Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09004-3

Vision-based human activity recognition: a survey

Djamila Romaissa Beddiar[1,3] · Brahim Nini[1] · Mohammad Sabokrou[2] · Abdenour Hadid[3]

Received: 10 June 2019 / Revised: 12 March 2020 / Accepted: 1 May 2020 /
Published online: 15 August 2020

Abstract
Human activity recognition (HAR) systems attempt to automatically identify and analyze human activities using acquired information from various types of sensors. Although several extensive review papers have already been published in the general HAR topics, the growing technologies in the field as well as the multi-disciplinary nature of HAR prompt the need for constant updates in the field. In this respect, this paper attempts to review and summarize the progress of HAR systems from the computer vision perspective. Indeed, most computer vision applications such as human computer interaction, virtual reality, security, video surveillance and home monitoring are highly correlated to HAR tasks. This establishes new trend and milestone in the development cycle of HAR systems. Therefore, the current survey aims to provide the reader with an up to date analysis of vision-based HAR related literature and recent progress in the field. At the same time, it will highlight the main challenges and future directions.

Keywords Human activity recognition · Behavior understanding · Action representation · Action detection · Computer Vision · Survey

1 Introduction

Human activity recognition is often associated to the process of determining and naming activities using sensory observations [213]. More specifically, a human activity (HA) refers to the movement (s) of one or several parts of the person's body. This can be either atomic or composed of many primitive actions performed in some sequential order. Therefore, human activity recognition should allow labeling the same activity with the same label even when

✉ Djamila Romaissa Beddiar
ad_beddiar@esi.dz

[1] Research Laboratory on Computer Science's Complex Systems, Larbi Ben M'hidi University, Oum El Bouaghi, Algeria

[2] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran Province, Iran

[3] Center for Machine Vision Research, Computer Science and Engineering, University of Oulu, Oulu, Finland

performed by different persons under different conditions or styles. HAR systems attempt to automatically analyze and recognize such HAs using the acquired information from the various types of sensors [4, 7]. Besides, HAR outputs can be employed to guide subsequent decision-support systems. For instance, authors in [103] proposed an HAR system that can help a teacher to control a multi-screen and multi-touch teaching tool, such as sweeping right or left to access the previous or next slide, call the eraser tool to rub out the wrong content, among others. Similarly, the work of [223] aims at ensuring a good implementation of various Human Computer interaction systems. To this end, the underlined HAR systems are generally preceded by an activity detection task. This consists of the temporal identification and localization of such activity in the scene in a way to boost the understanding of the ongoing event. Therefore, the activity recognition task can be divided into two classes: classification and detection.

HAR has become a hot scientific topic in computer vision community. It is involved in the development of many important applications such as human computer interaction (HCI) [65], virtual reality [164], security [171], video surveillance and home monitoring [14, 138, 145, 156–161, 163, 223]. Therefore, the wide range of the activity recognition methods is directly linked to the application domain to which they are implemented [145].

In this respect, HAR systems implementation is guided by two main streams of human computer interaction technologies: (1) Contact-based and, (2) Remote methods [214]. Contact-based systems require the physical interaction of the user with the command acquisition machine or device [131]. These methods are also straightforwardly impacted by the nature of data issued from the various sensory modalities and sources, e.g., accelerometers, multi-touch screens, body-mounted sensors or wearable sensors such as data gloves to analyze the human behavior. Nevertheless, contact-based systems are more and more abandoned because the physical contact requires some skills and sophisticated equipment that make them accessible only to experimented users. Besides, in order to enable implementation in real world application scenarios, wearable sensors need as well to be easy, efficient, of adequate size, and should benefit from user's acceptability and willingness to perform continuous monitoring tasks. Among the currently developed HAR, one shall distinguish the vision-based (Remote methods). The latter attempts to simplify the human computer interaction task by allowing the human to use natural and intuitive manner in communication [214]. In fact, human activities enable a user to convey ideas or thoughts through his gestures or combination of such activities [118, 214, 223]. Intuitively, since vision-based systems use captured images or recorded video sequences to recognize activities, this can provide an edge to alternative approaches to win societal trust. Unlike contact-based activity recognition systems, vision-based systems do not require ordinary users to wear several and uncomfortable devices on different parts of their body. So, the "non-intrusive" character of these last systems allowed them to gain acceptability of use among the society.

Human activity recognition has been studied significantly in the literature. In some previous works [4, 12, 25, 32, 60, 74, 107, 143, 189, 204, 237], HAR systems have been extensively reviewed and discussed. Nevertheless the rapid development of the technology and emergence of new methodological approaches call for a constant update in the field and, thereby, new reviews in a way to benefit the growing HAR community researchers. In this respect, the current survey completes and updates the aforementioned existing surveys. Figure 1 illustrates the major aspects studied in this review. The main contributions of our survey, which make it different from the previous works are as follows:

- Our survey discusses the most significant advances reported recently in the literature covering both the general aspects of human activities recognition and the specific

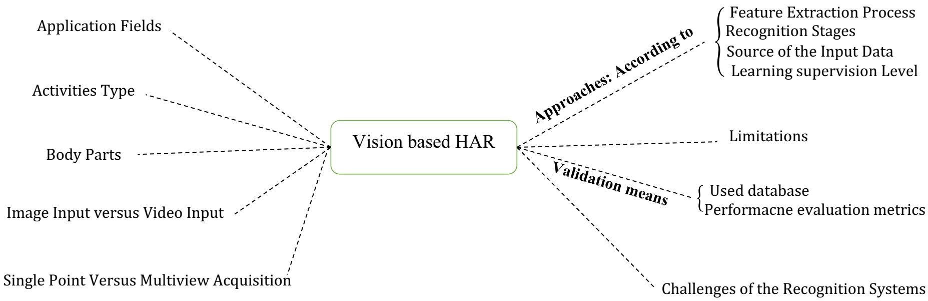
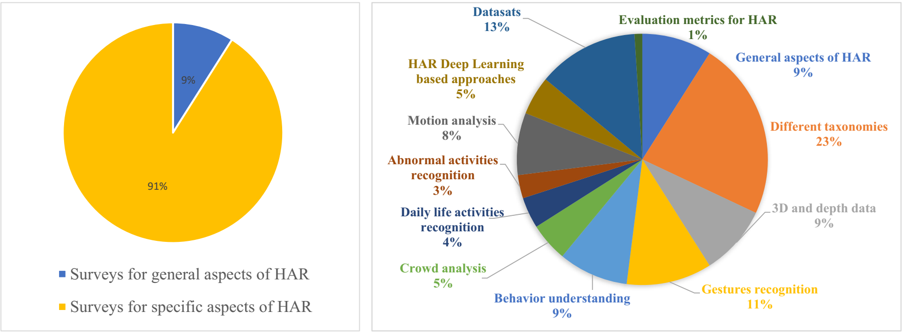

Application Fields
Activities Type
Body Parts
Image Input versus Video Input
Single Point Versus Multiview Acquisition
Vision based HAR
Approaches: According to
Feature Extraction Process
Recognition Stages
Source of the Input Data
Learning supervision Level
Limitations
Validation means
Used database
Performacne evaluation metrics
Challenges of the Recognition Systems

Fig. 1 In-depth analysis of vision-based Human activities recognition

vision-based HAR systems. One can see from Fig. 2a that only 9% of the existing surveys from 2010 to present are devoted to discuss the general framework of HAR while 91% are rather presenting specific taxonomies or domain specific. In addition, Fig. 2b shows most commonly discussed HAR subjects and the percentage of surveys covering each subject for the ten past years.

- This paper presents a deeper analysis of human activities recognition by discussing various applications of HAR in different fields, analyzing the proposed approaches, defining abstraction levels of activities and categorizing human action representation methods. In an attempt to quantify the in-depth characterization of the HAR survey papers reported in our study (this paper), we manually scrutinize the reported survey papers for the last ten years to distinguish in-depth and comprehensive surveys from domain-specific or light surveys. The results are presented in Fig. 3. It can be seen from this figure that there are only few surveys similar to our work and, thereby, there is a need for updated new comprehensive review for HAR systems. Strictly speaking, the previous quantification (of in depth-survey versus standard-survey) is rather based on the number of technologies and methodologies, structure of taxonomy, in-depth comparative analysis undertaken by the underlined review paper.



(a) (b)

Fig. 2 HAR related surveys from 2010 until 2019: **a** The percentage of surveys representing general aspects of HAR compared to surveys presenting specific taxonomies and application domains of HAR, **b** Distribution of major subjects of HAR covered by recent surveys

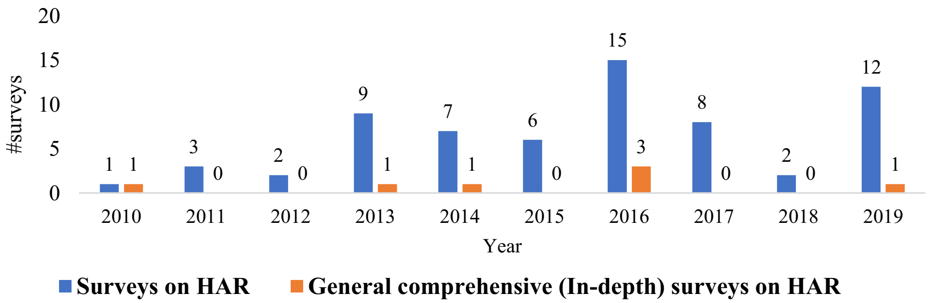


Fig. 3 Comparison between the number of HAR surveys covered by our study and the number of surveys presenting general and comprehensive analysis of HAR for the period of 2010 to present

- In addition to reviewing existing human activities recognition approaches and common related datasets, we classify them according to the modalities used when acquiring data and to the commonly employed three-stage activity recognition process: detection, tracking and classification.
- We also identify challenging issues and provide useful recommendations to provide useful insights to HAR system development community.

We explore in this paper, the state-of-the-art of HAR methods from different point of views, and categorize the general HAR task according to several criteria which are discussed in the following sections. In Section 2, we provide an overview of survey papers on HAR tasks. In Section 3, we detail the main applications of HAR systems, ranging from simple desktop applications to advanced robotic. Subsequently, we provide in Section 4 four taxonomies for classification of HAR approaches according to; (1) the feature extraction process with a categorization of human activity representation methods in Section 4.1, (2) the three-stages recognition system followed by the methods used in implementing each of them in Section 4.2, (3) the input data modalities in Section 4.3 and (4) the supervision level of the machine learning in Section 4.4. Afterwards, we describe in Section 5 a taxonomy of activity types and the body parts which can be used to identify the various human activities. We also classify the works in Sections 6 and 7 according to two criteria: (1) the nature of processed data, and (2) the data acquisition modality. We explore in Section 8 the used means to validate different approaches. We enumerate the standard and popular benchmarks and evaluation metrics for HAR task. Next, the limitations of HAR methods and the challenges associated to enhancing HAR system performances, in Sect. 9 and Section 10 respectively. Finally, we finish with a conclusion.

2 Related surveys

The state-of-the-art methods of HAR task are studied and surveyed in different papers [3, 4, 12, 25, 32, 60, 74, 75, 77, 81, 107, 136, 143, 189, 195, 201, 204, 213, 233, 237]. These survey papers introduce the recent advances in automated human activity recognition topic. A few number of these works such as [12, 74, 136, 143, 189, 201, 231] provide a comprehensive review on different aspects of HAR methods, while most of them look at HAR task from a specific point of view. For example, [213] classifies HAR approaches according to the spatial-temporal characteristics of actions, video segmentation and recognition systems and camera modalities. Likewise, [3] advocates a taxonomy-based approach and compares the advantages and limitations of each method. The authors examined both simple human actions and high-level activities. In addition, the authors of [81] review HAR approaches according to the complexity of features involved in the action recognition process. Similarly, [195] categorizes human activities according to their complexity into action and activity and then reviews the major approaches for recognizing human actions and activities. Authors of [60] summarize existing methods for human activity recognition from still images and categorize them into two big categories according to the level of abstraction and the type of features each method uses. Otherwise, [25, 77] compare techniques related to image segmentation, feature extraction and activity classification, and then discuss advantages and limitations of each of them. Furthermore, the authors of [204] categorize HAR approaches into two broad categories: uni-modal and multi-modal with regards to the source channel each of these approaches employs for human activities recognition. They also reviewed the existing publicly available human activity datasets and examined the requirements for building both ideal HAR dataset and system.

On the other hand, [4] focused on techniques that use 3D and depth data, while [63, 140, 230] surveyed 3D skeleton-based human representation and action recognition approaches. Interestingly, [237] represents the semantic-based human recognition methods using still images and videos. It identifies semantic space and semantic-based features such as pose, poselet, related objects, attributes, and scene context. It also briefly discusses the potential applications of semantic approaches. [107, 233] provide a classification of common Kinect-based motion recognition approaches and review each approach accordingly. They highlight Microsoft Kinect sensor applications in various domains as well as the publicly available Kinect datasets.

Overviews of current progress in human activities recognition are also presented in [7, 21, 122, 174]. These surveys analyze popular techniques used for object segmentation and activity recognition. The authors discussed as well the merits and demerits of such methods and proposed possible future scopes in this area of research. Authors in [129] were interested into knowledge-based human activity recognition methodologies which are not well covered in the literature. More specifically, they survey methods and techniques used in the literature to represent and integrate knowledge and reasoning into the recognition process. These methods are categorized in terms of statistical, syntactic and description-based approaches.

On the other hand, reviews of vision-based hand gesture recognition methods were presented in [33, 34, 45, 64, 119, 145, 153, 227]. The purpose of these reviews was to introduce the field of gesture recognition as a mechanism for interacting with computers and provide researchers with a summary of advances in hand gesture recognition to help identify areas where further research is needed. Authors in [145] focus on gesture taxonomies, their representations, recognition techniques, software platforms and frameworks as well as hand gesture recognition applications. Other researchers are interested in behavior and event understanding rather than actions and gestures. We can quote as examples [8, 22, 69, 135, 151, 193, 209]. Authors in [22, 135] discuss the advantages and the drawbacks of existing approaches for behavior understanding as well as related available datasets. They also highlight open research challenges and several important future directions. [8] examines complex event recognition techniques. They identify a number of limitations with respect to the employed languages, probabilistic models and their performance as compared to the purely deterministic cases. Based on those limitations, promising directions for future work are then highlighted. Authors in [193] attempt to summarize techniques in understanding activities of a person and/or a group as well as their social interactions. For instance, understanding crowd behavior is deemed essential to surveillance and security purposes. Motivated by this fact, many surveys are devoted to crowd analysis such as [1, 23, 59, 186], in order to provide an overview of the major techniques applicable for classifying abnormal behavior in a crowded scene scenario. On the other hand, reviews on vision-based Ambient Assisted Living, patient monitoring like [123, 127, 139], fall detection and abnormal human activity recognition such as [18, 39] were proposed in the literature. Other researchers were interested in motion analysis to recognize human activities and many reviews were presented in this field. We can mention for instance [2, 31, 47, 94, 116, 117]. Authors in [31] attempt to summarize human motion analysis algorithms that use depth imagery. [2] discusses three sub-topics of human motion analysis: human body parts-based motion analysis, moving human tracking and image sequences-based human recognition. [116] views human body motion as a hierarchical process with four steps: initialization, tracking, pose estimation and recognition. A comparative analysis of methods based on handcrafted representations and solutions that involve learning architectures is carried out in [68] and [236]. In both surveys, the authors discuss recent advancement in human action representations alongside the associated pros and cons. Reviews on deep-learning based methods of human activities recognition were provided in [16, 132, 210, 225]. They analyzed the advantages and limitations of current existing techniques and discussed the potential directions for future research. Authors in [210] were interested particularly to RGB-D-based human motion recognition using deep learning architecture and focused on three architectures of neural networks: CNN, RNN and other structured networks. On the other hand, the authors of [132] enumerate a list of datasets in different complexity levels and compare the performance of deep learning-based approaches to other existing works.

Surveys on dataset benchmarks for human action recognition from visual data constitute another field of research tackled in [28, 44, 61, 66, 98, 181]. They aim to guide researchers in the selection of the most suitable dataset for benchmarking their algorithms. These surveys also present the best performance scores achieved by various HAR methods on these benchmark dataset. The performance analysis includes the number of activity classes, complexity of events, application domain and impact of the ground truth. In addition, [66] presents a summary of the results obtained on the recent ASLAN benchmark [79], which was designed to reflect on the variety of challenges that modern activity recognition systems are expected to overcome. Authors in [44] propose a novel dataset, called CONVERSE, that represents complex conversational interactions between two individuals via 3D pose. Similarly, authors in [20, 46, 180, 181, 232] present a set of comprehensive reviews of the most commonly used RGB-D video-based activity recognition datasets. Relevant information in each category is extracted in order to help researchers to easily choose appropriate data for their needs. Moreover, the reviews highlight the evaluation protocols, and the limitations of the publicly available datasets. A guidance on future creation of datasets and establishment of standard evaluation protocols for specific purposes is also provided.

Authors in [114] examine another aspect of human activity recognition by analyzing several factors that influence the evaluation of activity recognition approaches. Especially, they reviewed many of the commonly used metrics, outlined the sources of errors in such systems and presented different methods for detecting and labeling these errors.

However, all these surveys discussed above do not cover all aspects of human activity recognition. They highlight different taxonomies, applications and specific theoretical angle of HAR (See Table 2 in the Appendix section for a comparative analysis between our survey and the surveys mentioned above). This motivates the current paper where a deeper analysis of human activity recognition is provided, although the review pays a special attention to vision-based HAR systems. We analyze approaches proposed in the literature. We define the activity and its related abstraction levels and categorize human action representation methods. Furthermore, we determine the corresponding human body parts that are tracked to identify activities and classify HAR methods according to the acquisition device and to the three basic stages. In addition, we discuss the nature of the processed data and the point of view. We enumerate means used for the evaluation of approaches, datasets of human activities and performance evaluation metrics. Finally, we point out limitations and open research challenges that require special attention from the HAR research community.

3 HAR applications

HAR has been widely used for various applications such as human-computer interaction systems HCI, computer vision and augmented reality [14, 19, 53, 65, 138, 222, 223]. It is important to stress the interest of having interactive and natural interfaces, where the user can use his performed actions to provide instructions to the machine. For instance, a speaker can control the presentation of the slides with the movements of his hand [24]. Likewise, the recognition of human activity from static images or video sequences has several potential applications in many fields. Examples of human activity recognition applications include monitoring and evaluation of processes in industry as well as machines and devices control [27, 202, 223], fraud detection [7, 21], extraction of information from videos [4, 198], video assistance and surveillance [4, 7, 19, 21, 27, 198] and public security [171] where crowds' movements are tracked to detect violent or criminal situations. Given the increasing involvement of robots in our life, it is essential to equip them with the ability to understand intentions, emotions and behaviors of individuals. Thus, Robotics and video games have also taken benefits from the progress made in HAR systems [4, 7, 19, 21, 104]. Finally, other applications of HAR systems touch in medical environments to ensure surgical operations or patient monitoring, interpretation of language signs [7, 21, 68, 73, 119, 177] as well as supervision of medication [19]. For instance, a combination feature extraction method based on human activities recognition is introduced in [73] making possible to classify static signs of the sign language.

The interest in the development of these human activity-based applications can be justified by the fact that they provide very valuable and useful means of communication. However, the progress of the research in this field is also affected by the considerable changes in the technology trend and overall ecosystems.

4 HAR approaches

HAR methods are composed of three important components: (1) Video frame segmentation for action detection, (2) Action representation with respect to posture and motion of the human body, and (3) Learning process that recognizes these actions. We categorize in this section, HAR approaches according to feature extraction process in the first subsection, to the three stages of recognition process in the second subsection, to the source modalities of input data in the third subsection and finally to machine learning supervision level in the fourth subsection.

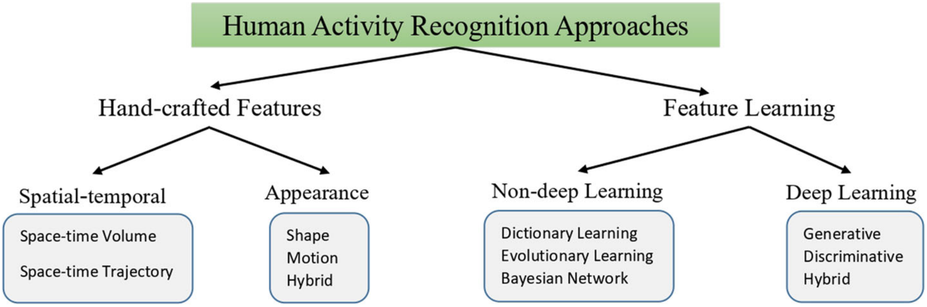


Fig. 4 Vision-based Human activities recognition approaches

4.1 HAR approaches according to feature extraction process

In this subsection, we present a classification of HAR approaches according to feature extraction process into handcrafted representation based and learning based approaches. Figure 4 summarizes the HAR methods as follows:

Methods based on Handcrafted features rely on human ingenuity and prior knowledge to extract discriminating features. These types of methods involve three major steps: (1) Foreground detection that corresponds to action segmentation, (2) Feature selection and extracting by an expert and (3) Classification of action represented by the extracted features [24]. The input images or video frames are analyzed to extract the most significant features that are used, then to build the descriptor. The classification is performed using a generic trained classifier which makes this family of approaches low cost, flexible and does not rely on large sets of samples for training. Methods based on handcrafted features are: spatial-temporal-based approaches as discussed below, appearance-based approaches, and other methods like local binary pattern LBP which is a visual descriptor used for texture classification and fuzzy logic.

Human activities can be seen either static or dynamic. Static activities are described using the orientation and the position of the limb in space while dynamic activities are described as movements of these static activities [145]. Hence, action recognition can be based either on spatial or temporal cues that describe and recognize these actions [213]. Both spatial and temporal representations can be categorized into three classes (see Fig. 5).

i) *Spatial representations*: This can also be categorized into three subclasses.

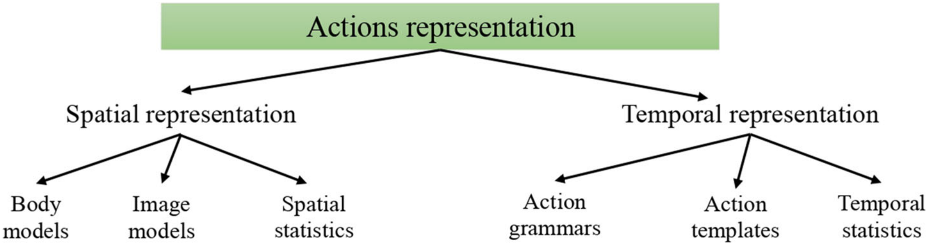


Fig. 5 Spatial and temporal representations of actions

(a) *Body models*: This allows recovering human body pose from features using the spatial structure of the action with reference to the human body. One potential method is the reconstruction of the 3D body model that aims at representing the body as a kinematic joint model [54, 182] and recognizing the underlined action using joint trajectories. An alternative method is the direct recognition from 2D models without employing 3D models.

(b) *Image models*: This corresponds to a holistic representation of actions that use a regular grid bounded by a region of interest centred around the person to detect and compute features. For instance, we can quote silhouettes [13, 56, 72], contours [72], motion history images (MHI) [6, 70], motion energy images (MEI) [6, 70] and optical flow [71, 93, 165, 190] that are used to describe actions and movements.

(c) *Spatial statistics*: This enables us to represent local actions through a set of statistics of local features from the surrounding regions. These regions are obtained after either dense or sparse decomposition of the underlined image or video. For instance, we can mention statistical methods and space-time interest points [6, 108, 142] which calculate Spatial-Temporal Interest Points (STIP) of the image and assign each region to a set of features, to provide spatial distribution of local image features. Methods based on spatial-temporal can be classified as follows:

□ *Volume-based*: These approaches represent video as spatial-temporal volume and may rely on features like texture, color, posture, histograms of optical flow, histograms of oriented gradients...etc. Actions are recognized using similarity between the two volumes. Volume-based approaches can not work efficiently when the scene is crowded, they are only suitable for simple action or gesture recognition. For example, Scale Invariant Feature Transformation (SIFT) is used as a 2D interest point detector in [105] while corners and Laplacian of Gaussian are used in [42] to detect 3D interest points. Figure 6a shows a representation of the human body with space-time volumes based methods.

□ *Trajectory based approaches*: This family of approaches represent joint positions of the body with 2D or 3D points which are further tracked along the video to compute action trajectories. These tracked changes in the posture are used to construct the 3D representation of the activity which is considered to be a set of spatial-temporal trajectories. These methods are powerful against noise, view and/or illumination changes, and are useful for recognizing complex activities. For instance, [205] calculates several descriptors (HOG, HOF and motion boundary histogram MBH) and trajectories by tracking densely sampled points using optical flow. Figure 6b shows a representation of the human body using spatial-temporal trajectories based methods.

ii) *Temporal representations*:

(a) *Action grammars*: they represent the action as a sequence of moments. Each moment is described by its own appearance and dynamics. To perform action recognition task, features are grouped into similar configurations called states and temporal transition between these states are learned [134]. Hidden Markov Models [82, 83], CRF [113, 179], regressive models [49, 115, 130] and context-free grammars [40, 149] are examples of action grammars.

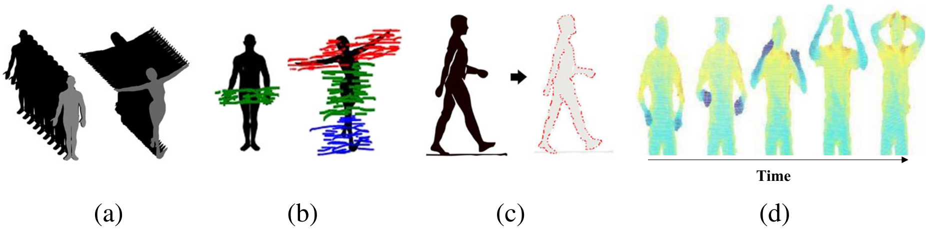


Fig. 6 Examples of handcrafted feature extraction approaches: **a** Space-time volumes, **b** Space-time trajectories, **c** Shape-based methods: Contour features, **d** Motion-based methods [141]

(b) *Action templates*: This consists of a set of temporal representations that allow representing the appearance of temporal blocks of features and dynamics called templates [70, 208]. Templates are computed over long sequences of frames and dynamics are represented using features of several frames loaded in feature vector or in space time volume. Typically, Fourier Transform [198, 199, 207], Wavelet representations [58, 229] and trajectories of body parts [176, 197] are templates that can be used for action recognition.

(c) *Temporal statistics*: They are based on statistical models and are as well temporal representation of action. Statistical models [173] are used to describe the distributions of unstructured features over time. These features represent the appearance of actions.

iii) *Appearance based approaches*: They use 2D or 3D depth images and are based on shape features, motion features or any combination of both features. These methods have been significantly simplified due to the emergence of depth cameras. A quick advance is felt in the skeleton-based recognition approaches as well with the advent of depth sensors and algorithms of real-time skeleton estimation [177]. According to [198], they can be classified into two categories:

□ Joint locations, which consider the skeleton as a set of points.

□ Joint angles, which assume the human body as a system of rigid connected segments and the movement as an evolution of their spatial configuration [7, 29, 128].

As an appearance based approach, [198] proposes an efficient skeleton representation for activity recognition based on the body parts. The authors model the 3D geometric relationships between the body parts using rotations and translations. The work proposed in [234] is based on the Microsoft Kinect sensor and presents a new method of HAR while using the machine training. Actually, the growing interest in the recognition, generated by the release of Kinect, is due to the fact that the skeleton information can be deduced from depth images [124]. A transparent standardized input interface, making it possible to reduce HCI constraints was proposed in [118]. It is used to imitate in real time the motions carried out by the user, using an articulated 3D model. The system is based on the image analysis to detect the trajectory of the hand, to determine its configurations and interpret them like 3D postures. This solution is very fast and computationally efficient, and can be used as a control entry of a mobile robot.

Generally, the appearance based approaches can be classified according to either shape or motion based characteristics.

(a) *Shape based methods*: They capture local shape features such as contour points, local region, silhouette and geometric features from the human image or video using foreground segmentation.For example, [223] proposes a process for hand gesture recognition in the 3D point cloud, which explicitly uses 3D information of depth maps where the color information is ignored. It submits a standardized descriptor of features, making it possible to effectively represent the various positions of the hand. Figure 6c shows a representation of the human body using contour points.

(b) *Motion based methods*: They include optical flow and motion history volume, which are then used for action representation [71, 190, 212]. Then, a generic classifier on top of this representation is implemented for action recognition. The research of [191] proposes to recognize actions using vector quantization of motion descriptors. This method is related to a meta-algorithm combining the histograms of optical flow and classifiers of bag-of-words. An example of action recognition proposed by [141] is given in Fig. 6d.

(c) *Hybrid methods*: they combine shape and motion features to represent actions [5, 72, 85, 96, 172].

Recently, feature learning [97, 196] has started to become popular in different computer vision applications such as pedestrian detection [26, 87], image classification [11, 88], vision-based anomaly detection [35], ...etc. Besides, many of the learning-based action recognition approaches rely on the end-to-end learning which consists of transformations from pixel-level to action classes. The feature-learning based methods for the HAR task can further be grouped into (1) Traditional and, (2) Deep-learning-based methods.

i) *Traditional approaches*: This includes methods like genetic programming [36, 100], dictionary learning [110, 226] and Bayesian networks [92, 101].

(a) *Dictionary learning*: This provides a sparse representation of the input data by a linear combination of basis dictionary atoms. They use an end-to-end unsupervised learning to learn the dictionary and the corresponding classifier within a single learning procedure. The concept of these methods is similar to visual Bag of Words model (BoW) that generates global data representations. For instance, [235] presents a weakly-supervised cross-domain dictionary learning approach to adapt knowledge of one action dataset to another action dataset. The proposed method learns a reconstructive, discriminating and domain-adaptive dictionary pair and the corresponding classifier parameters without using any prior information. As aforementioned, methods based on spatial-temporal features are very popular, and have achieved state-of-the-art results on activities classification task. However, the use of over-complete dictionaries proves to be more interesting because they can produce even more compact representations. Therefore, the authors of [222] have combined these two concepts to propose a solution of HAR from multi-part missing video.

(b) *Genetic programming (GP)*: GP is an evolutionary method inspired by the process of biological evolution and its fundamental mechanism [24]. The principle of genetic programming is to search a space of possible solutions without having any prior knowledge and it allows discovering functional relationships between features in data enabling its classification. Genetic programming has been used to construct holistic descriptors that allow to maximize the performance of action recognition tasks [100]. In [100], the authors developed an adaptive learning methodology using GP to evolve discriminating spatial-temporal representations, which simultaneously fuse the color and motion information, for high-level action recognition tasks.

(c) *Bayesian networks*: These methods are probabilistic graphical models that infer the conditional dependencies using directed acyclic graphs [24]. The graph nodes and edges represent the random variables and their associated conditional dependencies, respectively. The probability computations are performed using the Bayesian inference. In [92], the authors adopted a Bayesian Network (BN) to represent and capture the semantic relationships among action units, as well as the correlations of the action unit intensities, to more accurately and robustly measure the intensity of spontaneous facial actions.

ii) *Deep-learning*: Feature learning approaches based on deep-learning methods are widely explored for HAR task because of their promising performance, robustness in extracting features and their generalization ability for different types of data. These methods are very data harvesting in the training process. They aim to learn multiple levels of representation and abstraction that allow a fully automated feature extraction process. Deep learning-based methods can be considered as trainable feature extractors, which allow the recognition of high-level activities with complex structures. However, high computational complexity and huge data requirements for the training phase are still among ongoing challenging problems. Deep learning-based approaches can be categorized into:

(a) *Generative methods*: These are unsupervised models which are based on the famous quote of Richard Feynman: “What I cannot create, I do not understand”. [55]. They use unsupervised learning to represent any kind of unlabeled data distribution. The new representation reduces the data dimensionality and comply from the data distribution. Therefore, the main aim of the generative models is to understand the data distribution including the features that belong to each class, in order to replicate the initial true data distribution of the training set. The most commonly used and efficient approaches are: Auto-encoder [187, 221], Variational Autoencoders (VAE) [41] and Generative Adversarial Networks (GAN) [146]. As an example of generative models, [162] proposes an end-to-end deep learning model for abnormal activity recognition in videos. The proposed architecture is similar to GAN, where the two networks compete to learn and collaborate in the detection task.

(b) *Discriminative methods*: these are supervised models that use a hierarchical learning model composed of different hidden layers to classify raw data input into various output categories. The most used are Deep Neural Networks (DNN), Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). As an example, in [197], the authors aim to demonstrate the advantages of long-term temporal convolutions and the importance of high-quality optical flow estimation for learning accurate video representations for human action recognition. For that, they investigate the learning of long-term video representations. They consider space-time convolutional neural networks and study architectures with Long-term Temporal Convolutions (LTC).

(c) *Hybrid models*: these methods integrate generative and discriminative models to gain performances and advantages of both models such as [112, 187].

4.2 HAR approaches according to the recognition stages

Human activities recognition systems are, in general, composed of three main steps: Detection which consists in determining the part of the body to follow or to recognize; Tracking that provides a connection between the successive images, and; Recognition which consists in interpreting the semantics of the localization, the posture, and the activity. In order to investigate further HAR methods, this subsection is devoted to represent different state of the art vision-based HAR techniques associated to each stage. This assumes that the choice of the algorithm depends on the selected representation of activities. We classify methods for human activities recognition into the following.

i) *First stage methods (detection)*:

  (a) *Skin color*: This can be used to detect the desired body part. The problems of skin color-based methods reside mainly in choosing the relevant color space and with false detections because of the objects of the scene whose color is close to that of the skin. Very common solutions to these problems are the separation of brightness and chromaticity components, elimination of shadow effect and background subtraction [37, 164, 194]. Other problems may influence the results of these methods, like the inherent discrepancy of skin color variations according to ethnicity group, lighting conditions and the sensitivity of the employed acquisition devices.

  (b) *Shape*: The contours of the body part shape can be extracted to follow and recognize human activities. To increase the reliability of these techniques, sophisticated approaches of post-processing are used for the extraction. Shape-based methods are typically independent of the camera view, skin color and conditions of lighting but present the difficulty of classifier's implementation. Moreover, background objects can be confusing or can occlude the forms to be detected. For instance, [234] uses the skeleton shape compared to information of the skeleton embarked in MICROSOFT SDK in order to detect human actions. In [223], the detection is based on appearance through a combination of the color and the shape to ensure a region segmentation that contains the person's hand. Shape-based methods were also used by [15, 214].

  (c) *Pixel values*: The appearance can be expressed in terms of pixel values change between images of a sequence according to the activity. Indeed, it is evident that the difference in appearance between various activities is more noticeable than among people performing the same activity. So, techniques based on the appearance of body parts are proposed as in [203].

  (d) *3D models*: They attempt to build matches between characteristics of the model based on various features of the images. These techniques are independent of the viewpoint, but present some limitations in terms of accurate positioning characteristics. For instance, [50] uses a 3D model method for hand gesture tracking.

  (e) *Motion*: Motion can be detected using the difference in brightness of pixels of two successive images. For example, authors in [118, 183] use several motion characteristics based methods to detect body parts.

  (f) *Anisotropic-diffusion*: This is based on the extension of the successful anisotropic diffusion based segmentation to the whole video sequence to improve the detection of object contours and regions of interest that may trigger specific activity, see [102, 219]. Therefore such concept can be used to detect and describe activity patterns. For instance, the work in [95] uses anisotropic diffusion based approach to enable the recognition of crowd activities using a two-step clustering scheme that extracts the semantic regions and the coherent motion. In [102], an anisotropic filter is used for noise discrimination of 3D human activity recognition.

ii) *Second stage methods (Tracking)*:

Various methods, independently from the ones used for the detection phase, can be used to track the human body. If the detection method is fast enough to operate at the frame acquisition rate, it may be the same in both stages. Tracking methods can also be categorized into the following.

(a) *Template-based methods*: these techniques use models to follow the body parts [15]. They require continuous learning with high frame rate and are classified into two categories:

□ *Features tracking based on the correlation*: Regions of an image containing the body parts to be followed are used as a prototype to be detected in the following image. These techniques require that the part being tracked remains in the same neighbourhood in the successive images. They are influenced by the lighting variations and may not be efficient in real time systems. In [223], the authors use 3D information of depth maps to ensure the follow-up of the human body. Optical flow characteristics of two successive images are employed in [191] using the algorithm of Kanade Lucas Tomasi [106, 175]. In [118], the hand gestures are followed using an articulated 3D model.

□ *Contours based tracking*: They are deformable contour techniques (snakes) which initially place a contour close to the area of interest, then it is warped in an iterative way using active shape models in each frame to make the snake converge [217]. These techniques are more efficient if a contrast between the object to be tracked and the background exists. They can be used in real-time systems, may handle several targets at the same time and can also be adapted to complex postures. On the other hand, they are influenced by color intensity variations and are limited regarding smoothing and softening of contours.

(b) *Optimal estimation*: These methods evaluate the state of moving systems from series of measures [9, 218]. In the case of HAR systems, they are used to estimate the movements of the human body in similar way that Kalman Filter does [91]. They can be used in real-time systems, and can deal with uncertainty, but have limitations against cluttered backgrounds.

(c) *Particle Filters*: These methods consist in following the positions of the body parts and their configuration in complex environments. A particular location of the part of interest is modelled by a set of particles.

(d) *Cam Shift*: They are based on the mean shift algorithm which use the models of appearance based on density to represent the targets [106, 175]. Such methods consist in finding, in a sequence of images, the nearest model of distribution to the sample model using an iterative research. These methods are simple and have low-cost calculation. They have some limitations in complex scenes or scale variations.

iii) *Third stage methods (classification)*:

(a) *Support Vector Machine (SVM)*: The classification is performed by the construction of a set of hyper planes in a multidimensional space separating the elements from various classes with a vast margin. SVM classifier is the most used handcrafted classifier because it offers a very high rate of recognition performance and classification. It was used with a Gaussian, Radial basis function (RBF) or linear kernel with or without parameters in [21, 27, 124, 191, 222, 223].

(b) *Naive Bayesian classifier*: This is a probabilistic classifier based on the theorem of Bayes [10]. It consists in counting the number of occurrences of the key motion in a video sequence. To recognize a new action, the rule of Bayes is applied and the selection of the activity which has the greatest probability a posteriori is done. In [191], a comparison between this classifier and SVM is carried out.

(c) *Algorithm of K_Nearest Neighbor*: it consists in the classification of objects based on the majority of their neighbors' vote [21, 222]. To identify neighbours, objects are represented by position vectors in the multidimensional space of features. This algorithm is sensitive to local data structure. The work in [198] compares the obtained classification results using this algorithm and SVM.

(d) *K_means*: This is a clustering algorithm, which is sensitive to data structures and consists to iteratively calculate the k-distances from each class centroid to each datum [222]. The points are then assigned to the nearest cluster and the centers are re-evaluated to be the average of their class values. The process is repeated until the stop condition is reached, which can be the maximum number of iterations or a tolerance level. It is used by [191] for the classification of human actions.

(e) *Mean shift clustering*: These are non-configurable techniques of clustering that do not require prior knowledge about the number of clusters and do not limit their forms.

(f) *Machines finite state*: In this model, the static gestures and postures are represented by states, authorized changes where temporal and/or probabilistic constraints are represented by transitions, and finally the dynamic gestures are represented by arcs between the initial and the final states. These machine-finite states require the modification of the model whenever a new gesture appears, and have a high computational complexity because they are proportional to the used gestures.

(j) *Hidden Markov Models*: These models represent solutions to the segmentation problem and are among the most popular [7, 19, 21]. They constitute a generalization of Markov chains which are finite state automaton with a probability value on each arc. They have been discussed in [204].

(h) *Dynamic time warping*: This algorithm calculates the distances between each pair of possible points from two signals according to their associated characteristic values. This allows estimating and detecting the movement in a video sequence. It was used by authors of [124] in order to eliminate the problem of rate variations generated by the classification of human actions.

(i) *Neural networks*: They are mathematical models whose design is inspired from the functioning of biological neurons [19, 21]. They are generally optimized by probabilistic learning techniques, in particular, Bayesian. Neural networks allow creating fast classifications which can be applied to real-time systems. For instance, [78, 152, 200] use neural networks for activities recognition.

4.3 HAR according to the source of the input data

Methods of human activities recognition are categorized, in [127], with regard to the nature of the detector into two main categories: Uni-modal methods that consider an activity as a set of visual characteristics and allow the recognition from single modality data [109, 147]; Multi-modal methods, which combine collected features from various sources, and therefore several modalities are used [133, 168]. We can also mention the work of [27], where the authors combine speech and localization of human body for HAR.

Among the uni-modal methods, we can mention the space-time methods, which concatenate the time and the 3D representation of the body to locate activities in space, allowing a detailed analysis of human movements [80, 128]. Often sensitive to noise and occlusion, these methods are not adapted to recognize complex actions, such as in [191, 198]. On the other hand, the stochastic methods, which represent the activity by stochastic models like Markov Model, enable the modelling of human interactions and the recognition of complex activities. These methods yield approximate solutions whose training is difficult because of the large number of training parameters [82, 83]. We can also mention rule-based methods that characterize the activity using a set of rules or attributes [19, 211]. They present some difficulties during the generation of these rules and attributes, in the analysis of long video sequences and during the recognition of complex activities. Lastly, another family, shape-based methods has emerged. They use the shape features to represent and recognize activities [118, 191, 216, 223]. The existence of a great number of pose estimation devices at low cost makes these techniques very useful, although they depend on the viewpoint, occlusion, people clothing and are sensitive to lighting variations.

Among multi-modal methods, there are emotional methods that associate visual and textual features to classify the emotional states of individuals in static images [166, 171]. On the other hand, the behavioral methods aim to recognize the behavioral attributes like the emotions, mood, etc. These methods allow the recognition of complex human activities using complex classification models, which make the specification of emotional attributes difficult [144, 228]. Finally, we shall mention the methods based on social networks, which allow recognition of social events (wedding, birth...), and interactions. They are limited by the number of people in interaction and the modelling difficulty in complex scenes [43, 148].

4.4 HAR approaches according to the machine learning supervision level

HAR is based on machine learning approaches and can be further categorized according to the level of learning supervision into three sub-classes; supervised, unsupervised and semi-supervised methods.

Supervised methods: In supervised learning, the training takes place offline and is performed by the machine using the well labeled data in order to predict outcomes of unforeseen data. For HAR, supervised methods are used to classify and recognize short term actions. Methods of this sub-class are computationally simple, highly accurate and trustworthy. We can quote as examples of supervised learning: Support vector machine, Linear and logistics regression, random forest and neural networks. Works of [149, 161, 235] are examples of supervised approaches for HAR.

Unsupervised methods: This refers to machine learning techniques that do not need any supervision mechanism. These techniques allow the model to discover by itself the discriminative features of the input unlabelled data using a real-time learning such as K-means and K-nearest-neighbor. For HAR, unsupervised methods perform well for finding spatio-temporal patterns of motion and generating scene models in order to automatically localize activities. These methods are computationally complex, less accurate and trustworthy. For instance, [215] presents a new algorithm that models the human activities in a completely unsupervised setting enabling to recognize daily living and forgotten activities. The probabilistic model considers, both the short-range and the long-range action relations and shows considerable results in action segmentation and clustering. Similarly, unsupervised approaches are presented in [26, 109, 187]. Authors of [187] present a multi-layer Long Short Term Memory (LSTM) networks to learn representations of video sequences, while an auto-annotation framework to iteratively label pedestrian instances using multi-modal data is introduced in [26].

Semi-supervised methods: This refers to hybrid methods which combine supervised and unsupervised learning. The training is performed using both labeled and unlabeled data, mostly when the labeled data is not enough to product accurate model. For HAR, these methods can benefit from discriminative power of supervised approaches to distinguish between features and the ability of unsupervised methods to automatically localize actions. For instance, [120] presents a hybrid framework for online recognition of daily living activities. The authors provide a complete presentation of human activities by exploiting both unsupervised (to represent global motion patterns and localize activities) and supervised learning approaches (to distinguish between actions occurring under specific scene region). Other works such as [6, 115, 211] provide also semi or weakly supervised framework for HAR in videos.

5 Activities type

Human activities are regarded to be the means of communication between individuals, interactions with machines and with the environment in which we live. As mentioned earlier in this survey, activities refer to body parts or whole body movements and is composed of several elementary actions performed in a temporal sequential order. They can be accomplished by one person or a group of people. We present in this section, a hierarchy of human activities depending on their complexity scaling from simple action to more complex events. Figure 7 shows this hierarchy.

i) *Elementary human actions*: This consists of simple atomic activities which indicate voluntary and/or intentional body movements that form the basis for building other more complex actions such as “raising the left hand” or “walking”. They are easy to

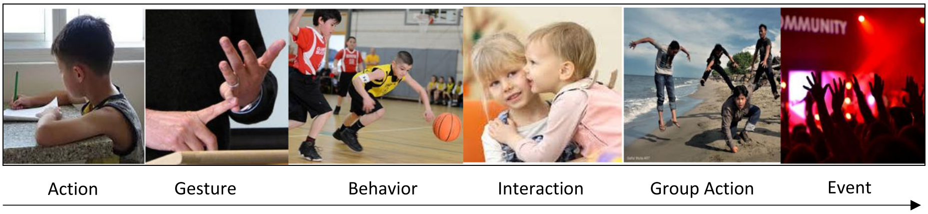


Fig. 7 Human activity types scaling from simple action to event

recognize and have been a center of interest of several research tasks like [21, 27, 124, 198].

ii) *Gestures*: Typically, a gesture is a language or part of the non-verbal communication which can be employed to express significant ideas or orders. The gestures are a second type of activities which may be conscious like "applauding", and unconscious like "hiding the face with hands when getting shy" or "pulling out the hand when touching a hot material". Some gestures are universal, whereas others are related to quite specific social and cultural contexts. Among the works that are interested in gestures recognition, we can mention [21, 118, 214, 223].

iii) *Behaviors*: These describe the set of physical actions and reactions of individuals in specific situations that are observable from the outside and which are relative to their emotions and psychological states. Proposals in [19, 21, 171] are examples of approaches that attempt to recognize human behaviors.

iv) *Interactions*: These are reciprocal actions or exchanges between two entities or more, which modify the behavior of individuals or objects involved in the interaction. They are, in general, complex activities of two types: human to human such as "kissing" or human to object such as "cooking" which involves various kitchen utensils. The works presented in [118, 178] focus on the recognition of interactions.

v) *Group actions*: constitute the activities carried out by a group of people like "cuddling". These activities are more or less complex and difficult to track or recognize. The approaches suggested in [21, 118] make it possible to recognize complex activities.

vi) *Events*: These are human activities taking place in a specific environment or high-level activities which represent social actions between individuals such as "weddings and parties" [148, 154].

Next, we introduce here another particular type of activities related daily living (ADL), because it has gained a particular attention in the computer vision community due to its importance in surveillance, assistance and patient monitoring environments. Understanding such activities is important because, in addition to providing information on the person's autonomy and ability to independent living, it provides personal safety to older people. ADL can be defined as activities that people are used to do every day without any assistance. In general, they are complex activities, composed of many simple activities performed in an indoor environment. Among research interested to activities of daily living, we can quote [19, 120, 123, 133, 202].

5.1 Body parts

Human activities can be performed using only one part or several parts of the body, and are interpreted differently according to the culture of the region. Recognition of human activities may require analyzing the movements of different body parts of the individual, e.g., hands, feet, head,...etc (see Fig. 8). The movements of a single limb or several at once enable to describe and give significant information on the underlined action. The hand, for example, can be tracked to detect the communication between individuals. The works presented in [21, 118, 145, 223] are interested in the recognition of hand gestures in particular. The foot is also a part which can be tracked to detect shifting and movements of people or other actions like walking, running, etc. The majority of studies focus on tracking the full body performed activities. [4, 21, 27, 124, 191, 198] conduct their studies on the recognition of postures and human actions through tracking the whole body. Other works, such as [51,

Fig. 8 Different human body parts used to perform actions

192], focus on the follow-up of facial expressions to interpret specific types of human activities, especially in the case of handicapped or disabled people who can move neither their hands nor other parts of their bodies.

6 Image input versus video input

The research in human activities recognition can be classified, according to the nature of input data, into two sub-classes:

i) *Human activities recognition based on static images*: where the system can recognize activities from images like: sitting, walking, eating, ...etc. The activity is distinguishable compared to others by its characteristics. Authors of [124, 125, 211, 223] are interested to recognize human activities from images;

ii) *Human activities recognition based on video*: Some activities cannot be recognized using solely a single current static image. There is a need to have access to extra information related to prior and post event occurring, through, for instance, examining previous and next frames. In this case, videos are more accurate, where the relation between two successive frames can be established. Works in [7, 21, 27, 118, 191, 198] attempt to recognize activities from videos using such approach.

7 Single viewpoint versus multi-view acquisition

The point of view is the place where the grabbing camera is located. Many researchers are interested to recognize activities from only one device placed in a suitable position while others consider many devices placed at different positions to have different views of the activity performed. We can classify works of the state of the art into two categories:

i) *Single view acquisition*: This considers only one viewpoint because of the nature of the acquisition device which captures only the front part as in [118, 124, 191, 223]. We give an example of single view recorded frames from the MSR Daily Activity 3D dataset [206] in Fig. 9a.

ii) *Multiple view acquisition*: In the sequel, multiple devices are used to cover multiple views of the scene as in [27, 52, 85, 204]. We give an example of multi-view recorded frames from the Caviar dataset in Fig. 9b.

8 Validation means

The validation of an approach is a vital step because it allows confirming by tangible proofs that its results are conforming and satisfying the requirements specified in the relative solution. In this context, we classify the research works according to the means used to check the proposals. To validate their approach, the authors of [27] used an experimental platform called DOMUS [137] which is designed and implemented by the MULTICOM team of the Informatics Laboratory of Grenoble. This functional apartment of 34m^2 is equipped with various types of sensors and actuators in order to act on the environment (lighting, shutters, security systems, heating, ventilation, audio-video control...). Real data from 21 people, who performed predefined scenarios of different actions, were collected. Moreover, the authors of [198] validate their proposal through a set of real experiments allowing them to test their method on three sets of actions: MSR 3D Action [90], UT Kinect Action [220] and Florence 3D Action [220]. For each set, half of the subjects was used for the training and the other half for the test. The authors of [222] have used the same validation means in order to support their proposal. However, they have used UT-interaction dataset [154] and UCF 50 dataset [148]. The authors of [54] have conducted their experiments on five benchmark datasets: MSR Action 3D [90], UT Kinect [220], MSR C12 [48], Multiview

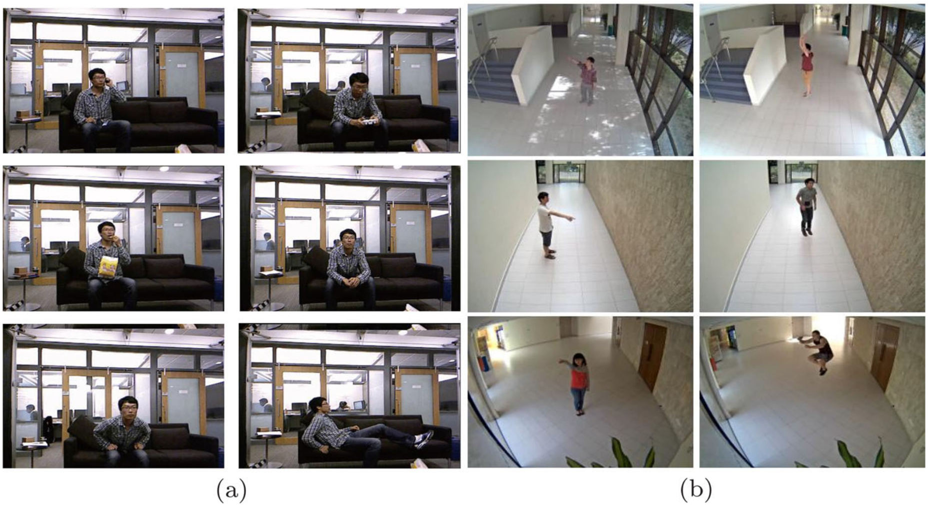


Fig. 9 Viewpoint of the acquisition device: **a** Samples from a single view dataset "MSR Daily Activity 3D [206]", **b** Samples from a multi-view dataset "the Caviar dataset"

3D Action dataset [62] and NTU RGB+D dataset [170]. The authors of [191] as well led experiments on the data extracted from the KTH database to validate and analyze the performance of their approach. They compared the results obtained by the naive Bayesian and SVM classifiers. In addition, [223] created a dataset of actions performed by the members of the laboratory using a WebCam in order to evaluate the performance of their suggested model. While, authors in [234] used the Kinect sensor in order to build a learning model of 22 human postures. These postures are divided into three categories: postures of the hand, foot and full body. Similarly, [223] executed the designed system on a dataset captured by Kinect, containing 1000 depth maps of hand gestures of 10 subjects with 10 catches for each gesture.

In addition, most of the works such as [191, 222] compare the results of their approaches with the results of other methods and approaches suggested in the literature to prove their effectiveness.

8.1 Open datasets

There exist many public datasets that can be used by researchers in order to validate their proposals and to evaluate their performance. According to [204], these datasets / databases can be grouped into several classes depending on the types of action they contain, the viewpoint as well as the nature of data: databases relating to movie scenes, social networks, human behaviors, human poses, atomic actions or daily life activities. [4] enumerates 13 sets of data captured using Kinect that can be used for training and testing. We quote the most used datasets in the literature and categorize them according to activity types. We consider in this classification only four types (levels): atomic action level, behavior level, interaction level and group activities level.

i) *Action level datasets*

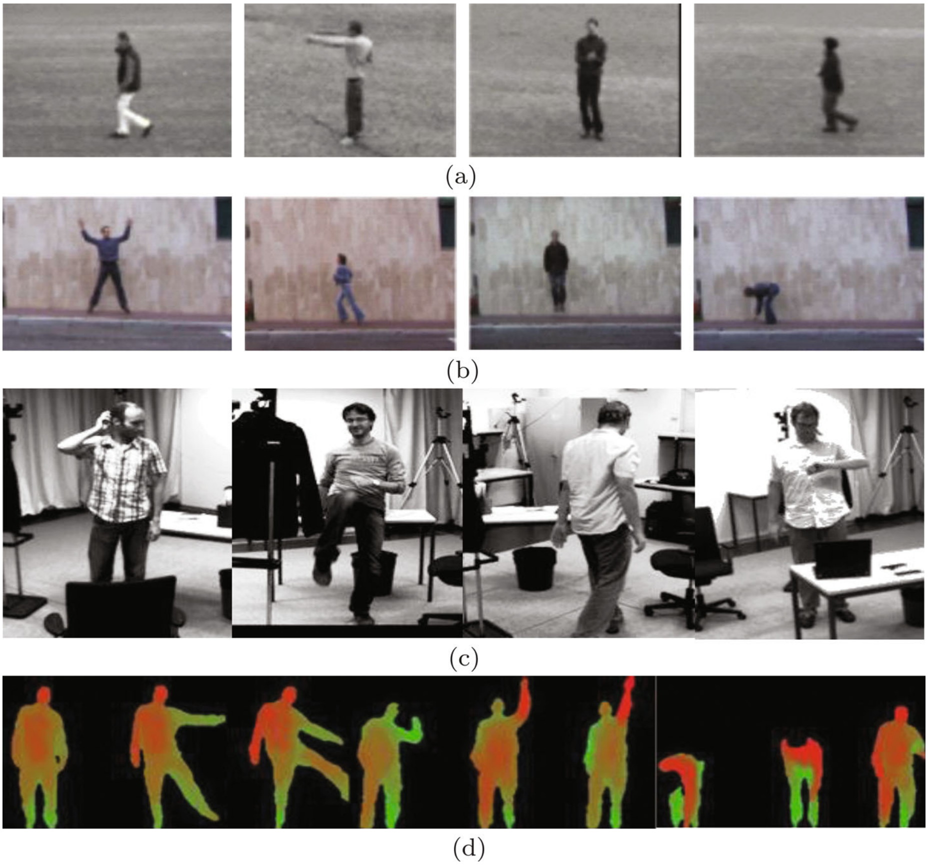
   (a) *KTH Human Action Dataset*: It was created in 2004 by the Royal Institute of Technology of Sweden [167]. It compromises 2391 sequences of six human action classes (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios. All sequences have a length of 4 seconds in average and were taken over homogeneous backgrounds with a static camera (Fig. 10a).

   (b) *Weizmann Human Action Dataset*: It was created by the Weizmann Institute of Science in 2005 [57]. It compromises 90 video sequences of 9 different people performing 10 simple actions (running, walking, skipping, jumping-jack, jumping forward on two legs, jumping in place on two legs, gallop-sideways, waving with two hands, waving with one hand and bending) (Fig. 10b).

   (c) *Stanford 40 Actions dataset*: It was created by the Stanford vision Lab [224]. It contains 9532 images of 40 different classes of actions.

   (d) *IXMAS dataset*: It is a multi-view for view-invariant human action recognition dataset which was created in 2006 [212]. It contains videos of 11 actors performing 13 daily live motions, 3 times each. These actions are recorded with 5 calibrated cameras from different views and include: crossing arms, stretching head, sitting down, ...etc. (Fig. 10c).

   (e) *MSR Action 3D*: It was created by Wanqing Li in the Microsoft Research Redmond [90]. It contains 567 depth map sequences of 10 subjects performing 20 action types twice or 3 times. The sequences were recorded using a Kinect device (Fig. 10d).


(a)

(b)

(c)

(d)

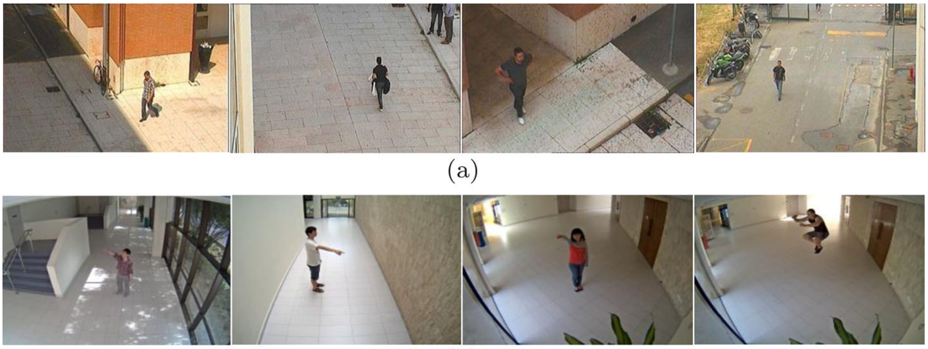
Fig. 10 Samples from action level datasets: **a** KTH Human Action Dataset [167], **b** Weizmann Human Action Dataset [57], **c** IXMAS dataset [212], **d** MSR Action 3D dataset [90]

ii) *Behavior level datasets*:

(a) *VISOR dataset*: It was created in 2005 by the Imagelab Laboratory of the University of Modena and Reggio Emilia [17]. It is composed of several types of videos sorted in different categories. The category "Videos for human action recognition in video surveillance" is used for human action and activity recognition and it contains 130 video sequences (Fig. 11a).

(b) *Caviar dataset*: It was created in 2004. It is composed of two sets: the first one is filmed with a wide-angle camera lens in the entrance lobby of the INRIA Labs in Grenoble, France and the second set also uses a wide-angle lens along and across the hallway in a shopping center in Lisbon. This dataset compromises a number of videos of people performing 9 activities in two different places (Fig 11b).

(c) *Multi-Camera Action Dataset (MCAD)*: was created in the National University of Singapore [89]. It was designed to evaluate the open-view classification problem under surveillance environment. 18 daily actions which are inherited from the KTH, IXMAS and TRECIVD dataset are recorded using 5 cameras and performed


(a)

(b)

Fig. 11 Samples from behavior level datasets: **a** Visor Dataset [17], **b** Caviar dataset

by 20 subjects. For each camera, each action is produced by a subject 8 times (4 times during the day and 4 times in the evening).

iii) *Interaction level datasets*:

(a) *MSR Daily Activity 3D Dataset*: It was created by Jiang Wang in the Microsoft Research Redmond [206]. It consists of 320 sequences for each channel: depth maps, skeleton joint positions and RGB video of 10 subjects performing 16 activities such as drinking, eating, reading... etc. Each activity is carried out twice, once in standing position and once in sitting position (Fig. 12a).

(b) *50 Salads dataset*: It was created at the University of Dundee [188]. It is composed of video sequences of 25 people preparing 2 mixed salads, having a total duration of 4hours.

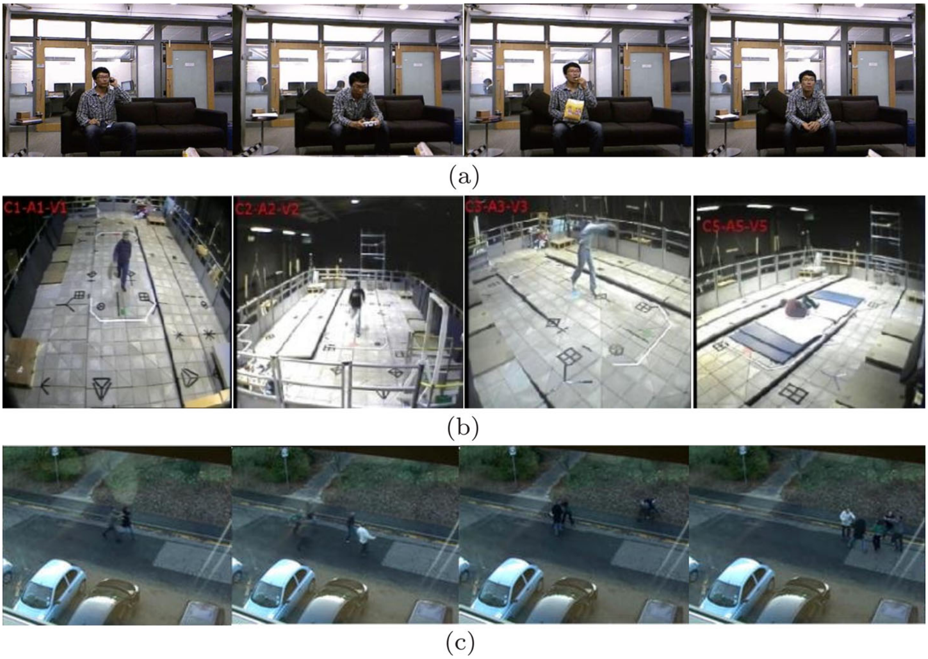
(c) *MuHAVI dataset or Multicamera Human Action Video Dataset*: It was created in 2010 by the Faculty of Science, Engineering and Computing of Kingston University. It aims at evaluating silhouette-based human action recognition methods. It consists of videos of 17 action classes performed by 14 actors several times. The data were recorded using 8 non-synchronized cameras located on 4 sides and 4 corners of a rectangular platform (Fig. 12b).

(d) *UCF50*: it was created by the center for research in Computer Vision, university of Central Florida, USA in 2012 [148]. It is composed of 50 action categories, collected from realistic YouTube videos. This dataset is an extension of YouTube Action dataset (UCF11) which has 11 action categories.

(e) *UCF Sports Action Dataset*: It was created by the center for research in computer vision, university of central florida, USA in 2008 [150, 184]. It is composed of 150 sequences of 11 action categories collected from various sports broadcasted on television channels.

(f) *ETISEO dataset*: it was created in 2005 by the INRIA Institute [121]. It aims at improving video surveillance algorithms. It provides videos of people carrying out several activities in 5 different scenarios: apron, building corridor, building entrance, metro and road.

(g) *Olympic Sports Dataset*: It was created in 2010 by the Stanford vision Lab [126]. It contains 50 videos of athletes practicing 16 different sports. All video sequences are gathered from YouTube.



(a)

(b)

(c)

Fig. 12 Samples from interaction level datasets: **a** MSR Daily Activity 3D dataset [206], **b** MuHAVI dataset, **c** UT-Interaction dataset [155]

(h) *UT-Interaction dataset*: It was created by the university of Texas [155] within the Contest on Semantic Description of Human Activities (SDHA), a research competition to recognize human activities in realistic scenarios, which was held in conjunction with the 20th International Conference on Pattern Recognition (ICPR 2010). It contains 20 video sequences of continuous executions of 6 classes of human-human interactions. Several participants with more than 15 different clothing conditions appear in the videos (Fig. 12c).

(i) *UT-Tower dataset*: It was created as well in the same context of UT-interaction dataset [30]. It consists of 108 video sequences of 9 types of actions. Each action was performed 12 times by 6 individuals. The dataset is composed of two types of scenes: concrete square and lawn.

iv) *Group activities level datasets*:

(a) *ActivityNet Dataset*: It was created in 2015 [67]. It consists of 849 video hours that illustrate 203 activity classes with 137 untrimmed videos for each activity class. It encompasses three scenarios to compare human activity understanding algorithms: untrimmed video classification, trimmed activity classification and activity detection. It is considered as a large-scale video dataset covering a wide range of complex human activities (Fig. 13a).

(b) *The Kinetics Human Action Video Dataset*: It was created by the DeepMind team in 2017 [76]. The initial release (Kinetics_400) contains 400 human action classes with at least 400 video clips for each action taken from different YouTube videos. Kinetics_600 is an approximate super-set of the initial Kinetics_400 dataset that

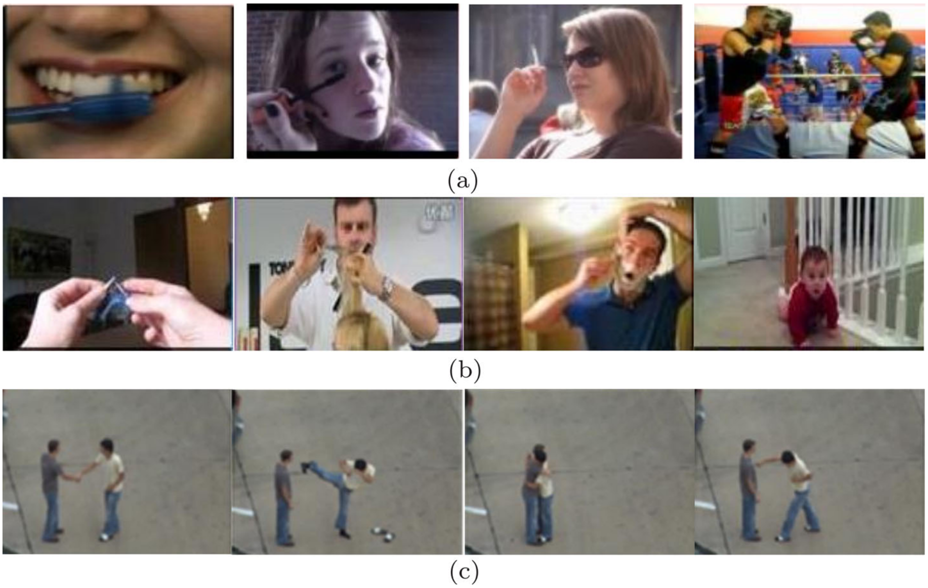


Fig. 13 Samples from group activities level datasets: **a** ActivityNet Dataset [67], **b** UCF-101 Action Recognition Dataset [185], **c** Behave dataset

covers 600 human action classes with at least 600 video clips for each action class. It consists of approximately 500,000 video clips, and each clip lasts around 10 seconds and is labeled with a single class. It is a large-scale, high-quality dataset of YouTube video URLs which include a diverse range of human focused actions.

(c) *HMDB-51 dataset*: It was created in 2011 by the Serre Lab, Brown university USA [84]. It consists of 6849 clips of 51 action categories collected from various sources (movies, public data-bases such as Prelinger archive, YouTube and Google videos). It is considered as one of the largest datasets of human activities recognition.

(d) *Hollywood dataset*: It was created in 2008 by INRIA (Institut national de recherche en informatique et en automatique) France [86]. It is composed of short sequences of human actions collected from realistic videos retrieved from 32 movies.

(e) *Hollywood2 dataset*: It was created as well by INRIA in 2009 [111]. It was proposed to provide realistic and challenging settings (multiple persons, cluttered background...). It is composed of 3669 video clips of 12 classes of human actions and 10 classes of scenes collected from 69 movies.

(f) *UCF-101 Action Recognition Dataset*: It was created by the centre for research in Computer Vision, university of Central Florida, USA in 2012 [185]. It is an extension of UCF50 dataset [148] which has 50 action categories. It is composed of 13 320 videos of 101 realistic action categories collected from YouTube. It gives the largest diversity in terms of actions and realistic settings (viewpoint, illumination conditions ...etc.) (Fig. 13b).

(g) *YouTube Action Dataset*: It was developed by the center for research in Computer Vision, university of Central Florida, USA in 2009 [99]. It contains 11 action categories and it is very challenging due to large variations in camera motion, viewpoints, illumination condition, ...etc.

(h) *Behave dataset*: It was created in 2004 by the School of Informatics of Edinburgh University. It aims at detecting unusual human activities. It is composed of two sets: optical flow data and multi-agent interaction data. The first set is composed of 30 optical flow sequences from the Waverly train station while the second one comprises two views of various scenarios of people interactions (Fig. 13c).

(i) *Video Web Dataset*: It was created in 2010 by the Video Computing Group, belonging to the Department of Electrical Engineering at the University of California Riverside (UCR) [38]. It consists of 2.5 hours of videos of 10 actors interacting with each other, with vehicles or with facilities. Each video is recorded using a camera network of minimum of 4 and maximum of 8 cameras.

For instance, the authors of [191] use the KTH dataset [167], which contains grey level video sequences of low-resolution (160 x 120 pixels), representing atomic human actions. The sets of test and training are separated and contain 461 and 328 images respectively. The authors of [198] use three datasets to test the performance of their approach; 3D MSR Action [90], UT Kinect [220] and 3D Florence Actions [169]. The two last datasets are captured using a fixed Kinect sensor, which is composed of 10 and 9 actions performed by 10 different subjects with a total of 199 and 215 sequences of actions respectively. The authors of [27] recorded video sequences of several activities (dressing, shopping, cleaning, listening to the radio...) performed by 21 persons in predefined scenarios. Moreover, [234] used the Kinect sensor to build a training model for 22 human postures. These postures are divided into three categories: postures of the hand, foot, and full body. For each posture, 100 samples are used. In the same way, [223] have built a training dataset captured by Kinect, which contains 1000 depth maps of hand gestures of 10 subjects with 10 catches for each gesture. We present in Table 1, a classification of benchmark datasets based on activity types, in order to enable the reader to better select the suitable dataset for his study.

8.2 Evaluation metrics

Several performance metrics used in different classification fields have been adapted and used for human activities recognition. Based on [114], we quote in this section frequently used metrics such as accuracy, precision, recall, ...etc. Before summarizing these metrics, we define firstly what would be True Positive, True Negative, False Positive and False Negative for action recognition (see Fig. 14b) as follows:

□ True Positive = actions where the actual and predicted transactions are correct.

□ False Negative = actions which belong to a particular class and are actually predicted to be not from this class.

□ True Negative = actions where the actual and predicted transactions do not correspond to the searched class.

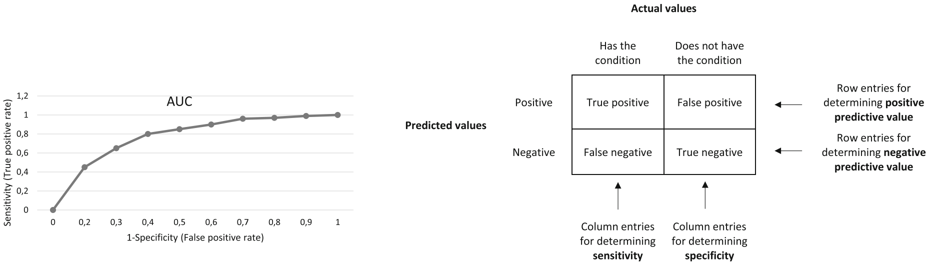
□ False Positive = actions where the actual transactions do not correspond to the searched class, but predicted to be from the searched class.

We describe frequently used performance metrics in the following:

Table 1 Classification of Benchmark datasets based on activity types

Dataset	Action	Behavior	Human-object interaction	Human-human interaction	Group activities
KTH	✓				
Weizmann	✓				
Stanford 40	✓				
IXMAS	✓				
MSR Action 3D	✓				
VISOR		✓			
Caviar		✓			
MCAD		✓			
MSR Daily Activity 3D	✓		✓		
50 Salads	✓		✓		
MuHAVI	✓	✓	✓		
UCF50	✓		✓		
UCF Sports	✓		✓		
ETISEO			✓	✓	
Olympic Sports	✓		✓		
UT-Interaction			✓	✓	
UT-Tower			✓	✓	
ActivityNet	✓	✓	✓	✓	✓
Kinetics	✓	✓	✓	✓	✓
HMDB-51	✓	✓	✓	✓	✓
Hollywood	✓	✓	✓	✓	✓
Hollywood2	✓	✓	✓	✓	✓
UCF-101	✓	✓	✓	✓	✓
YouTube Action	✓	✓	✓	✓	✓
Behave				✓	✓
Video Web			✓	✓	✓

i) *Sensitivity*: It is also called true positive rate, recall or probability of detection. It corresponds to actual positive cases predicted as positive. For action recognition, sensitivity measures the proportion of activities predicted in their classes. Likewise, (1 -



(a) (b)

Fig. 14 Examples of performance evaluation metrics (AUC and confusion matrix): **a** Diagram demonstrating how to calculate the AUC, **b** Structure of the confusion matrix

sensitivity) determines the failure of the system to detect actions. Mathematically, this can be expressed as:

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \tag{1}$$

ii) *Precision*: It is also called Positive Prediction Value (PPV) and corresponds to the likelihood of a detected instance of activity to its real occurrence. Likewise, (1 - precision) determines the probability of the recognizer incorrectly identifying a detected activity. Mathematically, this can be expressed as:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2}$$

iii) *Specificity*: It is also called true negative rate or false positive rate (FPR). It corresponds to actual negative cases predicted as negative. It measures the system sensitivity to negative class. Mathematically, specificity can be calculated as follows:

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \tag{3}$$

iv) *Negative Predictive Value (NPV)*: It is often referred to "negative precision" and measures the likelihood that a negative identification is correct relative to all negative identifications. Mathematically, this can be expressed as:

$$NPV = \frac{TrueNegative}{TrueNegative + FalseNegative} \tag{4}$$

v) *F_Measure*: It determines the harmonic mean of precision and recall. It gives information about the test's accuracy. Hence, F_measure determines at the same time, how precise and robust is the classifier. It reaches its best value at 1 and worst at 0. Mathematically, this can be expressed as:

$$F\_Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

vi) *Accuracy*: It measures the percentage of correct predictions relative to the total number of samples. The accuracy gives good results when the classes are equally sampled. Mathematically, this can be expressed as:

$$Accuracy = \frac{NumberOfCorrectPredictions}{TotalNumberOfPredictionsMade} \tag{6}$$

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalNumberOfSamples} \tag{7}$$

vii) *Likelihood Ratio*: computes the likelihood of an activity predicted when it matches the ground truth compared to the likelihood when it is predicted wrongly. It can be computed for both true positive and true negative results. Mathematically, this can be expressed as:

$$LR+ = \frac{Sensitivity}{1 - Specificity} \tag{8}$$

$$LR- = \frac{1 - Sensitivity}{Specificity} \tag{9}$$

viii) *Area Under Curve (AUC)*: It is widely used for binary classification problems. It measures the probability of the classifier to rank a positive randomly chosen sample higher than a negative randomly chosen sample. In perfect cases, its value is 1. AUC is the area under the curve of plot False Positive Rate (Specificity) vs True Positive Rate (Sensitivity) at different thresholds ranging in [0, 1] (see Fig. 14a).

ix) *Confusion Matrix*: It is also called error matrix and gives a summary of prediction results and describes the complete performance of the model. The confusion matrix shows the errors being made by the classifier as well as their types. Each row of the matrix represents instances of predicted class and each column represents instances in an actual class or vice versa (see Fig. 14b).

x) *Intersection over Union (IoU)*: called also Jaccard index or Jaccard similarity coefficient. It measures the accuracy of the detector on a particular dataset. If we refer to area of overlap between predicted bounding box and ground-truth bounding box with "Area of overlap" and the area encompassed by both the predicted bounding box and the ground-truth bounding box with "Area of Union", this ratio can be calculated as the following:

$$IoU = \frac{Areaofoverlap}{Areaofunion} \tag{10}$$

9 Limitations

This section is a presentation of various issues that may affect the effectiveness of HAR systems. Some of them are specific to the methods used during the various phases of the recognition process. Others are related to the acquisition devices, experimentation environments, or various applications of these systems. Lighting variation is the main difficulty facing vision-based recognition systems in general, because it affects the quality of images and thus, analyzed information. In the same way, perspective change is another limitation of the current systems which were implemented to operate using a single view acquisition devices. This problem reduces the amount of extracted information and provides a limited visualization of activities being analyzed. This includes occlusion with its different types: self occlusion where body parts occlude each other, occlusion of another object and partial occlusion of human body parts are major limitations to HAR systems. These problems are discussed in the works of [4, 118, 120, 124, 145, 198]. The variety of gestures linked to the complex structure of human activities and the similarity between classes of different actions can also be source of additional difficulties due to data association problem it may involve. This needs to be addressed in order to set up complete, robust and flexible HAR systems under various conditions or environments. The limitation of hand configurations, predefined actions or postures, and recognition of simple gestures or activities are discussed in [4, 27, 118, 145, 223]. Some methods of detection based on the form or the appearance, such as colorimetric segmentation, can confuse the human body parts with objects of the scene, as in [118, 145, 222, 223] or cannot operate correctly during variation in appearance or clothing of people such as in [4, 198]. These problems are associated with other problems related to methods of recognition or acquisition devices, such as noise [118], complex or moving backgrounds and unstructured scenes as in [4, 120, 124, 145, 198, 222], and scale variation when the person gets closer or move away from the camera [4, 198]. Finally, many researchers rely on their own recorded datasets to test the performance of their proposals. This raises the concerns about the limitation of available benchmark dataset for testing novel domain-specific applications. We can mention for instance the case of daily life activities and fall detection datasets which are not large enough to be used for training effective models.

10 Challenges of the recognition systems

The current HAR systems present a big number of challenges they have to cope with in order to provide the principal functions for which they are developed. For instance, most HCI or video surveillance systems based on HAR must provide continuous monitoring and generate reliable answers at the right time in order to ensure the performance of the provided results. This challenge is discussed in [118, 145, 223]. Furthermore, this becomes a major issue when modelling and analyzing interactions between people and objects with an appropriate level of accuracy which is still challenging. This would be very useful for surveillance and public security applications and may help to detect several abnormality scenarios.

The use of these systems for the aim of surveillance, elderly assistance and patient monitoring together with the increasing implementation costs raise also new societal challenges: acceptance by the society, privacy, side effects of installation of these devices at home as well as large scale applications. For instance, authors in [27, 118] discuss the challenge of device integration at home for tracking, which is considered as violation of intimacy and privacy. To address this last problem, it is interesting to explore the development of HAR systems on smartphones. This may contribute to overcome the user's privacy constraint since the recorded data would be stored on his own device and may reduce also the computational time related to the transmission between the distant server and the device. However, on-device implementation is a challenging issue as well due to memory constraint, high number of parameters needed by the recognition model and the battery life of the device [127]. Another challenge is related to implicit dependency of such systems with the physical and physiological abilities of the user. Intuitively, HAR systems should not depend on the user's age, color, size or capacity to use such systems. Both experienced user and beginner should be able to use these systems, and in the same way. This challenge is raised in [118, 145]. The gestures independence and gestures spotting from continuous data streams constitute also another type of challenge, since it is still difficult to localize temporally the gesture in long continuous videos. Indeed, HAR systems are not yet able to detect and recognize various gestures under different background conditions and are not tolerant with the scalability and growth of gestures. This is the major challenge which is studied and considered by many works such as: [4, 118, 145, 198, 223]. One more area that need to be explored is the ability of HAR systems to be context-aware. This may help to make use of proposed approaches and progress made in many application domains.

Understanding and detection of daily life activities in long-term videos is a challenging task. This is due to the fact that the long-term videos containing daily life activities are composed of several complex activities. These activities are difficult to model because of their complex structure as well as the big variation in ways of performing the same activity. Another issue is related to the overlapping between starting and ending time of each particular activity. This challenge is addressed by [120]. In addition, the discrimination between intentional and involuntary actions is still very challenging area to tackle.

Other challenges are discussed in [7, 21], which are human activities recognition through missed parts of video, recognition of more than one activity performed by one person at the same time, and early recognition and prediction of actions, especially in crowded environments. Nowadays, memory constraint, high number of parameters update, collection and fusion of large multi-modal variant data for the training process as well as deployment of different architectures of deep-learning based methods in smartphones or wearable devices are still unresolved issues in deep-learning HAR systems [127].

11 Conclusion

The need to understand and interpret effectively human activities has become unavoidable in several applications of computer vision, HCI, robotics, security and home monitoring. This paper aims to give an overview of the recent works in this field of research. It proposes a classification according to several criteria. It initially discusses the different applications of HAR, and the major objectives intended by these systems. Then, it presents an analysis of the used approaches in the state of the art, as well as the means used in their validation. It also provides a taxonomy of human activity types and the used methods for action representations. Moreover, another classification of the approaches, depending on the nature of the acquisition device (e.g., uni-modal and multi-modal methods), is given along with a categorization of these methods according to the stage of recognition in which they can be applied (detection, tracking and recognition). During this study, it is observed that almost every approach still suffers from certain limitations. However, we notice that deep learning-based approaches are getting more attention nowadays due to the progress they have made and the promising results in terms of detection and recognition performance. On the other hand, interactions and group activities recognition are among prominent research topics because interactions between people or with objects can provide useful information in many HAR application fields such as video surveillance, public security, abnormal activity detection, ... etc. Moreover, the integration of HAR systems in smartphones is to be explored as well, since the smartphones have worthy emerged to our daily life, and, are shown to be non-intrusive and widely accepted by the society, which overcome the privacy barriers seen in standard HAR systems. The key to a successful human action recognition system is the effective modelling, discriminative representation and accurate analysis of features of the recorded data, coupled with fast and accurate action detection results. As a conclusion, we can say that the developed techniques having as purposes the recognition and understanding of human activities are still facing several open problems, limitations, and still require big challenges to be addressed.

Acknowledgments The authors wish to thank Prof Mourad Oussalah from the center for machine vision and signal analysis, university of Oulu, who participated in writing and technical editing of the manuscript.

Funding Information Open access funding provided by University of Oulu including Oulu University Hospital.

Appendix

Table 2 Analysis of some state-of-the-art comprehensive surveys on HAR

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[145]	2015	Gesture taxonomies, hand gesture recognition applications and approaches.	Different fields	Gestures	Hands	Image and video	Not interested to different activities and focuses only on gestures.
[32]	2015	An approach-based taxonomy of the state-of-the-art research and advances in HAR.	Different fields	Different types	Different body parts	Image and video	Follows a specific taxonomy and doesn't cover a wide range of deep-learning based aproaches.
[107]	2015	Comprehensive survey on kinect-based motion recognition techniques and the underlying datasets.	Applications of the Kinect technology	Different types	Different body parts	Image and video	Devoted only to motion recognition using data captured by Kinect.
[204]	2015	Classification of HAR approaches regarding the source of input data into unimodal or multimodal methods and analysis of some publicly available human activity datasets.	Different fields	Different types	Different body parts	Image and video	Presents an approach-based taxonomy according to the source of input data but doesn't cover many general aspects of HAR.
[237]	2015	Semantic-based human recognition methods and a brief representation of their application fields.	Different fields	Actions and interactions	Different body parts	Image and video	Focuses on semantic-based HAR methods and doesn't include many general aspects of HAR.
[132]	2015	Video-based HAR using deep learning and classification of datasets according to different complexity levels.	Different fields	Different types	Different body parts	Image and video	Interested to deep learning based HAR and doesn't cover hand-crafted approaches.
[7]	2016	Analysis of different HAR methods and comparison between different action identification methods.	Different fields	Different types	Different body parts	Video	Doesn't cover many HAR approaches and general aspects.

Table 2 (continued)

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[12]	2016	Comprehensive survey on the recent techniques of HAR.	Different fields	Actions	Different body parts	Video	Doesn’t cover many aspects and benchmark databases of HAR.
[74]	2016	Classification of various action recognition and detection algorithms according to the extraction and encoding of features, as well as the classification processes.	Different fields	Different types	Different body parts	Image and video	Covers many aspects of HAR but many research works have emerged from 2016 till now.
[189]	2016	Overview of methodologies, challenges and issues of HAR systems.	Different fields	Actions and interactions	Different body parts	Image and video	Doesn’t provide an indepth study. It doesn’t cover many techniques and aspects of HAR.
[140]	2016	3D skeleton-based HAR approaches.	Different fields	Actions	Different body parts	Video	Focuses mostly on 3D skeleton-based HAR and omits a wide range of other approaches.
[174]	2016	Analysis of popular techniques used for object segmentation and recognition.	Different fields	Actions	Different body parts	Video	Devoted to object segmentation and detection in general and is not specific to HAR.
[21]	2016	Overview of HAR techniques in videos.	Surveillance, entertainment and healthcare.	Different types	Different body parts	Video	Doesn’t cover many HAR approaches and is limited to some specific application fields.
[122]	2016	Comprehensive survey on the recent development and challenges of human detection.	Different fields	Actions	Different body parts	Image and video	Devoted to human activities detection and doesn’t cover the whole process of HAR.

Table 2 (continued)

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[129]	2016	Knowledge-based HAR methodologies.	Different fields	Actions	Different body parts	Video	Interested to methods incorporating a priori knowledge and context information on the activity and doesn't cover many other approaches.
[33]	2016	Comprehensive survey of the emerging progress on 3D hand gesture recognition approaches and systems.	Human computer interaction	Gestures	Hands	Video	The emphasis is on 3D hand gesture recognition approaches. It doesn't cover other activity types.
[236]	2016	Comprehensive analysis and comparison between learning-based and handcrafted action representations.	Different fields	Different types	Different body parts	Image and video	Presents a human action representation based taxonomy and omits many other aspects of HAR.
[44]	2016	Current state of publicly available HAR datasets.	Different fields	Different types	Different body parts	Image and video	The focus of this survey is the available datasets for HAR. It doesn't cover HAR approaches.
[232]	2016	Comprehensive review of the most commonly used action recognition related RGB-D video datasets.	Different fields	Different types	Different body parts	Video	Presents only RGB-D video datasets and doesn't discuss HAR approaches.
[123]	2016	Review of the state of the art of vision-based systems for the recognition of daily life activities.	Daily life activities	Different types	Different body parts	Video	The focus is made on techniques related to daily life activities and omits many other application domains.

Table 2 (continued)

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[25]	2017	Categorization of video-based HAR techniques into hand-crafted feature-based and deep learning-based approaches.	Different fields	Actions	Different body parts	Video	The focus is automatic HAR techniques in videos. Still images are omitted.
[63]	2017	Survey of existing space-time action representations based on 3D skeletal data.	Different fields	Actions	Different body parts	Video	Focuses on 3D human representation based on skeletal data, and omits other data representations.
[45]	2017	State of the art of multimodal gesture recognition.	Machine learning and computer vision	Gestures	Hands	Image and video	Devoted to gesture recognition using multimodal data. It doesn't cover other activity types.
[8]	2017	Complex event recognition techniques.	Event recognition	Event	Different body parts	Image and video	Interested to complex event techniques and doesn't cover HAR approaches related to other activity types.
[68]	2017	Comprehensive review of the notable steps taken towards recognizing human actions.	Different fields	Actions	Different body parts	Image and video	Classifies methods of human action only and doesn't cover other activity types and action detection methods.
[16]	2017	Survey on current deep learning methodologies for action and gesture recognition.	Different fields	Actions and gestures	Different body parts	Video	Deep-learning based taxonomy for action and gesture recognition with particular interest on temporal dimension of data. Spatial features are not covered.

Table 2 (continued)

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[59]	2017	Discussion of research works focusing on identifying, tracking and understanding group activities, interactions, and abnormal activities detection in large crowds with a summary of underlying available datasets.	Crowd management, public space design, and visual surveillance	Interactions and group activities	Different body parts	Video	Specific to crowd analysis for video surveillance purposes and abnormality detection. It doesn't cover other application domais.
[210]	2018	RGB-D-based human motion recognition with deep learning focusing on three architectures of neural networks.	Different fields	Different types	Different body parts	Video	Dedicated to RGB-D-based human motion recognition using deep learning and doesn't cover many other approaches.
[180, 181]	2019	Presentation and comparison of different types of video datasets, challenges, and their related latest evaluation techniques.	Different fields	Different types	Different body parts	Video	Devoted only to datasets analysis and doesn't discuss many other aspects of HAR.
[231]	2019	Survey of HAR methods, including progress in both hand-designed and deep learning-based action feature representation methods.	Different fields	Actions and interactions	Different body parts	Image and video	Doesn't cover many HAR approaches and different activity types.
[34]	2019	Review of state-of-the-art techniques used in recent hand gesture and sign language recognition research.	Sign language recognition	Gestures	Hands	Image and video	Devoted only to hand gesture recognition techniques and doesn't include other activity types.

Table 2 (continued)

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[193]	2019	Summary of techniques of one person/ a group of people and their social interactions as well as their interaction with robots.	Robotics	Interactions and group activities	Different body parts	Video	Reviews recent approaches for recognizing human activities within the general framework of interaction with robots. It doesn't cover many other application domains.
[1]	2019	A comprehensive review on abnormal crowd behaviour detection methods.	Intelligent surveillance video systems	Interactions and group activities	Different body parts	Video	Focuses on detection abnormal activities methods in a crowded scene scenario. It omits many other application domains.
[186]	2019	Deep learning based techniques for various crowd video analysis methods.	Surveillance video analysis	Interactions and group activities	Different body parts	Video	Focuses on intelligent surveillance video analysis techniques and omits many other HAR application domains.
[23]	2019	State-of-the-art techniques on crowd behavior analysis, motion patterns, tracking, activity analysis and modeling. Evaluation metrics and datasets are also discussed.	Video surveillance	Interactions and group activities	Different body parts	Video	Devoted to crowd behavior analysis for video surveillance. It omits many other HAR approaches and application domains.
[139]	2019	Some insights of the state-of-the-art research works in the fields of Intelligent Video Surveillance, Wireless Sensor Network-based HAR, and camera-based health monitoring.	Intelligent Video Surveillance and health monitoring.	Actions	Different body parts	Video	Devoted to specific applications of HAR and doesn't cover many HAR approaches.

Table 2 (continued)

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
[127]	2019	In-depth and comprehensive analysis of data fusion and multiple techniques for HAR.	Different fields	Actions	Different body parts	Image and video	Emphasis on data fusion and applications of HAR on mobile and wearable devices. It omits many HAR approaches and applications.
[39]	2019	Overview of existing abnormal human activity recognition approaches.	Smart home surveillance and public place security	Actions, interactions and group activities	Different body parts	Image and video	Focuses on abnormal human activity recognition and fall detection. It omits many other applications domains and approaches of HAR.
[225]	2019	Comprehensive review of the CNN-based action recognition methods.	Different fields	Actions	Different body parts	Image and video	Presents CNN-based HAR approaches and doesn't include any of the hand-crafted methods.

References

1. Afiq A, Zakariya M, Saad M, Nurfarzana A, Khir MHM, Fadzil A, Jale A, Gunawan W, Izuddin Z, Faizari M (2019) A review on classifying abnormal behavior in crowd scene. J Vis Commun Image Represent 58:285
2. Aggarwal J, Cai Q (1999) Human motion analysis: a review. Comput Vis Image Understand 73(3):428
3. Aggarwal J, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv (CSUR) 43(3):16
4. Aggarwal J, Xia L (2014) Human activity recognition from 3d data: a review. Pattern Recogn Lett 48:70
5. Ahmad M, Lee SW (2008) Human action recognition using shape and clg-motion flow from multi-view image sequences. Pattern Recogn 41(7):2237–2252
6. Ahsan U, Sun C, Essa I (2018) Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks, arXiv:1801.07230
7. Akansha UA, Shailendra M, Singh N (2016) Analytical review on video-based human activity recognition. In: Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, pp 3839–3844
8. Alevizos E, Skarlatidis A, Artikis A, Paliouras G (2017) Probabilistic complex event recognition: a survey. ACM Comput Surv (CSUR) 50(5):71
9. Ali A, Aggarwal J (2001) Segmentation and recognition of continuous human activity. In: 2001. Proceedings. IEEE Workshop on Detection and recognition of events in video. IEEE, pp 28–35
10. Ali S, Shah M (2010) Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans Pattern Anal Mach intell 32(2):288
11. AlMubarak HA, Stanley J, Guo P, Long R, Antani S, Thoma G, Zuna R, Frazier S, Stoecker W (2019) A hybrid deep learning and handcrafted feature approach for cervical cancer digital histology image classification. Int J Healthcare Inf Syst Inf (IJHISI) 14(2):66
12. Amirbandi EJ, Shamsipour G (2016) Exploring methods and systems for vision based human activity recognition. In: 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC). IEEE, pp 160–164
13. Angelini F, Fu Z, Velastin S, Chambers JA, Naqvi SM (2018) 3d-hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 4219–4223
14. Antoshchuk S, Kovalenko M, Sieck J (2018) Gesture recognition-based human–computer interaction interface for multimedia applications. In: Digitisation of Culture: Namibian and International Perspectives. Springer, pp 269–286
15. Argyros AA, Lourakis MI (2006) Binocular hand tracking and reconstruction based on 2d shape matching. In: 2006. ICPR 2006. 18th International Conference on Pattern Recognition, vol 1. IEEE, pp 207–210
16. Asadi-Aghbolaghi M, Clapes A, Bellantonio M, Escalante HJ, Ponce-López V, Baró X, Guyon I, Kasaei S, Escalera S (2017) A survey on deep learning based approaches for action and gesture recognition in image sequences. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, pp 476–483
17. Ballan L, Bertini M, Del Bimbo A, Seidenari L, Serra G (2009) Effective codebooks for human action categorization. In: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, pp 506–513
18. Beddiar DR, Nini B (2017) Vision based abnormal human activities recognition: An overview. 2017 8th International Conference on Information Technology (ICIT), pp 548–553
19. Ben Youssef M, Trabelsi I, Bouhlel M (2016) Human action analysis for assistance with daily activities. International Journal on Human Machine Interaction
20. Berger K (2013) The role of rgb-d benchmark datasets: an overview, arXiv:1310.2053
21. Bhardwaj R, Singh PK (2016) Analytical review on human activity recognition in video. In: 2016 6th International Conference Cloud System and Big Data Engineering (Confluence). IEEE, pp 531–536
22. Borges PVK, Conci N, Cavallaro A (2013) Video-based human behavior understanding: a survey. IEEE Trans Circ Syst Video Technol 23(11):1993
23. Bour P, Cribelier E, Argyriou V (2019) Crowd behavior analysis from fixed and moving cameras. In: Multimodal Behavior Analysis in the Wild. Elsevier, pp 289–322
24. Bux A (2017) Vision-based human action recognition using machine learning Techniques. Ph.d. thesis, Lancaster University
25. Bux A, Angelov P, Habib Z (2017) Vision based human activity recognition: a review. In: Advances in Computational Intelligence Systems. Springer, pp 341–371
26. Cao Y, Guan D, Huang W, Yang J, Cao Y, Qiao Y (2019) Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. Inf Fusion 46:206

27. Chahuara P, Fleury A, Vacher M, Portet F (2012) Méthodes SVM et MLN pour la reconnaissance automatique d'activités humaines dans les habitats perceptifs: tests et perspectives. In: RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle), Lyon, pp 978–2–9539,515–2–3. https://hal.archives-ouvertes.fr/hal-00656557. Session "Posters"
28. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633
29. Chaudhry R, Ofli F, Kurillo G, Bajcsy R, Vidal R (2013) Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 471–478
30. Chen CC, Ryoo MS, Aggarwal J (2010) UT-Tower Dataset: Aerial View Activity Classification Challenge. http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html
31. Chen L, Wei H, Ferryman J (2013) A survey of human motion analysis using depth imagery. Pattern Recogn Lett 34(15):1995
32. Cheng G, Wan Y, Saudagar AN, Namuduri K, Buckles BP (2015) Advances in human action recognition: A survey, arXiv:1501.05964
33. Cheng H, Yang L, Liu Z (2016) Survey on 3d hand gesture recognition. IEEE Trans Circ Syst Video Techn 26(9):1659
34. Cheok MJ, Omar Z, Jaward MH (2019) A review of hand gesture and sign language recognition techniques. Int J Mach Learn Cybern 10(1):131
35. Chu W, Xue H, Yao C, Cai D (2019) Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos. IEEE Trans Multimed 21(1):246
36. Chu WT, Chu HA (2019) A genetic programming approach to integrate multilayer cnn features for image classification. In: International Conference on Multimedia Modeling. Springer, pp 640–651
37. Cornacchia M, Ozcan K, Zheng Y, Velipasalar S (2017) A survey on activity detection and classification using wearable sensors. IEEE Sens J 17(2):386
38. Denina G, Bhanu B, Nguyen HT, Ding C, Kamal A, Ravishankar C, Roy-Chowdhury A, Ivers A, Varda B (2011) Videoweb dataset for multi-camera activities and non-verbal communication. In: Distributed Video Sensor Networks. Springer, pp 335–347
39. Dhiman C, Vishwakarma DK (2019) A review of state-of-the-art techniques for abnormal human activity recognition. Eng Appl Artif Intell 77:21
40. Dixon S (2018) Human activity workflow parsing
41. Doersch C (2016) Tutorial on variational autoencoders, arXiv:1606.05908
42. Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: 2005. 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. IEEE, pp 65–72
43. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118
44. Edwards M, Deng J, Xie X (2016) From pose to activity: Surveying datasets and introducing converse. Comput Vis Image Underst 144:73
45. Escalera S, Athitsos V, Guyon I (2017) Challenges in multi-modal gesture recognition. In: Gesture Recognition. Springer, pp 1–60
46. Firman M (2016) Rgbd datasets: Past, present and future. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 19–31
47. Forsyth DA, 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
48. Fothergill S, Mentis H, Kohli P, Nowozin S (2012) Instructing people for training gestural interactive systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp 1737–1746
49. Fu J, Xiong L, Song X, Yan Z, Xie Y (2017) Identification of finger movements from forearm surface emg using an augmented probabilistic neural network. In: 2017 IEEE/SICE International Symposium on System Integration (SII). IEEE, pp 547–552
50. Fu Y, Hospedales TM, Xiang T, Gong S (2014) Learning multimodal latent attributes. IEEE Trans Pattern Anal Mach Intell 36(2):303
51. Gan C, Wang N, Yang Y, Yeung DY, Hauptmann AG (2015) Devnet: A deep event network for multimedia event detection and evidence recounting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2568–2577
52. Garcia-Ceja E, Galván-Tejada CE, Brena R (2018) Multi-view stacking for activity recognition with sound and accelerometer data. Inf Fusion 40:45

53. Gavrilova ML, Wang Y, Ahmed F, Paul PP (2018) Kinect sensor gesture and activity recognition: New applications for consumer cognitive systems. IEEE Consum Electron Mag 7(1):88
54. Ghorbel E, Boutteau R, Boonaert J, Savatier X, Lecoeuche S (2018) Kinematic spline curves: a temporal invariant descriptor for fast action recognition. Image Vis Comput 77:60
55. Gleick J, Dyson FJ (1992) Genius: The life and science of richard feynman. Phys Today 45:87
56. Gonzalez L, Velastin S, Acuna G (2018) Silhouette-based human action recognition with a multi-class support vector machine
57. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247
58. Goyani M, Patel N (2017) Multi-level haar wavelet based facial expression recognition using logistic regression. Indian Journal of Science and Technology 10(9):976–990
59. Grant JM, Flynn PJ (2017) Crowd scene understanding from video: a survey. ACM Trans Multimed Comput Commun Appl (TOMM) 13(2):1
60. Guo G, Lai A (2014) A survey on still image based human action recognition. Pattern Recogn 47(10):3343
61. Guo Y, Zhang J, Lu M, Wan J, Ma Y (2014) Benchmark datasets for 3d computer vision. In: 2014 IEEE 9th Conference on Industrial Electronics and Applications (ICIEA). IEEE, pp 1846–1851
62. Hammouche M, Ghorbel E, Fleury A, Ambellouis S (2016) Toward a real time view-invariant 3d action recognition. In: International joint conference on computer vision, imaging and computer graphics theory and applications (VISIGRAPP)
63. Han F, Reily B, Hoff W, Zhang H (2017) Space-time representation of people based on 3d skeletal data: a review. Comput Vis Image Underst 158:85
64. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans Cybern 43(5):1318
65. Haria A, Subramanian A, Asokkumar N, Poddar S, Nayak JS (2017) Hand gesture recognition for human computer interaction. Procedia Comput Sci 115:367
66. Hassner T (2013) A critical review of action recognition benchmarks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 245–250
67. Heilbron FC, Escorcia V, Ghanem B, Niebles JC (2015) Activitynet: A large-scale video benchmark for human activity understanding. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 961–970. https://doi.org/10.1109/CVPR.2015.7298698
68. Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: a survey. Image Vis Comput 60:4
69. Hu W, Tan T, Wang L, Maybank S (2004) A survey on visual surveillance of object motion and behaviors. IEEE Trans Syst Man Cybern Part C (Appl Rev) 34(3):334
70. Ijjina EP, Chalavadi KM (2017) Human action recognition in rgb-d videos using motion sequence information and deep learning. Pattern Recogn 72:504
71. Inkawhich N, Inkawhich M, Chen Y, Li H (2018) Adversarial attacks for optical flow-based action recognition classifiers, arXiv:1811.11875
72. Islam S, Qasim T, Yasir M, Bhatti N, Mahmood H, Zia M (2018) Single-and two-person action recognition based on silhouette shape and optical point descriptors, Signal. Image Video Process 12(5):853
73. Jadooki S, Mohamad D, Saba T, Almazyad AS, Rehman A (2017) Fused features mining for depth-based hand gesture recognition to classify blind human communication. Neural Comput Appl 28(11):3285
74. Kang SM, Wildes RP (2016) Review of action recognition and detection methods, arXiv:1610.06906
75. Kang W, Deng F (2007) Research on intelligent visual surveillance for public security. In: 2007. ICIS 2007. 6th IEEE/ACIS International Conference on Computer and Information Science. IEEE, pp 824–829
76. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al (2017) The kinetics human action video dataset, arXiv:1705.06950
77. Ke SR, Thuc H, Lee YJ, Hwang JN, Yoo JH, Choi KH (2013) A review on video-based human activity recognition. Computers 2(2):88
78. Khaire P, Kumar P, Imran J (2018) Combining cnn streams of rgb-d and skeletal data for human activity recognition. Pattern Recognition Letters
79. Kliper-Gross O, Hassner T, Wolf L (2012) The action similarity labeling challenge. IEEE Trans Pattern Anal Mach Intell 34(3):615
80. Kong Y, Jia Y, Fu Y (2014) Interactive phrases: Semantic descriptionsfor human interaction recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence (50) 2:1775

81. Krüger V, Kragic D, Ude A, Geib C (2007) The meaning of action: a review on action recognition and mapping. Adv Robot 21(13):1473
82. Kuehne H, Arslan A, Serre T (2014) The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 780–787
83. Kuehne H, Gall J, Serre T (2016) An end-to-end generative framework for video segmentation and recognition. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 1–8
84. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV)
85. Kumar V, Chaturvedi A, Rai AK (2018) A framework using multiple features to detect multi-view human activity. In: Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), pp 26–27
86. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE Conference on computer vision & pattern recognition
87. Li C, Song D, Tong R, Tang M (2019) Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recogn 85:161
88. Li J, Zhang B, Lu G, Zhang D (2019) Generative multi-view and multi-feature learning for classification. Inf Fusion 45:215
89. Li W, Wong Y, Liu AA, Li Y, Su YT, Kankanhalli M (2016) Multi-camera action dataset (MCAD): a dataset for studying non-overlapped cross-camera action recognition. CoRR arXiv:1607.06408
90. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, pp 9–14
91. Li Y (2012) Hand gesture recognition using kinect. In: 2012 IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS). IEEE, pp 196–199
92. Li Y, Mavadati SM, Mahoor MH, Zhao Y, Ji Q (2015) Measuring the intensity of spontaneous facial action units with dynamic bayesian network. Pattern Recogn 48(11):3417
93. Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CG (2018) Videolstm convolves, attends and flows for action recognition. Comput Vis Image Underst 166:41
94. Liang Y, Zhou X, Yu Z, Guo B (2014) Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare. Mob Netw Appl 19(3):303–317
95. Lin W, Mi Y, Wang W, Wu J, Wang J, Mei T (2016) A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. IEEE Trans Image Process 25(4):1674
96. Lin Z, Jiang Z, Davis LS (2009) Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 444–451
97. Liu AA, Xu N, Nie WZ, Su YT, Zhang YD (2019) Multi-domain and multi-task learning for human action recognition. IEEE Trans Image Process 28(2):853
98. Liu H, Feris R, Sun MT (2011) Benchmarking datasets for human activity recognition. In: Visual Analysis of Humans. Springer, pp 411–427
99. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild". In: 2009. CVPR 2009. IEEE conference on Computer vision and pattern recognition. IEEE, pp 1996–2003
100. Liu L, Shao L, Li X, Lu K (2016) Learning spatio-temporal representations for action recognition: a genetic programming approach. IEEE Trans Cybern 46(1):158
101. Liu L, Wang S, Hu B, Qiong Q, Wen J, Rosenblum DS (2018) Learning structures of interval-based bayesian networks in probabilistic generative model for human complex activity recognition. Pattern Recogn 81:545
102. Liu S, Chen C, Kehtarnavaz N (2016) A computationally efficient denoising and hole-filling method for depth image enhancement. In: Real-Time Image and Video Processing 2016, vol 9897. International Society for Optics and Photonics, pp 98970V
103. Liu T, Chen Z, Liu H, Zhang Z, Chen Y (2018) Multi-modal hand gesture designing in multi-screen touchable teaching system for human-computer interaction. In: Proceedings of the 2Nd International Conference on Advances in Image Processing, ICAIP '18. ACM, New York, ,pp 198–202. https://doi.org/10.1145/3239576.3239619
104. Lopes HCT (2017) Contextual game design: from interface development to human activity recognition
105. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91
106. Lucas BD, 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, , IJCAI'81, Vol 2. Morgan Kaufmann Publishers Inc., San Francisco, pp 674–679. http://dl.acm.org/citation.cfm?id=1623264.1623280

107. Lun R, Zhao W (2015) A survey of applications and human motion recognition with microsoft kinect. Int J Pattern Recogn Artif Intell 29(05):1555008
108. Ma S, Zhang J, Sclaroff S, Ikizler-Cinbis N, Sigal L (2018) Space-time tree ensemble for action recognition and localization. Int J Comput Vis 126(2-4):314
109. Machado IP, Gomes AL, Gamboa H, Paixao V, Costa RM (2015) Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. Inf Process Manag 51(2):204
110. Mademlis I, Tefas A, Pitas I (2019) Greedy salient dictionary learning for activity video summarization. In: International Conference on Multimedia Modeling. Springer, pp 578–589
111. Marszałek M, Laptev I, Schmid C (2009) Actions in context. In: IEEE Conference on computer vision & pattern recognition
112. Mathieu M, Couprie C, LeCun Y (2015) Deep multi-scale video prediction beyond mean square error, arXiv:1511.05440
113. Mavroudi E, Bhaskara D, Sefati S, Ali H, Vidal R (2018) End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding, arXiv:1801.09571
114. Minnen D, Westeyn T, Starner T, Ward J, Lukowicz P (2006) Performance metrics and evaluation issues for continuous activity recognition. Performance Metrics for Intelligent Systems **4**. pp303-317
115. Mirchev A, Ahmadi SA (2018) Classification of sparsely labeled spatio-temporal data through semi-supervised adversarial learning, arXiv:1801.08712
116. Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Understand 81(3):231
117. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Understand 104(2-3):90
118. Mollet N, Chellali R (2005) Détection et interprétation des gestes de la main. In: 2005 3rd International Conference on SETIT
119. Murthy G, Jadon R (2009) A review of vision based hand gestures recognition. Int J Inf Technol Knowl Manag 2(2):405
120. Negin F, Koperski M, Crispim CF, Bremond F, Coşar S, Avgerinakis K (2016) A hybrid framework for online recognition of activities of daily living in real-world settings. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 37–43. https://doi.org/10.1109/AVSS.2016.7738021
121. Nghiem AT, Bremond F, Thonnat M, Valentin V (2007) Etiseo, performance evaluation for video surveillance systems. In: 2007. AVSS 2007. IEEE Conference on Advanced Video and Signal Based Surveillance. IEEE, pp 476–481
122. Nguyen DT, Li W, Ogunbona P (2016) Human detection from images and videos: A survey. Pattern Recogn 51:148
123. Nguyen THC, Nebel JC, Florez-Revuelta F et al (2016) Recognition of activities of daily living with egocentric vision: a review. Sensors 16(1):72
124. Nguyen-Duc-Thanh N, Stonier D, Lee S, Kim DH (2011) A new approach for human-robot interaction using human body language. In: International Conference on Hybrid Information Technology. Springer, pp 762–769
125. Nicolaou MA, Pavlovic V, Pantic M (2014) Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. IEEE Trans Pattern Anal Mach Intell 36(7):1299. https://doi.org/10.1109/TPAMI.2014.16
126. Niebles JC, Chen CW, Fei-Fei L (2010) Modeling temporal structure of decomposable motion segments for activity classification. In: European conference on computer vision. Springer, pp 392–405
127. Nweke HF, Teh YW, Mujtaba G, Al-garadi MA (2019) Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Inf Fusion 46:147
128. Ohn-Bar E, Trivedi M (2013) Joint angles similarities and hog2 for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 465–470
129. Onofri L, Soda P, Pechenizkiy M, Iannello G (2016) A survey on using domain and contextual knowledge for human activity recognition in video streams. Expert Syst Appl 63:97
130. Oord Avd, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks, arXiv:1601.06759
131. Paulson B, Cummings D, Hammond T (2011) Object interaction detection using hand posture cues in an office setting. Int J Hum-Comput Stud 69(1-2):19
132. Pham HH, Khoudour L, Crouzil A, Zegers P, Velastin Carroza SA (2015) Video-based human action recognition using deep learning: a review

133. Pires IM, Garcia NM, Pombo N, Flórez-Revuelta F (2016) From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. Sensors 16(2):184
134. Pirsiavash H, Ramanan D (2014) Parsing videos of actions with segmental grammars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 612–619
135. Popoola OP, Wang K (2012) Video-based abnormal human behavior recognition—a review. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(6):865
136. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
137. Portet F, Vacher M, Golanski C, Roux C, Meillon B (2013) Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers Ubiquit Comput 17(1):127
138. Poulos A, Brown C, McCulloch D, Cole J (2017) Context-aware augmented reality object commands US Patent 9,791,921
139. Prati A, Shan C, Wang KIK, Sensors vision (2019) Networks From video surveillance to activity recognition and health monitoring. J Ambient Intell Smart Environ 11(1):5
140. Presti LL, La Cascia M (2016) 3D skeleton-based human action classification. A Surv Pattern Recogn 53:130
141. Rahmani H, Mahmood A, Huynh DQ, Mian A (2014) Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In: European conference on computer vision. Springer, pp 742–757
142. 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
143. Ramanathan M, Yau WY, Teoh EK (2014) Human action recognition with video data: research and evaluation challenges. IEEE Trans Hum-Mach Syst 44(5):650
144. Ramanathan V, Li C, Deng J, Han W, Li Z, Gu K, Song Y, Bengio S, Rosenberg C, Fei-Fei L (2015) Learning semantic relationships for better action retrieval in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1100–1109
145. Rautaray SS, Agrawal A (2015) Vision based hand gesture recognition for human computer interaction: A survey. Artif Intell Rev 43(1):1. https://doi.org/10.1007/s10462-012-9356-9
146. Ravanbakhsh M, Nabi M, Sangineto E, Marcenaro L, Regazzoni C, Sebe N (2017) Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE International Conference on Image Processing (ICIP) IEEE, pp 1577–1581
147. Ravi D, Wong C, Lo B, Yang GZ (2017) A deep learning approach to on-node sensor data analytics for mobile or wearable devices. IEEE J Biomed Health Inf 21(1):56
148. Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971
149. Richard A, Kuehne H, Gall J (2017) Action sets: Weakly supervised action segmentation without ordering constraints, arXiv:1706.00699
150. Rodriguez MD, Ahmed J, Shah M (2008) Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008. CVPR 2008. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 1–8
151. Rodríguez ND, Cué llar MP, Lilius J, Calvo-Flores MD (2014) A survey on ontologies for human behavior recognition. ACM Comput Surv (CSUR) 46(4):43
152. Rueda FM, Fink GA (2018) Learning attribute representation for human activity recognition, arXiv:1802.00761
153. Ruffieux S, Lalanne D, Mugellini E, Khaled OA (2014) A survey of datasets for human gesture recognition. In: International Conference on Human-Computer Interaction. Springer, pp 337–348
154. Ryoo MS, Aggarwal J (2009) Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: 2009 ieee 12th international conference on Computer vision. IEEE, pp 1593–1600
155. Ryoo MS, Aggarwal J (2010) Ut-interaction dataset, icpr contest on semantic description of human activities (sdha). In: IEEE International Conference on Pattern Recognition Workshops, vol 2, pp 4
156. Sabokrou M, Fathy M, Hoseini M (2016) Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electron Lett 52(13):1122
157. Sabokrou M, Fathy M, Hoseini M, Klette R (2015) Real-time anomaly detection and localization in crowded scenes. In: Proceedings of the IEEE CVPR Workshops, pp 56–62
158. Sabokrou M, Fathy M, Moayed Z, Klette R (2017) Fast and accurate detection and localization of abnormal behavior in crowded scenes. Mach Vis Appl 28(8):965
159. Sabokrou M, Fayyaz M, Fathy M, Klette R (2017) Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process 26(4):1992

160. Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R (2018) Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Comput Vis Image Underst 172:88
161. Sabokrou M, Khalooei M, Adeli E (2019) Self-supervised representation learning via neighborhood-relational encoding. International Conference on Computer Vision
162. Sabokrou M, Khalooei M, Fathy M, Adeli E (2018) Adversarially learned one-class classifier for novelty detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3379–3388
163. Sabokrou M, Pourreza M, Fayyaz M, Entezari R, Fathy M, Gall J, Adeli E (2018) Avid: Adversarial visual irregularity detection. In: Asian computer vision conference
164. Sagayam KM, Hemanth DJ (2017) Hand posture and gesture recognition techniques for virtual reality applications: a survey. Virt Real 21(2):91
165. Saha S, Singh G, Sapienza M, Torr PH, Cuzzolin F (2016) Deep learning for detecting multiple space-time action tubes in videos, arXiv:1608.01529
166. Samanta S, Chanda B (2014) Space-time facet model for human activity classification. IEEE Trans Multimed 16(6):1525
167. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: 2004. ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition, vol 3. IEEE, pp 32–36
168. Sebestyen G, Stoica I, Hangan A (2016) Human activity recognition and monitoring for elderly people. In: 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE, pp 341–347
169. Seidenari L, Varano V, Berretti S, Bimbo A, Pala P (2013) Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 479–485
170. Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019
171. Shao J, Kang K, Change Loy C, Wang X (2015) Deeply learned attributes for crowded scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4657–4666
172. Shao L, Ji L, Liu Y, Zhang J (2012) Human action segmentation and recognition via motion and shape analysis. Pattern Recogn Lett 33(4):438
173. Sharaf A, Torki M, Hussein ME, El-Saban M (2015) Real-time multi-scale action detection from 3d skeleton data. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 998–1005
174. Sharma A, Singh PK, Khurana P (2016) Analytical review on object segmentation and recognition. In: Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. IEEE, pp 524–530
175. Shi J, Tomasi C (1994) Good features to track. In: 1994 Proceedings of IEEE conference on computer vision and pattern recognition. IEEE, pp 593–600
176. Shi Y, Tian Y, Wang Y, Huang T (2017) Sequential deep trajectory descriptor for action recognition with three-stream cnn. IEEE Trans Multimed 19(7):1510
177. Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R (2013) Real-time human pose recognition in parts from single depth images. Commun ACM 56(1):116
178. Shu T, Xie D, Rothrock B, Todorovic S, Chun Zhu S (2015) Joint inference of groups, events and human roles in aerial videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4576–4584
179. Sigurdsson GA, Divvala SK, Farhadi A, Gupta A (2017) Asynchronous temporal fields for action recognition.. In: CVPR, vol 5, pp 7
180. Singh T, Vishwakarma DK (2019) Human activity recognition in video benchmarks: A survey. In: Advances in Signal Processing and Communication. Springer, pp 247–259
181. Singh T, Vishwakarma DK (2019) Video benchmarks of human action datasets: a review. Artif Intell Rev 52(2):1107
182. Song S, Lan C, Xing J, Zeng W, Liu J (2018) Spatio-temporal attention-based lstm networks for 3d action recognition and detection. IEEE Trans Image Process 27(7):3459
183. Soo Park H, Shi J (2015) Social saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4777–4785
184. Soomro K, Zamir AR (2014) Action recognition in realistic sports videos. In: Computer vision in sports. Springer, pp 181–208

185. Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv:1212.0402
186. Sreenu G, Durai MS (2019) Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data 6(1):48
187. Srivastava N, Mansimov E, Salakhudinov R (2015) Unsupervised learning of video representations using lstms. In: International conference on machine learning, pp 843–852
188. Stein S, McKenna SJ (2013) Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. ACM, pp 729–738
189. Subetha T, Chitrakala S (2016) A survey on human activity recognition from videos. In: 2016 International Conference on Information Communication and Embedded Systems (ICICES). IEEE, pp 1–7
190. Sun S, Kuang Z, Sheng L, Ouyang W, Zhang W (2018) Optical flow guided feature: a fast and robust motion representation for video action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 8
191. Tabia H, Gouiffes M, Lacassagne L, sur Yvette B (2012) Reconnaissance des activités humaines à partir des vecteurs de mouvement quantifiés
192. Tang K, Yao B, Fei-Fei L, Koller D (2013) Combining the right features for complex event recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2696–2703
193. Tapus A, Bandera A, Vazquez-Martin R, Calderita LV (2019) Perceiving the person and their interactions with the others for social robotics–a review. Pattern Recogn Lett 118:3
194. Tripathi RK, Jalal AS, Agrawal SC (2017) Suspicious human activity recognition: a review. Artificial Intelligence Review 50(2):1–57
195. Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: a survey. IEEE Trans Circ Syst Video Technol 18(11):1473
196. Ullah A, Muhammad K, Haq IU, Baik SW (2019) Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments. Futur Gener Comput Syst 96:386
197. Varol G, Laptev I, Schmid C (2018) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510
198. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington, CVPR '14, pp 588–595. https://doi.org/10.1109/CVPR.2014.82
199. Vemulapalli R, Chellapa R (2016) Rolling rotations for recognizing human actions from 3d skeletal data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4471–4479
200. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
201. Vishwakarma S, Agrawal A (2013) A survey on activity recognition and behavior understanding in video surveillance. Vis Comput 29(10):983
202. Vital JP, Faria DR, Dias G, Couceiro MS, Coutinho F, Ferreira NM (2017) Combining discriminative spatiotemporal features for daily life activity recognition using wearable motion sensing suit. Pattern Anal Appl 20(4):1179
203. Vrigkas M, Nikou C, Kakadiadis IA (2014) Classifying behavioral attributes using conditional random fields. In: Hellenic Conference on Artificial Intelligence. Springer, pp 95–104
204. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2:28
205. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
206. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, pp 1290–1297
207. Wang J, Liu Z, Wu Y, Yuan J (2014) Learning actionlet ensemble for 3d human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914
208. Wang L, Ge L, Li R, Fang Y (2017) Three-stream cnns for action recognition. Pattern Recogn Lett 92:33
209. Wang L, Hu W, Tan T (2003) Recent developments in human motion analysis. Pattern Recogn 36(3):585
210. Wang P, Li W, Ogunbona P, Wan J, Escalera S (2018) Rgb-d-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding

211. Wang S, Ma Z, Yang Y, Li X, Pang C, Hauptmann AG (2014) Semi-supervised multiple feature analysis for action recognition. IEEE Trans Multimed 16(2):289
212. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Understand 104(2-3):249
213. 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
214. Wenkai X, Lee EJ (2012) Continuous gesture trajectory recognition system based on computer vision. International Journal of Applied Mathematics and Information Sciences:339–346
215. Wu C, Zhang J, Sener O, Selman B, Savarese S, Saxena A (2017) Watch-n-patch: unsupervised learning of actions and relations. IEEE Trans Pattern Anal Mach Intell 40(2):467
216. Wu Q, Wang Z, Deng F, Chi Z, Feng DD (2013) Realistic human action recognition with multimodal feature selection and fusion. IEEE Trans Syst Man Cybern Syst 43(4):875
217. Wu Y, Huang TS (2000) View-independent recognition of hand postures. In: Cvpr. IEEE, pp 2088
218. Wu Y, Lin JY, Huang TS (2001) Capturing natural hand articulation. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol 2. IEEE, pp 426–432
219. Wu Y, Wang Y, Jia Y (2013) Adaptive diffusion flow active contours for image segmentation. Comput Vis Image Underst 117(10):1421
220. Xia L, Chen CC, Aggarwal J (2012) View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE computer society conference on Computer vision and pattern recognition workshops (CVPRW). IEEE, pp 20–27
221. Xing L, Qin-kun X (2018) Human action recognition using auto-encode and pnn neural network. Software Guide (1):4 1608.01529
222. Xu K, Qin Z, Wang G (2016) Recognize human activities from multi-part missing videos. In: IEEE International Conference on Multimedia and Expo,. ICME 2016, Seattle, pp 976–990. https://doi.org/10.1109/ICME.2016.7552941
223. Xu W, Lee EJ (2015) A novel method for hand posture recognition based on depth information descriptor. KSII Transactions on Internet & Information Systems 9(2)
224. Yao B, Jiang X, Khosla A, Lin AL, Guibas L, Fei-Fei L (2011) Human action recognition by learning bases of action attributes and parts. In: 2011 IEEE International Conference on Computer Vision (ICCV). IEEE, pp 1331–1338
225. Yao G, Lei T, Zhong J (2019) A review of convolutional-neural-network-based action recognition. Pattern Recogn Lett 118:14
226. Yao T, Wang Z, Xie Z, Gao J, Feng DD (2017) Learning universal multiview dictionary for human action recognition. Pattern Recogn 64:236
227. Ye M, Zhang Q, Wang L, Zhu J, Yang R, Gall J (2013) A survey on human motion analysis from depth data. In: Time-of-flight and depth imaging. sensors, algorithms, and applications. Springer, pp 149–187
228. Yu G, Yuan J (2015) Fast action proposals for human action detection and search. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1302–1311
229. Yudistira N, Kurita T (2018) Deep packet flow: Action recognition via multiresolution deep wavelet packet of local dense optical flows. Journal of Signal Processing Systems 42(6):1–17
230. Zhan B, Monekosso DN, Remagnino P, Velastin S, Xu LQ (2008) Crowd analysis: a survey. Mach Vis Appl 19(5-6):345
231. Zhang H, Zhang YX, Zhong B, Lei Q, Yang L, Du JX, Chen DS (2019) A comprehensive survey of vision-based human action recognition methods. Sensors 19(5):1005
232. Zhang J, Li W, Ogunbona P, Wang P, Tang C (2016) Rgb-d-based action recognition datasets: a survey. Pattern Recogn 60:86
233. Zhang Z (2012) Microsoft kinect sensor and its effect. IEEE Multimed 19(2):4
234. Zhang Z, Liu Y, Li A, Wang M (2014) A novel method for user-defined human posture recognition using kinect. In: 2014 7th International Congress on Image and Signal Processing (CISP). IEEE, pp 736–740
235. Zhu F, Shao L (2014) Weakly-supervised cross-domain dictionary learning for visual recognition. Int J Comput Vis 109(1-2):42
236. Zhu F, Shao L, Xie J, Fang Y (2016) From handcrafted to learned representations for human action recognition: a survey. Image Vis Comput 55:42
237. Ziaeefard M, Bergevin R (2015) Semantic human activity recognition: a literature review. Pattern Recogn 48(8):2329

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.