Technologies (CoDIT), pp. 946–951. 10.1109/CoDIT.2018.8394947.
Wang, Y., Mori, G., 2009. Max-margin hidden conditional random fields for human

action recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 872–879. 10.1109/CVPR.2009.5206709.
Wang, H., Schmid, C., 2013. Action recognition with improved trajectories. In: 2013 IEEE International Conference on Computer Vision, pp. 3551–3558. 10.1109/ICCV.2013.441.
Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British Machine Vision Conference. BMVA Press, 124–1.
Wang, H., Klser, A., Schmid, C., Liu, C., 2011. Action recognition by dense trajectories. In: CVPR 2011, pp. 3169–3176. 10.1109/CVPR.2011.5995407.
Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297. 10.1109/CVPR.2012.6247813.
Wang, H., Kläser, A., Schmid, C., Liu, C.-L., 2013a. Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. 103 (1), 60–79.
Wang, Z., Guan, G., Qiu, Y., Zhuo, L., Feng, D., 2013b. Semantic context based refinement for news video annotation. Multimed. Tools Appl. 67 (3), 607–627.
Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S., 2014. Cross-view action modeling, learning, and recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2649–2656. 10.1109/CVPR.2014.339.
Wang, L., Qiao, Y., Tang, X., 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4305–4314.
Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S., 2018. Rgb-d-based human motion recognition with deep learning: a survey. Comput. Vis. Image Understand. 171, 118–139. https://doi.org/10.1016/j.cviu.2018.04.007. http://www.sciencedirect.com/science/article/pii/S1077314218300663. name="Line_manu_189".
Wei, P., Zhao, Y., Zheng, N., Zhu, S., 2013. Modeling 4d human-object interactions for event and object recognition. In: 2013 IEEE International Conference on Computer Vision, pp. 3272–3279. 10.1109/ICCV.2013.406.
Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Understand. 104 (2), 249–257. https://doi.org/10.1016/j.cviu.2006.07.013 special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour. http://www.sciencedirect.com/science/article/pii/S1077314206001081. name="Line_manu_198".
Weinland, D., Boyer, E., Ronfard, R., 2007. Action recognition from arbitrary views using 3d exemplars. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–7. https://doi.org/10.1109/ICCV.2007.4408849.
Weinland, D., Özuysal, M., Fua, P., 2010. Making action recognition robust to occlusions and viewpoint changes. In: European Conference on Computer Vision. Springer, pp. 635–648.
Weinland, D., Ronfard, R., Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Understand. 115 (2), 224–241. https://doi.org/10.1016/j.cviu.2010.10.002. http://www.sciencedirect.com/science/article/pii/S1077314210002171. name="Line_manu_183".
Willems, G., Tuytelaars, T., Van Gool, L., 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. In: European Conference on Computer Vision. Springer, pp. 650–663.
Wu, D., Shao, L., 2013. Silhouette analysis-based action recognition via exploiting human poses. IEEE Trans. Circuits Syst. Video Technol. 23 (2), 236–243. 10.1109/TCSVT.2012.2203731.
Wu, X., Xu, D., Duan, L., Luo, J., 2011. Action recognition using context and appearance distribution features. In: CVPR 2011, pp. 489–496. 10.1109/CVPR.2011.5995624.
Wu, X., Xu, D., Duan, L., Luo, J., Jia, Y., 2013. Action recognition using multilevel features and latent structural svm. IEEE Trans. Circuits Syst. Video Technol. 23 (8), 1422–1431. 10.1109/TCSVT.2013.2244794.
Wu, C., Zhang, J., Savarese, S., Saxena, A., 2015. Watch-n-patch: unsupervised understanding of actions and relations. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4362–4370. 10.1109/CVPR.2015.7299065.
Xia, L., Chen, C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–27. 10.1109/CVPRW.2012.6239233.
Xiao, Q., Song, R., 2018. Action recognition based on hierarchical dynamic bayesian network. Multimed. Tools Appl. 77 (6), 6955–6968.
Xie, X., Lam, K.-M., 2005. Face recognition under varying illumination based on a 2d face shape model. Pattern Recognit. 38 (2), 221–230. https://doi.org/10.1016/j.patcog.2004.07.002. http://www.sciencedirect.com/science/article/pii/S0031320304002754. name="Line_manu_296".
Xu, Q., Zhou, T., Zhou, L., Wu, Z., 2016. Exploring encoding and normalization methods on probabilistic latent semantic analysis model for action recognition. In: 2016 8th International Conference on Wireless Communications Signal Processing (WCSP), pp. 1–5. https://doi.org/10.1109/WCSP.2016.7752504.
Xu, G., et al., 2007. Viewpoint insensitive action recognition using envelop shape. In: Asian Conference on Computer Vision. Springer, pp. 477–486.
Yan, Pingkun, Khan, S.M., Shah, M., 2008. Learning 4d action feature models for arbitrary view action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. https://doi.org/10.1109/CVPR.2008.4587737.
Yang, Y., Saleemi, I., Shah, M., 2013a. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. IEEE Trans. Pattern Anal. Mach. Intell. 35 (7), 1635–1648. 10.1109/TPAMI.2012.253.
Yang, Z., Zicheng, L., Hong, C., 2013b. Rgb-depth feature for 3d human activity recognition. China Communications 10 (7), 93–103. 10.1109/CC.2013.6571292.
Ye, J., Qi, G., Zhuang, N., Hu, H., Hua, K.A., 2019. Learning compact features for human activity recognition via probabilistic first-take-all. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1.
Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L., 2018. Every moment counts: dense detailed labeling of actions in complex videos. Int. J. Comput. Vis. 126 (2–4), 375–389.
Yilmaz, Alper, Shah, Mubarak, 2005a. Actions sketch: a novel action representation. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 984–989. https://doi.org/10.1109/CVPR.2005.58 vol. 1.
Yilmaz, A., Shah, M., 2005b. Recognizing human actions in videos acquired by uncalibrated moving cameras. In: Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 1, pp. 150–157. https://doi.org/10.1109/ICCV.2005.201 vol. 1.
Yu, G., Liu, Z., Yuan, J., 2014. Discriminative orderlet mining for real-time recognition of human-object interaction. In: Asian Conference on Computer Vision. Springer, pp. 50–65.
Zaidenberg, S., Bilinski, P., Brmond, F., 2014. Towards unsupervised sudden group movement discovery for video surveillance. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 388–395.
Zelnik-Manor, L., Irani, M., 2001. Event-based analysis of video. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 2. https://doi.org/10.1109/CVPR.2001.990935. II–II.
Zhang, H., Parker, L.E., 2011. 4-dimensional local spatio-temporal features for human activity recognition. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2044–2049. 10.1109/IROS.2011.6094489.
Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S., Shi, C., 2013a. Cross-view action recognition via a continuous virtual path. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2690–2697. 10.1109/CVPR.2013.347.
Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., Li, Z., 2017b. A review on human activity recognition using vision-based method. Journal of healthcare engineering 1–31. https://doi.org/10.1155/2017/3090343.
Zhang, J., Yao, B., Wang, Y., 2013b. Auto learning temporal atomic actions for activity classification. Pattern Recognit. 46 (7), 1789–1798. https://doi.org/10.1016/j.patcog.2012.10.016. http://www.sciencedirect.com/science/article/pii/S0031320312004505. name="Line_manu_267".
Zhang, Y., Zhang, Y., Swears, E., Larios, N., Wang, Z., Ji, Q., 2013c. Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35 (10), 2468–2483. 10.1109/TPAMI.2013.33.
Zhang, W., Zhu, M., Derpanis, K.G., 2013d. From actemes to action: a strongly-supervised representation for detailed action understanding. In: 2013 IEEE International Conference on Computer Vision, pp. 2248–2255. 10.1109/ICCV.2013.280.
Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C., 2016. Rgb-d-based action recognition datasets: a survey. Pattern Recognit. 60, 86–105. https://doi.org/10.1016/j.patcog.2016.05.019. http://www.sciencedirect.com/science/article/pii/S0031320316301029. name="Line_manu_176".
Zhang, J., Han, Y., Tang, J., Hu, Q., Jiang, J., 2017a. Semi-supervised image-to-video adaptation for video action recognition. IEEE Transactions on Cybernetics 47 (4), 960–973. 10.1109/TCYB.2016.2535122.
Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., Chen, D.-S., 2019. A comprehensive survey of vision-based human action recognition methods. Sensors 19 (5), 1005.
Zhao, R., Ji, Q., 2018. An adversarial hierarchical hidden markov model for human pose modeling and generation. In: Thirty Second AAAI Conference on Artificial Intelligence, pp. 2636–2643.
H. Zhao, Z. Yan, H. Wang, L. Torresani, A. Torralba, Slac: A Sparsely Labeled Dataset for Action Classification and Localization, arXiv preprint arXiv:1712.09374.
Zhou, F., De la Torre, F., 2012. Generalized time warping for multi-modal alignment of human motion. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1282–1289. 10.1109/CVPR.2012.6247812.