frontiers
in Robotics and AI

REVIEW
published: 16 November 2015
doi: 10.3389/frobt.2015.00028

A Review of Human Activity Recognition Methods

Michalis Vrigkas[1], *Christophoros Nikou*[1]* *and Ioannis A. Kakadiaris*[2]

[1] Department of Computer Science and Engineering, University of Ioannina, Ioannina, Greece, [2] Computational Biomedicine Laboratory, Department of Computer Science, University of Houston, Houston, TX, USA

Recognizing human activities from video sequences or still images is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In this work, we provide a detailed review of recent and state-of-the-art research advances in the field of human activity classification. We propose a categorization of human activity methodologies and discuss their advantages and limitations. In particular, we divide human activity classification methods into two large categories according to whether they use data from different modalities or not. Then, each of these categories is further analyzed into sub-categories, which reflect how they model human activities and what type of activities they are interested in. Moreover, we provide a comprehensive analysis of the existing, publicly available human activity classification datasets and examine the requirements for an ideal human activity recognition dataset. Finally, we report the characteristics of future research directions and present some open issues on human activity recognition.

Keywords: human activity recognition, activity categorization, activity datasets, action representation, review, survey

OPEN ACCESS

Edited by:
Venkatesh Babu Radhakrishnan,
Indian Institute of Science, India

Reviewed by:
Stefano Berretti,
University of Florence, Italy
Xinlei Chen,
Carnegie Mellon University, USA

****Correspondence:***
Christophoros Nikou
cnikou@cs.uoi.gr

Specialty section:
This article was submitted to Vision Systems Theory, Tools and Applications, a section of the journal Frontiers in Robotics and AI

Received: *09 July 2015*
Accepted: *29 October 2015*
Published: *16 November 2015*

Citation:
Vrigkas M, Nikou C and Kakadiaris IA (2015) A Review of Human Activity Recognition Methods. Front. Robot. AI 2:28. doi: 10.3389/frobt.2015.00028

1. INTRODUCTION

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations. Because it provides information about the identity of a person, their personality, and psychological state, it is difficult to extract. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning. As a result of this research, many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system.

Among various classification techniques two main questions arise: "What action?" (i.e., the recognition problem) and "Where in the video?" (i.e., the localization problem). When attempting to recognize human activities, one must determine the kinetic states of a person, so that the computer can efficiently recognize this activity. Human activities, such as "walking" and "running," arise very naturally in daily life and are relatively easy to recognize. On the other hand, more complex activities, such as "peeling an apple," are more difficult to identify. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event (Gupta and Davis, 2007).

Most of the work in human activity recognition assumes a figure-centric scene of uncluttered background, where the actor is free to perform an activity. The development of a fully automated human activity recognition system, capable of classifying a person's activities with low error, is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting and appearance, and frame resolution. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. Moreover, intra- and interclass similarities make the problem amply challenging. That is, actions within the same class may be expressed by different people with different body movements, and actions between different classes may be difficult to distinguish as they may be represented by similar information. The way that humans perform an activity depends on their habits, and this makes the problem of identifying the underlying activity quite difficult to determine. Also, the construction of a visual model for learning and analyzing human movements in real time with inadequate benchmark datasets for evaluation is challenging tasks.

To overcome these problems, a task is required that consists of three components, namely: (i) background subtraction (Elgammal et al., 2002; Mumtaz et al., 2014), in which the system attempts to separate the parts of the image that are invariant over time (background) from the objects that are moving or changing (foreground); (ii) human tracking, in which the system locates human motion over time (Liu et al., 2010; Wang et al., 2013; Yan et al., 2014); and (iii) human action and object detection (Pirsiavash and Ramanan, 2012; Gan et al., 2015; Jainy et al., 2015), in which the system is able to localize a human activity in an image.

The goal of human activity recognition is to examine activities from video sequences or still images. Motivated by this fact, human activity recognition systems aim to correctly classify input data into its underlying activity category. Depending on their complexity, human activities are categorized into: (i) gestures; (ii) atomic actions; (iii) human-to-object or human-to-human interactions; (iv) group actions; (v) behaviors; and (vi) events. **Figure 1** visualizes the decomposition of human activities according to their complexity.

Gestures are considered as primitive movements of the body parts of a person that may correspond to a particular action of this person (Yang et al., 2013). Atomic actions are movements of a person describing a certain motion that may be part of more complex activities (Ni et al., 2015). Human-to-object or human-to-human interactions are human activities that involve two or more persons or objects (Patron-Perez et al., 2012). Group actions are activities performed by a group or persons (Tran et al., 2014b). Human behaviors refer to physical actions that are associated with the emotions, personality, and psychological state of the individual (Martinez et al., 2014). Finally, events are high-level activities that describe social actions between individuals and indicate the intention or the social role of a person (Lan et al., 2012a).

The rest of the paper is organized as follows: in Section 2, a brief review of previous surveys is presented. Section 3 presents the proposed categorization of human activities. In Sections 4 and 5, we review various human activity recognition methods and analyze the strengths and weaknesses of each category separately. In Section 6, we provide a categorization of human activity classification datasets and discuss some future research directions. Finally, conclusions are drawn in Section 7.

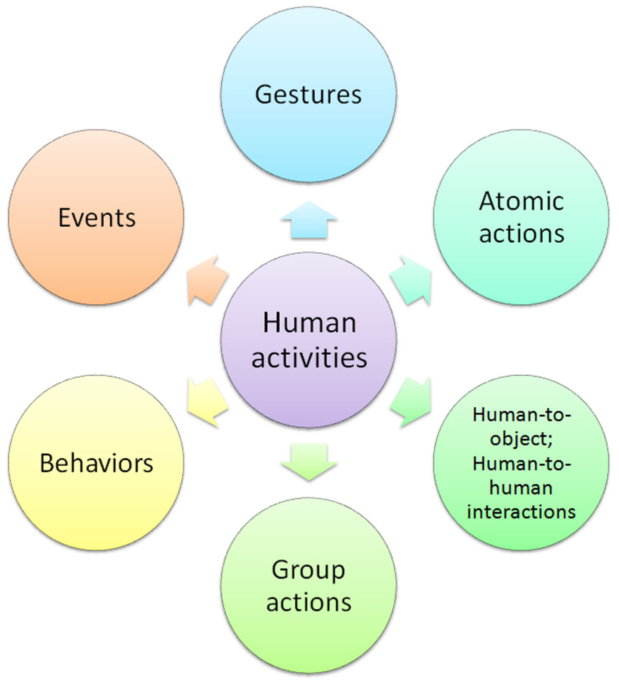

FIGURE 1 | Decomposition of human activities.

2. PREVIOUS SURVEYS AND TAXONOMIES

There are several surveys in the human activity recognition literature. Gavrila (1999) separated the research in 2D (with and without explicit shape models) and 3D approaches. In Aggarwal and Cai (1999), a new taxonomy was presented focusing on human motion analysis, tracking from single view and multiview cameras, and recognition of human activities. Similar in spirit to the previous taxonomy, Wang et al. (2003) proposed a hierarchical action categorization hierarchy. The survey of Moeslund et al. (2006) mainly focused on pose-based action recognition methods and proposed a fourfold taxonomy, including initialization of human motion, tracking, pose estimation, and recognition methods.

A fine separation between the meanings of "action" and "activity" was proposed by Turaga et al. (2008), where the activity recognition methods were categorized according to their degree of activity complexity. Poppe (2010) characterized human activity recognition methods into two main categories, describing them as "top-down" and "bottom-up." On the other hand, Aggarwal and Ryoo (2011) presented a tree-structured taxonomy, where the human activity recognition methods were categorized into two big sub-categories, the "single layer" approaches and the "hierarchical" approaches, each of which have several layers of categorization.

Modeling 3D data is also a new trend, and it was extensively studied by Chen et al. (2013b) and Ye et al. (2013). As the human body consists of limbs connected with joints, one can model these parts using stronger features, which are obtained from depth cameras, and create a 3D representation of the human body, which is more informative than the analysis of 2D activities carried out in the image plane. Aggarwal and Xia (2014) recently presented a categorization of human activity recognition methods from 3D stereo and motion capture systems with the main focus on methods that exploit 3D depth data. To this end, Microsoft Kinect has played a significant role in motion capture of articulated body skeletons using depth sensors.

Although much research has been focused on human activity recognition systems from video sequences, human activity recognition from static images remains an open and very challenging task. Most of the studies of human activity recognition are associated with facial expression recognition and/or pose estimation techniques. Guo and Lai (2014) summarized all the methods for human activity recognition from still images and categorized them into two big categories according to the level of abstraction and the type of features each method uses.

Jaimes and Sebe (2007) proposed a survey for multimodal human computer interaction focusing on affective interaction methods from poses, facial expressions, and speech. Pantic and Rothkrantz (2003) performed a complete study in human affective state recognition methods that incorporate non-verbal multimodal cues, such as facial and vocal expressions. Pantic et al. (2006) studied several state-of-the-art methods of human behavior recognition including affective and social cues and covered many open computational problems and how they can be efficiently incorporated into a human-computer interaction system. Zeng et al. (2009) presented a review of state-of-the-art affective recognition methods that use visual and audio cues for recognizing spontaneous affective states and provided a list of related datasets for human affective expression recognition. Bousmalis et al. (2013a) proposed an analysis of non-verbal multimodal (i.e., visual and auditory cues) behavior recognition methods and datasets for spontaneous agreements and disagreements. Such social attributes may play an important role in analyzing social behaviors, which are the key to social engagement. Finally, a thorough analysis of the ontologies for human behavior recognition from the viewpoint of data and knowledge representation was presented by Rodríguez et al. (2014).

Table 1 summarizes the previous surveys on human activity and behavior recognition methods sorted by chronological order. Most of these reviews summarize human activity recognition methods, without providing the strengths and the weaknesses of each category in a concise and informative way. Our goal is not only to present a new classification for the human activity recognition methods but also to compare different state-of-the-art studies and understand the advantages and disadvantages of each method.

TABLE 1 | Summary of previous surveys.

Authors and year	Area of interest
Aggarwal and Cai (1999)	Human motion analysis and tracking from single and multiview data
Gavrila (1999)	Shape model analysis from 2D and 3D data
Pantic and Rothkrantz (2003)	Multimodal human affective state recognition
Wang et al. (2003)	Human detection, tracking, and activity recognition
Moeslund et al. (2006)	Motion initialization, tracking, pose estimation, and recognition
Pantic et al. (2006)	Investigation of affective and social behaviors for human-computer interactions
Jaimes and Sebe (2007)	Multimodal affective interaction analysis for human-computer interactions
Turaga et al. (2008)	Categorization of actions and activities according to their complexity
Zeng et al. (2009)	Audio-visual affective recognition analysis
Poppe (2010)	Action classification according to global or local representation of data
Aggarwal and Ryoo (2011)	Gestures, human activities, actions, and interactions analysis
Bousmalis et al. (2013a)	Audio-visual behavior analysis of spontaneous agreements and disagreements
Chen et al. (2013b)	Human body part motion analysis from depth image data
Ye et al. (2013)	Human activity analysis from skeletal poses using depth data
Aggarwal and Xia (2014)	Human activity analysis from stereo, motion capture, and depth sensors 3D data
Guo and Lai (2014)	Understanding human activities from still images
Rodríguez et al. (2014)	Representation of human behavior ontologies from knowledge-based techniques

3. HUMAN ACTIVITY CATEGORIZATION

The human activity categorization problem has remained a challenging task in computer vision for more than two decades. Previous works on characterizing human behavior have shown great potential in this area. First, we categorize the human activity recognition methods into two main categories: (i) *unimodal* and (ii) *multimodal* activity recognition methods according to the nature of sensor data they employ. Then, each of these two categories is further analyzed into sub-categories depending on how they model human activities. Thus, we propose a hierarchical classification of the human activity recognition methods, which is depicted in **Figure 2**.

Unimodal methods represent human activities from data of a single modality, such as images, and they are further categorized as: (i) *space-time*, (ii) *stochastic*, (iii) *rule-based*, and (iv) *shape-based methods*.

Space-time methods involve activity recognition methods, which represent human activities as a set of spatiotemporal features (Shabani et al., 2011; Li and Zickler, 2012) or trajectories (Li et al., 2012; Vrigkas et al., 2013). Stochastic methods recognize activities by applying statistical models to represent human actions (e.g., hidden Markov models) (Lan et al., 2011; Iosifidis et al., 2012a). Rule-based methods use a set of rules to describe human activities (Morariu and Davis, 2011; Chen and Grauman, 2012). Shape-based methods efficiently represent activities with high-level reasoning by modeling the motion of human body parts (Sigal et al., 2012b; Tran et al., 2012).

Multimodal methods combine features collected from different sources (Wu et al., 2013) and are classified into three categories: (i) *affective*, (ii) *behavioral*, and (iii) *social networking methods*.

Affective methods represent human activities according to emotional communications and the affective state of a person

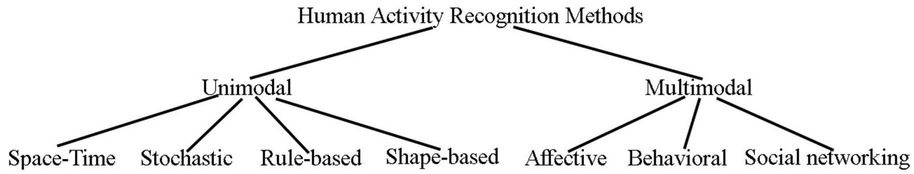

FIGURE 2 | Proposed hierarchical categorization of human activity recognition methods.

(Liu et al., 2011b; Martinez et al., 2014). Behavioral methods aim to recognize behavioral attributes, non-verbal multimodal cues, such as gestures, facial expressions, and auditory cues (Song et al., 2012a; Vrigkas et al., 2014b). Finally, social networking methods model the characteristics and the behavior of humans in several layers of human-to-human interactions in social events from gestures, body motion, and speech (Patron-Perez et al., 2012; Marín-Jiménez et al., 2014).

Usually, the terms "activity" and "behavior" are used interchangeably in the literature (Castellano et al., 2007; Song et al., 2012a). In this survey, we differentiate between these two terms in the sense that the term "activity" is used to describe a sequence of actions that correspond to specific body motion. On the other hand, the term "behavior" is used to characterize both activities and events that are associated with gestures, emotional states, facial expressions, and auditory cues of a single person. Some representative frames that summarize the main human action classes are depicted in **Figure 3**.

4. UNIMODAL METHODS

Unimodal human activity recognition methods identify human activities from data of one modality. Most of the existing approaches represent human activities as a set of visual features extracted from video sequences or still images and recognize the underlying activity label using several classification models (Kong et al., 2014a; Wang et al., 2014). Unimodal approaches are appropriate for recognizing human activities based on motion features. However, the ability to recognize the underlying class only from motion is on its own a challenging task. The main problem is how we can ensure the continuity of the motion along time as an action occurs uniformly or non-uniformly within a video sequence. Some approaches use snippets of motion trajectories (Matikainen et al., 2009; Raptis et al., 2012), while others use the full length of motion curves by tracking the optical flow features (Vrigkas et al., 2014a).

We classify unimodal methods into four broad categories: (i) *space-time*, (ii) *stochastic*, (iii) *rule-based*, and (iv) *shape-based approaches*. Each of these sub-categories describes specific attributes of human activity recognition methods according to the type of representation each method uses.

4.1. Space-Time Methods

Space-time approaches focus on recognizing activities based on space-time features or on trajectory matching. They consider an activity in the 3D space-time volume, consisting of concatenation of 2D spaces in time. An activity is represented by a set of space-time features or trajectories extracted from a video sequence. **Figure 4** depicts an example of a space-time approach based on dense trajectories and motion descriptors (Wang et al., 2013).

A plethora of human activity recognition methods based on space-time representation have been proposed in the literature (Efros et al., 2003; Schuldt et al., 2004; Jhuang et al., 2007; Fathi and Mori, 2008; Niebles et al., 2008). A major family of methods relies on optical flow, which has proven to be an important cue. Efros et al. (2003) recognized human actions from low-resolution sports' video sequences using the nearest neighbor classifier, where humans are represented by windows of height of 30 pixels. The approach of Fathi and Mori (2008) was based on mid-level motion features, which are also constructed directly from optical flow features. Moreover, Wang and Mori (2011) employed motion features as input to hidden conditional random fields (HCRFs) (Quattoni et al., 2007) and support vector machine (SVM) classifiers (Bishop, 2006). Real time classification and prediction of future actions was proposed by Morris and Trivedi (2011), where an activity vocabulary is learned through a three-step procedure. Other optical flow-based methods which gained popularity were presented by Dalal et al. (2006), Chaudhry et al. (2009), and Lin et al. (2009). An invariant in translation and scaling descriptor was introduced by Oikonomopoulos et al. (2009). Spatiotemporal features based on B-splines are extracted in the optical flow field. To model this descriptor, a Bag-of-Words (BoW) technique is employed, whereas, classification of activities is performed using relevant vector machines (RVM) (Tipping, 2001).

The classification of a video sequence using local features in a spatiotemporal environment has also been given much focus. Schuldt et al. (2004) represented local events in a video using space-time features, while an SVM classifier was used to recognize an action. Gorelick et al. (2007) considered actions as 3D space-time silhouettes of moving humans. They took advantage of the Poisson equation solution to efficiently describe an action by using spectral clustering between sequences of features and applying nearest neighbor classification to characterize an action. Niebles et al. (2008) addressed the problem of action recognition by creating a codebook of space-time interest points. A hierarchical approach was followed by Jhuang et al. (2007), where an input video was analyzed into several feature descriptors depending on their complexity. The final classification was performed by a multiclass SVM classifier. Dollár et al. (2005) proposed spatiotemporal features based on cuboid descriptors. Instead of encoding human motion for action classification, Jainy et al. (2015) proposed to incorporate information from human-to-objects interactions and combined several datasets to transfer information from one dataset to another.

An action descriptor of histograms of interest points, relying on the work of Schuldt et al. (2004), was presented by Yan and

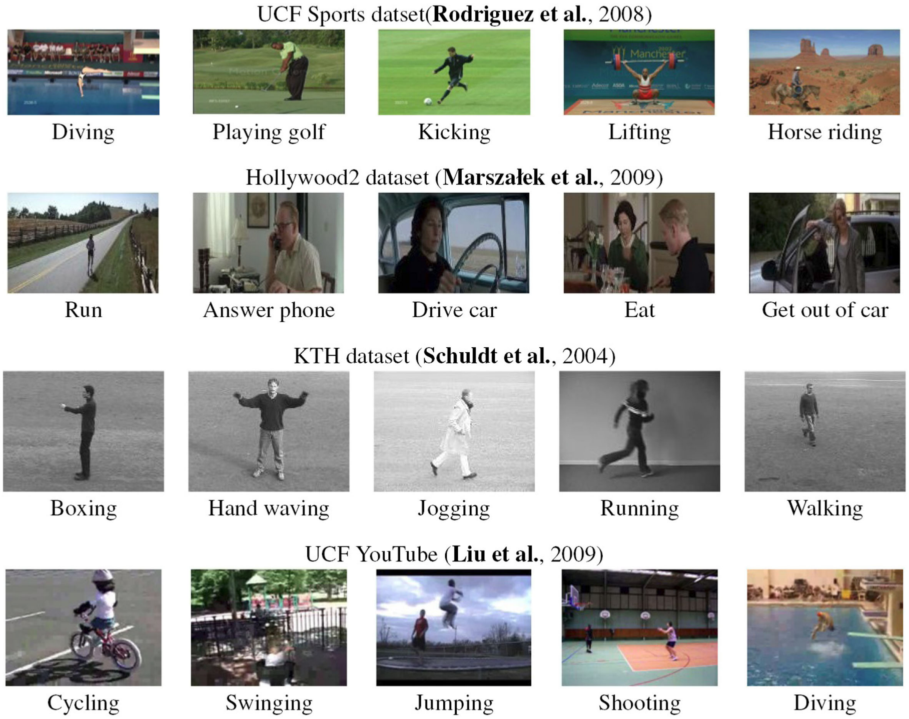

FIGURE 3 | Representative frames of the main human action classes for various datasets.

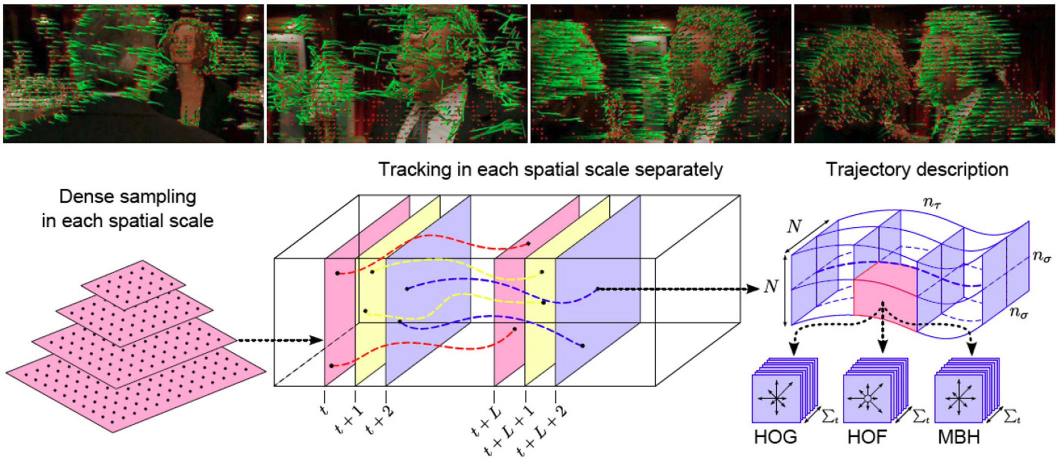

FIGURE 4 | Visualization of human actions with dense trajectories (top row). Example of a typical human space-time method based on dense trajectories (bottom row). First, dense feature sampling is performed for capturing local motion. Then, features are tracked using dense optical flow, and feature descriptors are computed (Wang et al., 2013).

Luo (2012). Random forests for action representation have also attracted widespread interest for action recognition Mikolajczyk and Uemura (2008) and Yao et al. (2010). Furthermore, the key issue of how many frames are required to recognize an action was addressed by Schindler and Gool (2008). Shabani et al. (2011) proposed a temporally asymmetric filtering for feature detection and activity recognition. The extracted features were more robust under geometric transformations than the features described by a Gabor filter (Fogel and Sagi, 1989). Sapienza et al. (2014) used a bag of local spatiotemporal volume features approach to recognize and localize human actions from weakly labeled video sequences using multiple instance learning.

The problem of identifying multiple persons simultaneously and performing action recognition was presented by Khamis et al. (2012). The authors considered that a person could first be detected by performing background subtraction techniques. Based on the histograms of oriented Gaussians, Dalal and Triggs (2005) were able to detect humans, whereas classification of actions was made by training an SVM classifier. Wang et al. (2011b) performed human activity recognition by associating the context between interest points based on the density of all features observed. A multiview activity recognition method was presented by Li and Zickler (2012), where descriptors from different views were connected together to construct a new augmented feature that contains the transition between the different views. Multiview action recognition has also been studied by Rahmani and Mian (2015). A non-linear knowledge transfer model based on deep learning was proposed for mapping action information from multiple camera views into one single view. However, their method is computationally expensive as it requires a two-step sequential learning phase prior to the recognition step for analyzing and fusing the information of multiviews.

Tian et al. (2013) employed spatiotemporal volumes using a deformable part model to train an SVM classifier for recognizing sport activities. Similar in spirit, the work of Jain et al. (2014) used a 3D space-time volume representation of human actions obtained from super-voxels to understand sport activities. They used an agglomerative approach to merge super-voxels that share common attributes and localize human activities. Kulkarni et al. (2015) used a dynamic programing approach to recognize sequences of actions in untrimmed video sequences. A per-frame time-series representation of each video and a template representation of each action were proposed, whereas dynamic time warping was used to sequence alignment.

Samanta and Chanda (2014) proposed a novel representation of human activities using a combination of spatiotemporal features and a facet model (Haralick and Watson, 1981), while they used a 3D Haar wavelet transform and higher order time derivatives to describe each interest point. A vocabulary was learned from these features and SVM was used for classification. Jiang et al. (2013) used a mid-level feature representation of video sequences using optical flow features. These features were clustered using K-means to build a hierarchical template tree representation of each action. A tree search algorithm was used to identify and localize the corresponding activity in test videos. Roshtkhari and Levine (2013) also proposed a hierarchical representation of video sequences for recognizing atomic actions by building a codebook of spatiotemporal volumes. A probe video sequence was classified into its underlying activity according to its similarity with each representation in the codebook.

Earlier approaches were based on describing actions by using dense trajectories. The work of Le et al. (2011) discovered the action label in an unsupervised manner by learning features directly from video data. A high-level representation of video sequences, called "action bank," was presented by Sadanand and Corso (2012). Each video was represented by a set of action descriptors, which were put in correspondence. The final classification was performed by an SVM classifier. Yan and Luo (2012) also proposed a novel action descriptor based on spatial temporal interest points (STIP) (Laptev, 2005). To avoid overfitting, they proposed a novel classification technique combining Adaboost and sparse representation algorithms. Wu et al. (2011) used visual features and Gaussian mixture models (GMM) (Bishop, 2006) to efficiently represent the spatiotemporal context distributions between the interest points at several space and time scales. The underlying activity was represented by a set of features extracted by the interest points over the video sequence. A new type of feature called the "hankelet" was presented by Li et al. (2012). This type of feature, which was formed by short tracklets, along with a BoW approach, was able to recognize actions under different viewpoints without requiring any camera calibration.

The work of Vrigkas et al. (2014a) focused on recognizing human activities by representing a human action with a set of clustered motion trajectories. A Gaussian mixture model was used to cluster the motion trajectories, and the action labeling was performed using a nearest neighbor classification scheme. Yu et al. (2012) proposed a propagative point-matching approach using random projection trees, which can handle unlabeled data in an unsupervised manner. Jain et al. (2013) used motion compensation techniques to recognize atomic actions. They also proposed a new motion descriptor called "divergence-curl-shear descriptor," which is able to capture the hidden properties of flow patterns in video sequences. Wang et al. (2013) used dense optical flow trajectories to describe the kinematics of motion patterns in video sequences. However, several intraclass variations caused by missing data, partial occlusion, and the sort duration of actions in time may harm the recognition accuracy. Ni et al. (2015) discovered the most discriminative groups of similar dense trajectories for analyzing human actions. Each group was assigned a learned weight according to its importance in motion representation.

An unsupervised method for learning human activities from short tracklets was proposed by Gaidon et al. (2014). They used a hierarchical clustering algorithm to represent videos with an unordered tree structure and compared all tree-clusters to identity the underlying activity. Raptis et al. (2012) proposed a mid-level approach extracting spatiotemporal features and constructing clusters of trajectories, which could be considered as candidates of an action. Yu and Yuan (2015) extracted bounding box candidates from video sequences, where each candidate may contain human motion. The most significant action paths were estimated by defining an action score. Due to the large spatiotemporal redundancy in videos, many candidates may overlap. Thus, estimation of the maximum set coverage was applied to address this problem. However, the maximum set coverage problem is NP-hard, and thus the estimation requires approximate solutions.

An approach that exploits the temporal information encoded in video sequences was introduced by Li et al. (2011). The temporal data were encoded into a trajectory system, which measures the similarity between activities and computes the angle between the associated subspaces. A method that tracks features and produces a number of trajectory snippets was proposed by Matikainen et al. (2009). The trajectories were clustered by an SVM classifier. Motion features were extracted from a video sequence by Messing et al. (2009). These features were tracked with respect to their velocities, and a generative mixture model was employed to learn the velocity history of these trajectories and classify each video clip. Tran et al. (2014a) proposed a scale and shape invariant method for localizing complex spatiotemporal events in video sequences. Their method was able to relax the tight constraints of bounding box tracking, while they used a sliding window technique to track spatiotemporal paths maximizing the summation score.

An algorithm that may recognize human actions in 3D space by a multicamera system was introduced by Holte et al. (2012a). It was based on the synergy of 3D space and time to construct a 4D descriptor of spatial temporal interest points and a local description of 3D motion features. The BoW technique was used to form a vocabulary of human actions, whereas agglomerative information bottleneck and SVM were used for action classification. Zhou and Wang (2012) proposed a new representation of local spatiotemporal cuboids for action recognition. Low-level features were encoded and classified via a kernelized SVM classifier, whereas a classification score denoted the confidence that a cuboid belongs to an atomic action. The new feature could act as complementary material to the low-level feature. The work of Sanchez-Riera et al. (2012) recognized human actions using stereo cameras. Based on the technique of BoW, each action was presented by a histogram of visual words, whereas their approach was robust to background clutter.

The problem of temporal segmentation and event recognition was examined by Hoai et al. (2011). Action recognition was performed by a supervised learning algorithm. Satkin and Hebert (2010) explored the effectiveness of video segmentation by discovering the most significant portions of videos. In the sense of video labeling, the study of Wang et al. (2012b) leveraged the shared structural analysis for activity recognition. The correct annotation was given in each video under a semisupervised scheme. Bag-of-video words have become very popular. Chakraborty et al. (2012) proposed a novel method applying surround suppression. Human activities were represented by bag-of-video words constructed from spatial temporal interest points by suppressing the background features and building a vocabulary of visual words. Guha and Ward (2012) employed a technique of sparse representations for human activity recognition. An overcomplete dictionary was constructed using a set of spatiotemporal descriptors. Classification over three different dictionaries was performed.

Seo and Milanfar (2011) proposed a method based on space-time locally adaptive regression kernels and the matrix cosine measure. They extracted features from space-time descriptors and compared them against features of the target video. A vocabulary based approach has been proposed by Kovashka and Grauman (2010). The main idea is to find the neighboring features around the detected interest points, quantize them, and form a vocabulary. Ma et al. (2015) extracted spatiotemporal segments from video sequences that correspond to whole or part human motion and constructed a tree-structured vocabulary of similar actions. Fernando et al. (2015) learned to arrange human actions in chronological order in an unsupervised manner by exploiting temporal ordering in video sequences. Relevant information was summarized together through a ranking learning framework.

The main disadvantage of using a global representation, such as optical flow, is the sensitivity to noise and partial occlusions. Space-time approaches can hardly recognize actions when more than one person is present in a scene. Nevertheless, space-time features focus mainly on local spatiotemporal information. Moreover, the computation of these features produces sparse and varying numbers of detected interest points, which may lead to low repeatability. However, background subtraction can help overcome this limitation.

Low-level features usually used with a fixed length feature vector (e.g., Bag-of-Words) failed to be associated with high-level events. Trajectory-based methods face the problem of human body detection and tracking, as these are still open issues. Complex activities are more difficult to recognize when space-time feature based approaches are employed. Furthermore, viewpoint invariance is another issue that these approaches have difficulty in handling.

4.2. Stochastic Methods

In recent years, there has been a tremendous growth in the amount of computer vision research aimed at understanding human activity. There has been an emphasis on activities, where the entity to be recognized may be considered as a stochastically predictable sequence of states. Researchers have conceived and used many stochastic techniques, such as hidden Markov model (HMMs) (Bishop, 2006) and hidden conditional random fields (HCRFs) (Quattoni et al., 2007), to infer useful results for human activity recognition.

Robertson and Reid (2006) modeled human behavior as a stochastic sequence of actions. Each action was described by a feature vector, which combines information about position, velocity, and local descriptors. An HMM was employed to encode human actions, whereas recognition was performed by searching for image features that represent an action. Pioneering this task, Wang and Mori (2008) were among the first to propose HCRFs for the problem of activity recognition. A human action was modeled as a configuration of parts of image observations. Motion features were extracted forming a BoW model. Activity recognition and localization via a figure-centric model was presented by Lan et al. (2011). Human location was treated as a latent variable, which was extracted from a discriminative latent variable model by simultaneous recognition of an action. A real-time algorithm that models human interactions was proposed by Oliver et al. (2000). The algorithm was able to detect and track a human movement, forming a feature vector that describes the motion. This vector was given as input to an HMM, which was used for action classification. Song et al. (2013) considered that human action sequences of various temporal resolutions. At each level of abstraction, they learned a hierarchical model with latent variables to group similar semantic attributes of each layer. Representative stochastic models are presented in **Figure 5**.

A multiview person identification was presented by Iosifidis et al. (2012a). Fuzzy vector quantization and linear discriminant analysis were employed to recognize a human activity. Huang et al. (2011) presented a boosting algorithm called LatentBoost. The authors trained several models with latent variables to recognize human actions. A stochastic modeling of human activities on a shape manifold was introduced by Yi et al. (2012). A human activity was extracted as a sequence of shapes, which is considered as one realization of a random process on a manifold. The piecewise Brownian motion was used to model human activity on the respective manifold. Wang et al. (2014) proposed a semisupervised framework for recognizing human actions combining different visual features. All features were projected onto a common subspace, and a boosting technique was employed to recognize human actions from labeled and unlabeled data.

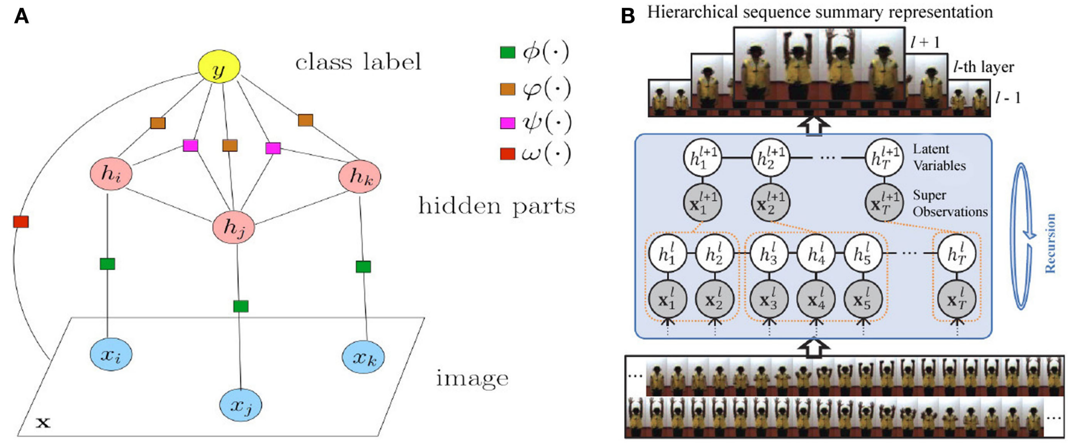
A
y
class label
🟩 $\phi(\cdot)$
🟧 $\varphi(\cdot)$
🟪 $\psi(\cdot)$
🟥 $\omega(\cdot)$
h_i
h_j
h_k
hidden parts
x_i
x_j
x_k
image
$\mathbf{x}$
B Hierarchical sequence summary representation
$l+1$
l-th layer
$l-1$
h_1^{l+1} h_2^{l+1} ... h_T^{l+1}
Latent Variables
$\mathbf{x}_1^{l+1}$ $\mathbf{x}_2^{l+1}$ $\mathbf{x}_T^{l+1}$
Super Observations
h_1^l h_2^l h_3^l h_4^l h_5^l ... h_T^l
$\mathbf{x}_1^l$ $\mathbf{x}_2^l$ $\mathbf{x}_3^l$ $\mathbf{x}_4^l$ $\mathbf{x}_5^l$ $\mathbf{x}_T^l$
Recursion

FIGURE 5 | Representative stochastic approaches for action recognition. **(A)** Factorized HCRF model used by Wang and Mori (2008). Circle nodes correspond to variables, and square nodes correspond to factors. **(B)** Hierarchical latent discriminative model proposed by Song et al. (2013).

Yang et al. (2013) proposed an unsupervised method for recognizing motion primitives for human action classification from a set of very few examples.

Sun and Nevatia (2013) treated video sequences as sets of short clips rather than a whole representation of actions. Each clip corresponded to a latent variable in an HMM model, while a Fisher kernel technique (Perronnin and Dance, 2007) was employed to represent each clip with a fixed length feature vector. Ni et al. (2014) decomposed the problem of complex activity recognition into two sequential sub-tasks with increasing granularity levels. First, the authors applied human-to-object interaction techniques to identify the area of interest, then used this context-based information to train a conditional random field (CRF) model (Lafferty et al., 2001) and identify the underlying action. Lan et al. (2014) proposed a hierarchical method for predicting future human actions, which may be considered as a reaction to a previous performed action. They introduced a new representation of human kinematic states, called "hierarchical movements," computed at different levels of coarse to fine-grained level granularity. Predicting future events from partially unseen video clips with incomplete action execution has also been studied by Kong et al. (2014b). A sequence of previously observed features was used as a global representation of actions and a CRF model was employed to capture the evolution of actions across time in each action class.

An approach for group activity classification was introduced by Choi et al. (2011). The authors were able to recognize activities such as a group of people talking or standing in a queue. The proposed scheme was based on random forests, which could select samples of spatiotemporal volumes in a video that characterize an action. A probabilistic Markov random field (MRF) (Prince, 2012) framework was used to classify and localize the activities in a scene. Lu et al. (2015) also employed a hierarchical MRF model to represent segments of human actions by extracting super-voxels from different scales and automatically estimated the foreground motion using saliency features of neighboring super-voxels.

The work of Wang et al. (2011a) focused on tracking dense sample points from video sequences using optical flow based on HCRFs for object recognition. Wang et al. (2012c) proposed a probabilistic model of two components. The first component modeled the temporal transition between action primitives to handle large variation in an action class, while the second component located the transition boundaries between actions. A hierarchical structure, which is called the sum-product network, was used by Amer and Todorovic (2012). The BoW technique encoded the terminal nodes, the sum nodes corresponded to mixtures of different subsets of terminals, and the product nodes represented mixtures of components.

Zhou and Zhang (2014) proposed a robust to background clutter, camera motion, and occlusions' method for recognizing complex human activities. They used multiple-instance formulation in conjunction with an MRF model and were able to represent human activities with a bag of Markov chains obtained from STIP and salient region feature selection. Chen et al. (2014) addressed the problem of identifying and localizing human actions using CRFs. The authors were able to distinguish between intentional actions and unknown motions that may happen in the surroundings by ordering video regions and detecting the actor of each action. Kong and Fu (2014) addressed the problem of human interaction classification from subjects that lie close to each other. Such a representation may be erroneous to partial occlusions and feature-to-object mismatching. To overcome this problem the authors proposed a patch-aware model, which learned regions of interacting subjects at different patch levels.

Shu et al. (2015) recognized complex video events and group activities from aerial shoots captured from unmanned aerial vehicles (UAVs). A preprocessing step prior to the recognition process was adopted to address several limitations of frame capturing, such as low resolution, camera motion, and occlusions. Complex events were decomposed into simpler actions and modeled using a spatiotemporal CRF graph. A video segmentation approach for video activities and a decomposition into smaller clips task that contained sub-actions was presented by Wu et al. (2015). The authors modeled the relation of consecutive actions by building a graphical model for unsupervised learning of the activity label from depth sensor data.

Often, human actions are highly correlated to the actor, who performs a specific action. Understanding both the actor and the action may be vital for real life applications, such as robot navigation and patient monitoring. Most of the existing works do not take into account the fact that a specific action may be performed in different manner by a different actor. Thus, a simultaneous inference of actors and actions is required. Xu et al. (2015) addressed these limitations and proposed a general probabilistic framework for joint actor-action understanding while they presented a new dataset for actor-action recognition.

There is an increasing interest in exploring human-object interaction for recognition. Moreover, recognizing human actions from still images by taking advantage of contextual information, such as surrounding objects, is a very active topic (Yao and Fei-Fei, 2010). These methods assume that not only the human body itself, but the objects surrounding it, may provide evidence of the underlying activity. For example, a soccer player interacts with a ball when playing soccer. Motivated by this fact, Gupta and Davis (2007) proposed a Bayesian approach that encodes object detection and localization for understanding human actions. Extending the previous method, Gupta et al. (2009) introduced spatial and functional constraints on static shape and appearance features and they were also able to identify human-to-object interactions without incorporating any motion information. Ikizler-Cinbis and Sclaroff (2010) extracted dense features and performed tracking over consecutive frames for describing both motion and shape information. Instead of explicitly using separate object detectors, they divided the frames into regions and treated each region as an object candidate.

Most of the existing probabilistic methods for human activity recognition may perform well and apply exact and/or approximate learning and inference. However, they are usually more complicated than non-parametric methods, since they use dynamic programing or computationally expensive HMMs for estimating a varying number of parameters. Due to their Markovian nature, they must enumerate all possible observation sequences while capturing the dependencies between each state and its corresponding observation only. HMMs treat features as conditionally independent, but this assumption may not hold for the majority of applications. Often, the observation sequence may be ignored due to normalization leading to the label bias problem (Lafferty et al., 2001). Thus, HMMs are not suitable for recognizing more complex events, but rather an event is decomposed into simpler activities, which are easier to recognize.

Conditional random fields, on the other hand, overcome the label bias problem. Most of the aforementioned methods do not require large training datasets, since they are able to model the hidden dynamics of the training data and incorporate prior knowledge over the representation of data. Although CRFs outperform HMMS in many applications, including bioinformatics, activity, and speech recognition, the construction of more complex models for human activity recognition may have good generalization ability but is rather impractical for real time applications due to the large number of parameter estimations and the approximate inference.

4.3. Rule-Based Methods

Rule-based approaches determine ongoing events by modeling an activity using rules or sets of attributes that describe an event. Each activity is considered as a set of primitive rules/attributes, which enables the construction of a descriptive model for human activity recognition.

Action recognition of complex scenes with multiple subjects was proposed by Morariu and Davis (2011). Each subject must follow a set of certain rules while performing an action. The recognition process was performed over basketball game videos, where the players were first detected and tracked, generating a set of trajectories that are used to create a set of spatiotemporal events. Based on the first-order logic and probabilistic approaches, such as Markov networks, the authors were able to infer which event has occurred. **Figure 6** summarizes their method using primitive rules for recognizing human actions. Liu et al. (2011a) addressed the problem of recognizing actions by a set of descriptive and discriminative attributes. Each attribute was associated with the characteristics describing the spatiotemporal nature of the activities. These attributes were treated as latent variables, which capture the degree of importance of each attribute for each action in a latent SVM approach.

A combination of activity recognition and localization was presented by Chen and Grauman (2012). The whole approach was based on the construction of a space-time graph using a high-level descriptor, where the algorithm seeks to find the optimal subgraph that maximizes the activity classification score (i.e., find the maximum weight subgraph, which in the general case is an NP-complete problem). Kuehne et al. (2014) proposed a structured temporal approach for daily living human activity recognition. The author used HMMs to model human actions as action units and then used grammatical rules to form a sequence of complex actions by combining different action units. When temporal grammars are used for action classification, the main problem consists in treating long video sequences due to the complexity of the models. One way to cope with this limitation is to segment video sequences into smaller clips that contain sub-actions, using a hierarchical approach (Pirsiavash and Ramanan, 2014). The generation of short description from video sequences (Vinyals et al., 2015) based on convolutional neural networks (CNN) (Ciresan et al., 2011) was also used for activity recognition (Donahue et al., 2015).

Intermediate semantic features representation for recognizing unseen actions during training were proposed (Wang and Mori, 2010). These intermediate features were learned during training, while parameter sharing between classes was enabled by capturing the correlations between frequently occurring low-level features (Akata et al., 2013). Learning how to recognize new classes that were not seen during training, by associating intermediate features and class labels, is a necessary aspect for transferring knowledge between training and test samples. This problem is generally known as zero-shot learning (Palatucci et al., 2009). Thus, instead of learning one classifier per attribute, a two-step classification method has been proposed by Lampert et al. (2009). Specific attributes are predicted from already learned classifiers and are mapped into a class-level score.

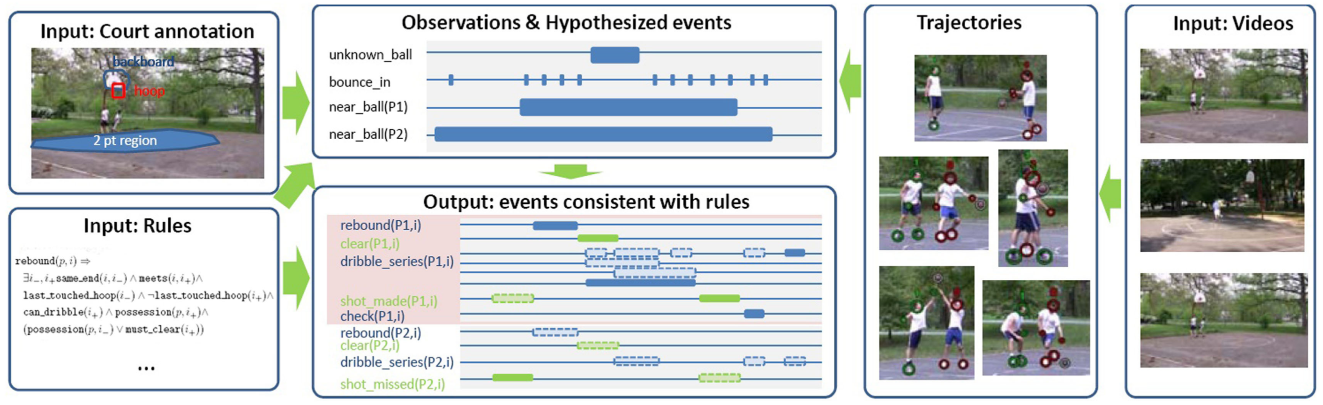

FIGURE 6 | Relation between primitive rules and human actions (Morariu and Davis, 2011).

Action classification from still images by learning semantic attributes was proposed by Yao et al. (2011). Attributes describe specific properties of human actions, while parts of actions, which were obtained from objects and human poses, were used as bases for learning complex activities. The problem of attribute-action association was reported by Zhang et al. (2013). The authors proposed a multitask learning approach Evgeniou and Pontil (2004) for simultaneously coping with low-level features and action-attribute relationships and introduced attribute regularization as a penalty term for handling irrelevant predictions. A robust to noise representation of attribute-based human action classification was proposed by Zhang et al. (2015). Sigmoid and Gaussian envelopes were incorporated into the loss function of an SVM classifier, where the outliers are eliminated during the optimization process. A GMM was used for modeling human actions, and a transfer ranking technique was employed for recognizing unseen classes. Ramanathan et al. (2015) were able to transfer semantic knowledge between classes to learn human actions from still images. The interaction between different classes was performed using linguistic rules. However, for high-level activities, the use of language priors is often not adequate, thus simpler and more explicit rules should be constructed.

Complex human activities cannot be recognized directly from rule-based approaches. Thus, decomposition into simpler atomic actions is applied, and then combination of individual actions is employed for the recognition of complex or simultaneously occurring activities. This limitation leads to constant feedback by the user of rule/attribute annotations of the training examples, which is time consuming and sensitive to errors due to subjective point of view of the user defined annotations. To overcome this drawback, several approaches employing transfer learning (Lampert et al., 2009; Kulkarni et al., 2014), multitask learning (Evgeniou and Pontil, 2004; Salakhutdinov et al., 2011), and semantic/discriminative attribute learning (Farhadi et al., 2009; Jayaraman and Grauman, 2014) were proposed to automatically generate and handle the most informative attributes for human activity classification.

4.4. Shape-Based Methods

Modeling of human pose and appearance has received a great response from researchers during the last decades. Parts of the human body are described in 2D space as rectangular patches and as volumetric shapes in 3D space (see **Figure 7**). It is well known that activity recognition algorithms based on the human silhouette play an important role in recognizing human actions. As a human silhouette consists of limbs jointly connected to each other, it is important to obtain exact human body parts from videos. This problem is considered as a part of the action recognition process. Many algorithms convey a wealth of information about solving this problem.

A major focus in action recognition from still images or videos has been made in the context of scene appearance (Thurau and Hlavac, 2008; Yang et al., 2010; Maji et al., 2011). More specifically, Thurau and Hlavac (2008) represented actions by histograms of pose primitives, and n-gram expressions were used for action classification. Also, Yang et al. (2010) combined actions and human poses together, treating poses as latent variables, to infer the action label in still images. Maji et al. (2011) introduced a representation of human poses, called the "poselet activation vector," which is defined by the 3D orientation of the head and torso and provided a robust representation of human pose and appearance. Moreover, action categorization based on modeling the motion of parts of the human body was presented by Tran et al. (2012), where a sparse representation was used to model and recognize complex actions. In the sense of template-matching techniques, Rodriguez et al. (2008) introduced the maximum average correlation height (MACH) filter, which was a method for capturing intraclass variabilities by synthesizing a single action MACH filter for a given action class. Sedai et al. (2013a) proposed a combination of shape and appearance descriptors to represent local features for human pose estimation. The different types of descriptors were fused at the decision level using a discriminative learning model. Nevertheless, identifying which body parts are most significant for recognizing complex human activities still remains a challenging task (Lillo et al., 2014). The classification model and some representative examples of the estimation of human pose are depicted in **Figure 8**.

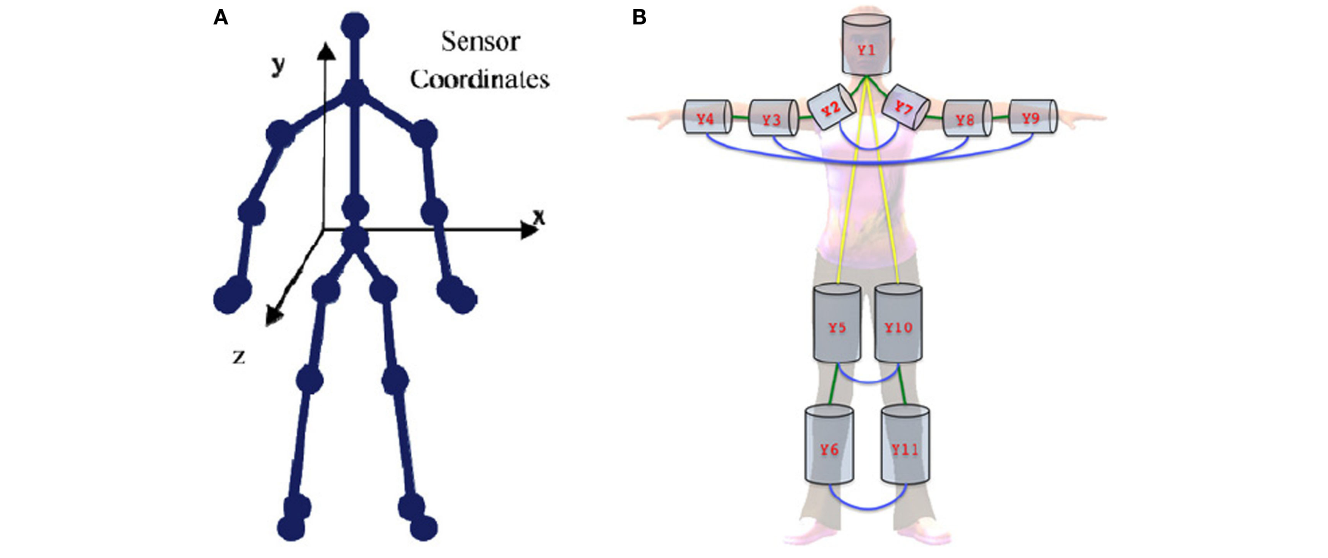

FIGURE 7 | Human body representations. **(A)** 2D skeleton model (Theodorakopoulos et al., 2014) and **(B)** 3D pictorial structure representation (Belagiannis et al., 2014).

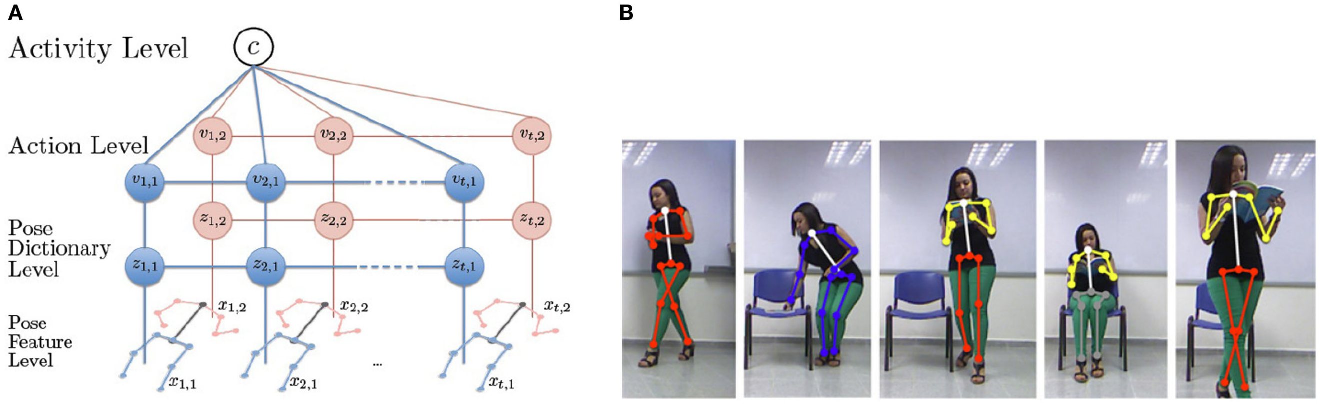

FIGURE 8 | Classification of actions from human poses (Lillo et al., 2014). **(A)** The discriminative hierarchical model for the recognition of human action from body poses. **(B)** Examples of correct human pose estimation of complex activities.

Ikizler and Duygulu (2007) modeled the human body as a sequence of oriented rectangular patches. The authors described a variation of BoW method called bag-of-rectangles. Spatially oriented histograms were formed to describe a human action, while the classification of an action was performed using four different methods, such as frame voting, global histogramming, SVM classification, and dynamic time warping (DTW) (Theodoridis and Koutroumbas, 2008). The study of Yao and Fei-Fei (2012) modeled human poses for human-object interactions by introducing a mutual context model. The types of human poses, as well as the spatial relationship between the different human parts, were modeled. Self organizing maps (SOM) (Kohonen et al., 2001) were introduced by Iosifidis et al. (2012b) for learning human body posture, in conjunction with fuzzy distances, to achieve time invariant action representation. The proposed algorithm was based on multilayer perceptrons, where each layer was fed by an associated camera, for view-invariant action classification. Human interactions were addressed by Andriluka and Sigal (2012). First, 2D human poses were estimated from pictorial structures from groups of humans and then each estimated structure was fitted into 3D space. To this end, several 2D human pose benchmarks have been proposed for the evaluation of articulated human pose estimation methods (Andriluka et al., 2014).

Action recognition using depth cameras was introduced by Wang et al. (2012a), where a new feature type called "local occupancy pattern" was also proposed. This feature was invariant to translation and was able to capture the relation between human body parts. The authors also proposed a new model for human actions called "actionlet ensemble model," which captured the intraclass variations and was robust to errors incurred by depth cameras. 3D human poses have been taken into consideration in recent years and several algorithms for human activity recognition have been developed. A recent review on 3D pose estimation and activity recognition was proposed by Holte et al. (2012b). The authors categorized 3D pose estimation approaches aimed at presenting multiview human activity recognition methods. The work of Shotton et al. (2011) modeled 3D human poses and performed human activity recognition from depth images by mapping the pose estimation problem into a simpler pixel-wise classification problem. Graphical models have been widely used in modeling 3D human poses. The problem of articulated 3D human pose estimation was studied by Fergie and Galata (2013), where the limitation of the mapping from the image feature space to the pose space was addressed using mixtures of Gaussian processes, particle filtering, and annealing (Sedai et al., 2013b). A combination of discriminative and generative models improved the estimation of human pose.

Multiview pose estimation was examined by Amin et al. (2013). The 2D poses for different sources were projected onto 3D space using a mixture of multiview pictorial structures models. Belagiannis et al. (2014) have also addressed the problem of multiview pose estimation. They constructed 3D body part hypotheses by triangulation of 2D pose detections. To solve the problem of body part correspondence between different views, the authors proposed a 3D pictorial structure representation based on a CRF model. However, building successful models for human pose estimation is not straightforward (Pishchulin et al., 2013). Combining both pose-specific appearance and the joint appearance of body parts helps to construct a more powerful representation of the human body. Deep learning has gained much attention for multisource human pose estimation (Ouyang et al., 2014) where the tasks of detection and estimation of human pose were jointly learned. Toshev and Szegedy (2014) have also used deep learning for human pose estimation. Their approach relies on using deep neural networks (DNN) (Ciresan et al., 2012) for representing cascade body joint regressors in a holistic manner.

Despite the vast development of pose estimation algorithms, the problem still remains challenging for real time applications. Jung et al. (2015) presented a method for fast estimation of human pose with 1,000 frames per second. To achieve such a high computational speed, the authors used random walk sub-sampling methods. Human body parts were handled as directional tree-structured representations and a regression tree was trained for each joint in the human skeleton. However, this method depends on the initialization of the random walk process.

Sigal et al. (2012b) addressed the multiview human-tracking problem where the modeling of 3D human pose consisted of a collection of human body parts. The motion estimation was performed by non-parametric belief propagation (Bishop, 2006). On the other hand, the work of Livne et al. (2012) explored the problem of inferring human attributes, such as gender, weight, and mood, by the scope of 3D pose tracking. Representing activities using trajectories of human poses is computationally expensive due to many degrees of freedom. To this end, efficient dimensionality reduction methods should be applied. Moutzouris et al. (2015) proposed a novel method for reducing dimensionality of human poses called "hierarchical temporal Laplacian eigenmaps" (HTLE). Moreover, the authors were able to estimate unseen poses using a hierarchical manifold search method.

Du et al. (2015) divided the human skeleton into five segments and used each of these parts to train a hierarchical neural network. The output of each layer, which corresponds to neighboring parts, is fused and fed as input to the next layer. However, this approach suffers from the problem of data association as parts of the human skeleton may vanish through the sequential layer propagation and back projection. Nie et al. (2015) also divided human pose into smaller mid-level spatiotemporal parts. Human actions were represented using a hierarchical AND/OR graph and dynamic programing was used to infer the class label. One disadvantage of this method is that it cannot deal with self-occlusions (i.e., overlapping parts of human skeleton).

A shared representation of human poses and visual information has also been explored (Ferrari et al., 2009; Singh and Nevatia, 2011; Yun et al., 2012). However, the effectiveness of such methods is limited by tracking inaccuracies in human poses and complex backgrounds. To this end, several kinematic and part-occlusion constraints for decomposing human poses into separate limbs have been explored to localize the human body (Cherian et al., 2014). Xu et al. (2012) proposed a mid-level representation of human actions by computing local motion volumes in skeletal points extracted from video sequences and constructed a codebook of poses for identifying the action. Eweiwi et al. (2014) reduced the required amount of pose data using a fixed length vector of more informative motion features (e.g., location and velocity) for each skeletal point. A partial least squares approach was used for learning the representation of action features, which is then fed into an SVM classifier.

Kviatkovsky et al. (2014) mixed shape and motion features for online action classification. The recognition processes could be applied in real time using the incremental covariance update and the on-demand nearest neighbor classification schemes. Rahmani et al. (2014) trained a random decision forest (RDF) (Ho, 1995) and applied a joint representation of depth information and 3D skeletal positions for identifying human actions in real time. A novel part-based skeletal representation for action recognition was introduced by Vemulapalli et al. (2014). The geometry between different body parts was taken into account, and a 3D representation of human skeleton was proposed. Human actions are treated as curves in the Lie group (Murray et al., 1994), and the classification was performed using SVM and temporal modeling approaches. Following a similar approach, Anirudh et al. (2015) represented skeletal joints as points on the product space. Shape features were represented as high-dimensional non-linear trajectories on a manifold to learn the latent variable space of actions. Fouhey et al. (2014) exploited the interaction between human actions and scene geometry to recognize human activities from still images using 3D skeletal representation and adopting geometric representation constraints of the scenes.

The problem of appearance-to-pose mapping for human activity understanding was studied by Urtasun and Darrell (2008). Gaussian processes were used as an online probabilistic regressor for this task using sparse representation of data for reducing computational complexity. Theodorakopoulos et al. (2014) have also employed sparse representation of skeletal data in the dissimilarity space for human activity recognition. In particular, human actions are represented by vectors of dissimilarities and a set of prototype actions is built. The recognition is performed into the dissimilarity space using sparse representation-based classification. A publicly available dataset (UPCV Action dataset) consisting of skeletal data of human actions was also proposed.

A common problem in estimating human pose is the high-dimensional space (i.e., each limb may have a large number of degrees of freedom that need to be estimated simultaneously). Action recognition relies heavily on the obtained pose estimations. The articulated human body is usually represented as a tree-like structure, thus locating the global position and tracking each limb separately is intrinsically difficult, since it requires exploration of a large state space of all possible translations and rotations of the human body parts in 3D space. Many approaches, which employ background subtraction (Sigal et al., 2012a) or assume fixed limb lengths and uniformly distributed rotations of body parts (Burenius et al., 2013), have been proposed to reduce the complexity of the 3D space.

Moreover, the association of human pose orientation with the poses extracted from different camera views is also a difficult problem due to similar body parts of different humans in each view. Mixing body parts of different views may lead to ambiguities because of the multiple candidates of each camera view and false positive detections. The estimation of human pose is also very sensitive to several factors, such as illumination changes, variations in view-point, occlusions, background clutter, and human clothing. Low-cost devices, such as Microsoft Kinect and other RGB-D sensors, which provide 3D depth data of a scene, can efficiently leverage these limitations and produce a relatively good estimation of human pose, since they are robust to illumination changes and texture variations (Gao et al., 2015).

5. MULTIMODAL METHODS

Recently, much attention has been focused on multimodal activity recognition methods. An event can be described by different types of features that provide more and useful information. In this context, several multimodal methods are based on feature fusion, which can be expressed by two different strategies: early fusion and late fusion. The easiest way to gain the benefits of multiple features is to directly concatenate features in a larger feature vector and then learn the underlying action (Sun et al., 2009). This feature fusion technique may improve recognition performance, but the new feature vector is of much larger dimension.

Multimodal cues are usually correlated in time, thus a temporal association of the underlying event and the different modalities is an important issue for understanding the data. In that context, audio-visual analysis is used in many applications not only for audio-visual synchronization (Lichtenauer et al., 2011) but also for tracking (Perez et al., 2004) and activity recognition (Wu et al., 2013). Multimodal methods are classified into three categories: (i) *affective methods*, (ii) *behavioral methods*, and (iii) *methods based on social networking*. Multimodal methods describe atomic actions or interactions that may correspond to affective states of a person with whom he/she communicates and depend on emotions and/or body movements.

5.1. Affective Methods

The core of emotional intelligence is understanding the mapping between a person's affective states and the corresponding activities, which are strongly related to the emotional state and communication of a person with other people (Picard, 1997). Affective computing studies model the ability of a person to express, recognize, and control his/her affective states in terms of hand gestures, facial expressions, physiological changes, speech, and activity recognition (Pantic and Rothkrantz, 2003). This research area is generally considered to be a combination of computer vision, pattern recognition, artificial intelligence, psychology, and cognitive science.

A key issue in affective computing is accurately annotated data. Ratings are one of the most popular affect annotation tools. However, this is challenging to obtain for real world situations, since affective events are expressed in a different manner by different persons or occur simultaneously with other activities and feelings. Preprocessing affective annotations may be detrimental for generating accurate and ambiguous affective models due to biased representations of affect annotation. To this end, a study on how to produce highly informative affective labels has been proposed by Healey (2011). Soleymani et al. (2012) investigated the properties of developing a user-independent emotion recognition system that is able to detect the most informative affective tags from electroencephalogram (EEG) signals, pupillary reflex, and bodily responses that correspond to video stimulus. Nicolaou et al. (2014) proposed a novel method based on probabilistic canonical correlation analysis (PCCA) (Klami and Kaski, 2008) and DTW for fusing multimodal emotional annotations and performing temporal aligning of sequences.

Liu et al. (2011b) associated multimodal features (i.e., textual and visual) for classifying affective states in still images. The authors argued that visual information is not adequate for understanding human emotions, and thus additional information that describes the image is needed. Dempster-Shafer theory (Shafer, 1976) was employed for fusing the different modalities, while SVM was used for classification. Hussain et al. (2011) proposed a framework for fusing multimodal psychological features, such as heart and facial muscle activity, skin response, and respiration, for detecting and recognizing affective states. AlZoubi et al. (2013) explored the effect of the affective feature variations over time on the classification of affective states.

Siddiquie et al. (2013) analyzed four different affective dimensions, such as activation, expectancy, power, and valence (Schuller et al., 2011). To this end, they proposed joint hidden conditional random Fields (JHCRF) as a new classification scheme to take advantage of the multimodal data. Furthermore, their method uses late fusion to combine audio and visual information together. This may lead to significant loss of the intermodality dependence, while it suffers from propagating the classification error across different levels of classifiers. Although their method could efficiently recognize the affective state of a person, the computational burden was high as JHCRFs require twice as many hidden variables as the traditional HCRFs when features represent two different modalities.

Nicolaou et al. (2011) proposed a regression model based on SVMs for regression (SVR) (Smola and Schölkopf, 2004) for continuous prediction of multimodal emotional states, using facial expression, shoulder gesture, and audio cues in terms of arousal and valence (**Figure 9**). Castellano et al. (2007) explored the dynamics of body movements to identify affective behaviors using time series of multimodal data. Martinez et al. (2014) presented a detailed review of learning methods for the classification of affective and cognitive states of computer game players. They analyzed the properties of directly using affect annotations in classification models, and proposed a method for transforming such annotations to build more accurate models.

Multimodal affect recognition methods in the context of neural networks and deep learning have generated considerable recent research interest (Ngiam et al., 2011). In a more recent study, Martinez et al. (2013) could efficiently extract and select the most informative multimodal features using deep learning to model emotional expressions and recognize the affective states of a person. They incorporated psychological signals into emotional states, such as relaxation, anxiety, excitement, and fun, and demonstrated that deep learning was able to extract more informative features than feature extraction on psychological signals.

Although the understanding of human activities may benefit from affective state recognition, the classification process is extremely challenging due to the semantic gap between the low-level features extracted from video frames and high-level concepts, such as emotions, that need to be identified. Thus, building strong models that can cope with multimodal data, such as gestures, facial expressions and psychological data, depends on the ability of the model to discover relations between different modalities and generate informative representation on affect annotations. Generating such information is not an easy task. Users cannot always express their emotion with words, and producing satisfactory and reliable ground truth that corresponds to a given training instance is quite difficult as it can lead to ambiguous and subjective labels. This problem becomes more prominent as human emotions are continuous acts in time, and variations in human actions may be confusing or lead to subjective annotations. Therefore, automatic affective recognition systems should reduce the effort for selecting the proper affective label to better assess human emotions.

5.2. Behavioral Methods

Recognizing human behaviors from video sequences is a challenging task for the computer vision community (Candamo et al., 2010). A behavior recognition system may provide information about the personality and psychological state of a person, and its applications vary from video surveillance to human-computer interaction. Behavioral approaches aim at recognizing behavioral attributes, non-verbal multimodal cues, such as gestures, facial expressions, and auditory cues. Factors that can affect human behavior may be decomposed into several components, including emotions, moods, actions, and interactions, with other people. Hence, the recognition of complex actions may be crucial for understanding human behavior. One important aspect of human behavior recognition is the choice of proper features, which can be used to recognize behavior in applications, such as gaming and physiology. A key challenge in recognizing human behaviors is to define specific emotional attributes for multimodal dyadic interactions (Metallinou and Narayanan, 2013). Such attributes may be descriptions of emotional states or cognitive states, such

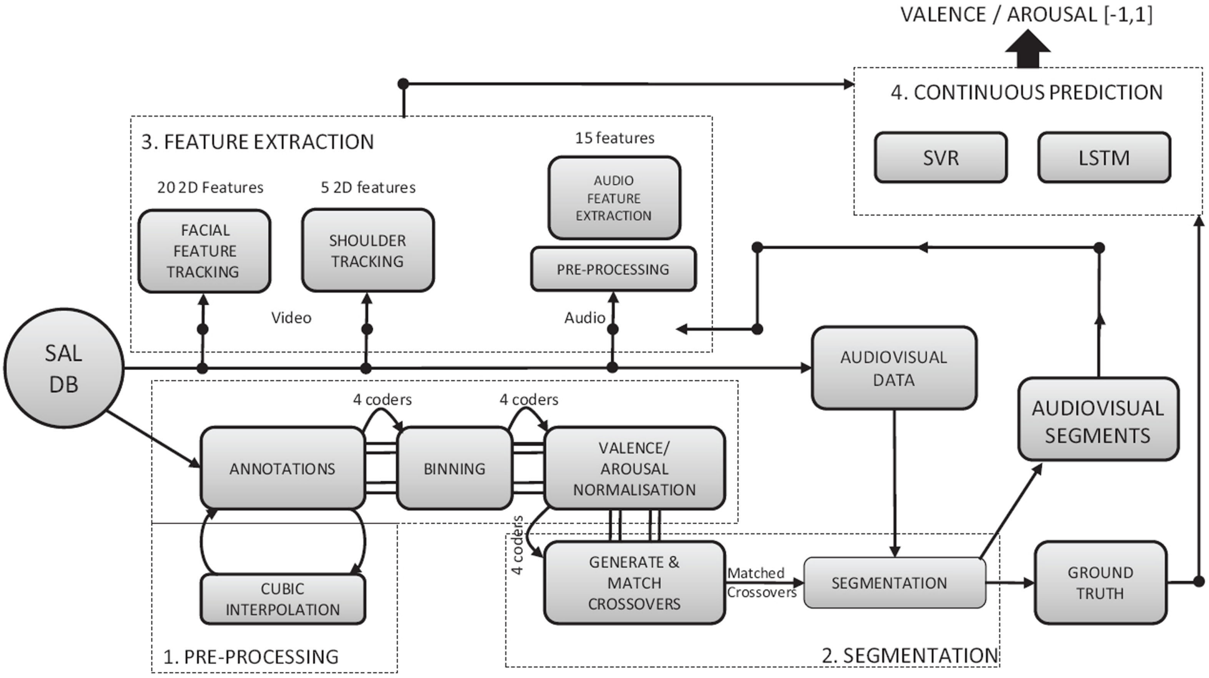

FIGURE 9 | Flow chart of multimodal emotion recognition. Emotions, facial expressions, shoulder gestures, and audio cues are combined for continuous prediction emotional states (Nicolaou et al., 2011).

as activation, valence, and engagement. A typical example of a behavior recognition system is depicted in **Figure 10**.

Audio-visual representation of human actions has gained an important role in human behavior recognition methods. Sargin et al. (2007) suggested a method for speaker identification integrating a hybrid scheme of early and late fusion of audio-visual features and used CCA (Hardoon et al., 2004) to synchronize the multimodal features. However, their method can cope with video sequences of frontal view only. Metallinou et al. (2008) proposed a probabilistic approach based on GMMs for recognizing human emotions in dyadic interactions. The authors took advantage of facial expressions as they can be expressed by the facial action coding system (FACS) (Ekman et al., 2002), which describes all possible facial expressions as a combination of action units (AU), and combines them with audio information, extracted from each participant, to identify their emotional state. Similarly, Chen et al. (2015) proposed a real-time emotion recognition system that modeled 3D facial expressions using random forests. The proposed method was robust to subjects' poses and changes in the environment.

Wu et al. (2010) proposed a human activity recognition system by taking advantage of the auditory information of the video sequences of the HOHA dataset (Laptev et al., 2008) and used late fusion techniques for combining audio and visual cues. The main disadvantage of this method is that it used different classifiers to separately learn the audio and visual context. Also, the audio information of the HOHA dataset contains dynamic backgrounds and the audio signal is highly diverse (i.e., audio shifts roughly from one event to another), which generates the need for developing audio feature selection techniques. Similar in spirit is the work of Wu et al. (2013), who used the generalized multiple kernel learning algorithm for estimating the most informative audio features. They applied fuzzy integral techniques to combine the outputs of two different SVM classifiers, increasing the computational burden of the method.

Song et al. (2012a) proposed a novel method for human behavior recognition based on multiview hidden conditional random fields (MV-HCRF) (Song et al., 2012b) and estimated the interaction of the different modalities by using kernel canonical correlation analysis (KCCA) (Hardoon et al., 2004). However, their method cannot deal with data that contain complex backgrounds, and due to the down-sampling of the original data the audio-visual synchronization may be lost. Also, their method used different sets of hidden states for audio and visual information. This property considers that the audio and visual features were *a priori* synchronized, while it increases the complexity of the model. Metallinou et al. (2012) employed several hierarchical classification models from neural networks to HMMs and their combinations to recognize audio-visual emotional levels of valence and arousal rather than emotional labels, such as anger and kindness.

Vrigkas et al. (2014b) employed a fully connected CRF model to identify human behaviors, such as friendly, aggressive, and neutral. To evaluate their method, they introduced a novel behavior dataset, called the *Parliament* dataset, which consists of political speeches in the Greek parliament. Bousmalis et al. (2013b) proposed a method based on hierarchical Dirichlet processes to automatically estimate the optimal number of hidden states in an HCRF model for identifying human behaviors. The proposed

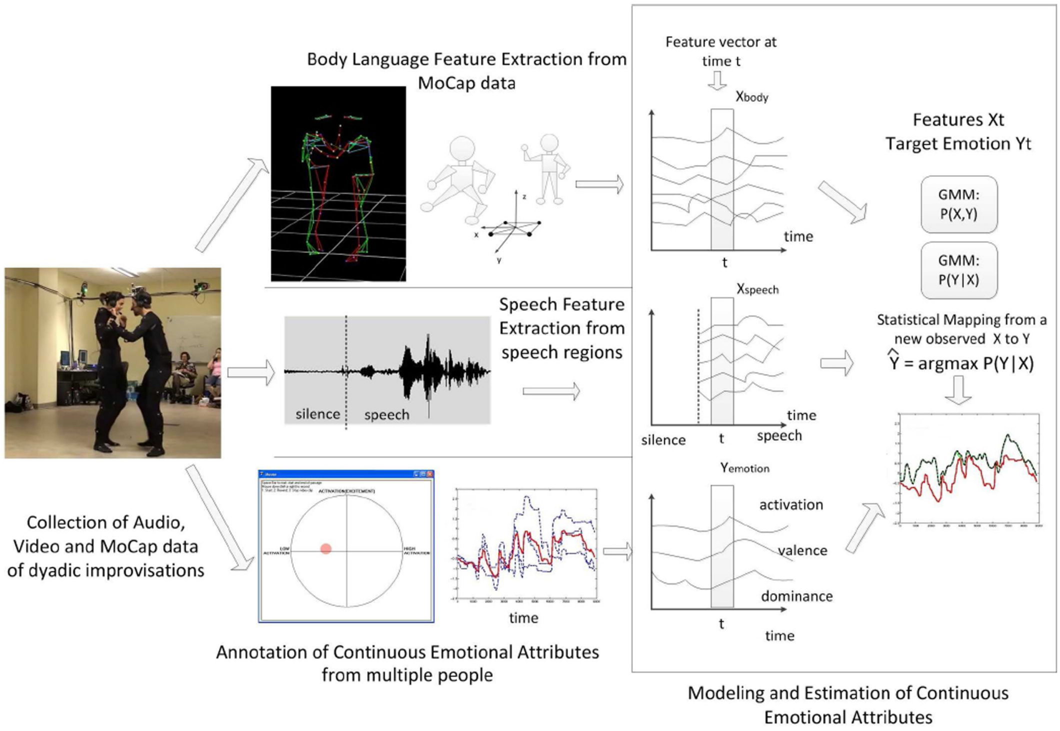
Body Language Feature Extraction from MoCap data
Speech Feature Extraction from speech regions
silence
speech
Collection of Audio, Video and MoCap data of dyadic improvisations
Annotation of Continuous Emotional Attributes from multiple people
time
Feature vector at time t
Xbody
time
t
Xspeech
time
silence
t
speech
Yemotion
activation
valence
dominance
t
time
Features Xt
Target Emotion Yt
GMM: P(X,Y)
GMM: P(Y|X)
Statistical Mapping from a new observed X to Y
$\hat{Y}$ = argmax P(Y|X)
Modeling and Estimation of Continuous Emotional Attributes

FIGURE 10 | Example of interacting persons. Audio-visual features and emotional annotations are fed into a GMM for estimating the emotional curves (Metallinou et al., 2013).

model, also known as infinite hidden conditional random field model (iHCRF), was employed to recognize emotional states, such as pain and agreement, and disagreement from non-verbal multimodal cues.

Baxter et al. (2015) proposed a human classification model that does not learn the temporal structure of human actions but rather decomposes human actions and uses them as features for learning complex human activities. The intuition behind this approach is a psycholinguistics phenomenon, where randomizing letters in the middle of words has almost no effect on understanding the underlying word if and only if the first and the last letters of this word remain unchanged (Rawlinson, 2007). The problem of behavioral mimicry in social interactions was studied by Bilakhia et al. (2013). It can be seen as an interpretation of human speech, facial expressions, gestures, and movements. Metallinou et al. (2013) applied mixture models to capture the mapping between audio and visual cues to understand the emotional states of dyadic interactions.

Selecting the proper features for human behavior recognition has always been a trial-and-error approach for many researchers in this area of study. In general, effective feature extraction is highly application dependent. Several feature descriptors, such as HOG3D (Kläser et al., 2008) and STIP (Laptev, 2005), are not able to sufficiently characterize human behaviors. The combination of visual features with other more informative features, which reflect human emotions and psychology, is necessary for this task. Nonetheless, the description of human activities with high-level contents usually leads to recognition methods with high computational complexity. Another obstacle that researchers must overcome is the lack of adequate benchmark datasets to test and validate the reliability, effectiveness, and efficiency of a human behavior recognition system.

5.3. Methods Based on Social Networking

Social interactions are an important part of daily life. A fundamental component of human behavior is the ability to interact with other people via their actions. Social interaction can be considered as a special type of activity where someone adapts his/her behavior according to the group of people surrounding him/her. Most of the social networking systems that affect people's behavior, such as Facebook, Twitter, and YouTube, measure social interactions and infer how such sites may be involved in issues of identity, privacy, social capital, youth culture, and education. Moreover, the field of psychology has attracted great interest in studying social interactions, as scientists may infer useful information about human behavior. A recent survey on human behavior recognition provides a complete summarization of up-to-date techniques for automatic human behavior analysis for single person, multiperson, and object-person interactions (Candamo et al., 2010).

Fathi et al. (2012) modeled social interactions by estimating the location and orientation of the faces of persons taking part in a social event, computing a line of sight for each face. This information was used to infer the location where an individual may be found. The type of interaction was recognized by assigning social roles to each person. The authors were able to recognize three types of social interactions: dialog, discussion, and monolog. To capture these social interactions, eight subjects wearing head-mounted cameras participated in groups of interacting persons analyzing their activities from the first-person point of view. **Figure 11** shows the resulting social network built from this method. In the sense of first-person scene understanding, Park and Shi (2015) were able to predict joint social interactions by modeling geometric relationships between groups of interacting persons. Although the proposed method could cope with missing information and variations in scene context, scale, and orientation of human poses, it is sensitive to localization of interacting members, which leads to erroneous predictions of the true class.

Human behavior on sport datasets was investigated by Lan et al. (2012a). The authors modeled the behavior of humans in a scene using social roles in conjunction with modeling low-level actions and high-level events. Burgos-Artizzu et al. (2012) discussed the social behavior of mice. Each video sequence was segmented into periods of activities by constructing a temporal context that combines spatiotemporal features. Kong et al. (2014a) proposed a new high-level descriptor called "interactive phrases" to recognize human interactions. This descriptor was a binary motion relationship descriptor for recognizing complex human interactions. Interactive phrases were treated as latent variables, while the recognition was performed using a CRF model.

Cui et al. (2011) recognized abnormal behaviors in human group activities. The authors represented human activities by modeling the relationships between the current behavior of a person and his/her actions. An attribute-based social activity recognition method was introduced by Fu et al. (2014). The authors were interested in classifying social activities of daily life, such as birthdays and weddings. A new social activity dataset has also been proposed. By treating attributes as latent variables,

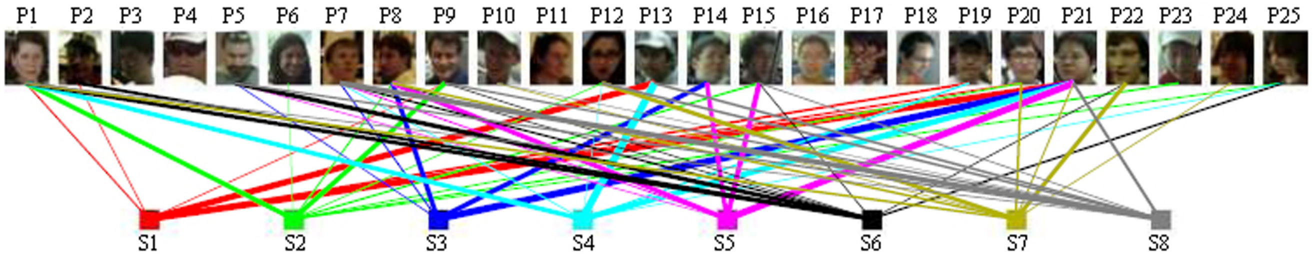

FIGURE 11 | Social network of interacting persons. The connections between the group of persons $P_1 \ldots P_{25}$ and the subjects wearing the cameras $S_1 \ldots S_8$ are weighted based on how often a person's face is captured by a subject's camera (Fathi et al., 2012).

the authors were able to annotate and classify video sequences of social activities. Yan et al. (2014) leveraged the problem of human tracking for modeling the repulsion, attraction, and non-interaction effects in social interactions. The tracking problem was decomposed into smaller tasks by tracking all possible configurations of interactions effects, while the number of trackers was dynamically estimated. Tran et al. (2014b) modeled crowded scenes as a graph of interacting persons. Each node represents one person and each edge on the graph is associated with a weight according to the level of the interaction between the participants. The interacting groups were found by graph clustering, where each maximal clique corresponds to an interacting group.

The work of Lu et al. (2011) focused on automatically tracking and recognizing players' positions (i.e., attacker and defender) in sports' videos. The main problem of this work was the low resolution of the players to be tracked (a player was roughly 15 pixels tall). Lan et al. (2012b) recognized group activities, which were considered as latent variables, encoding the contextual information in a video sequence. Two types of contextual information were explored: group-to-person interactions and person-to-person interactions. To model person-to-person interactions, one approach is to model the associated structure. The second approach is based on spatiotemporal features, which encode the information about an action and the behavior of people in the neighborhood. Finally, the third approach is a combination of the above two.

Much focus has also been given to recognizing human activities from real life videos, such as movies and TV shows, by exploiting scene contexts to localize activities and understand human interactions (Marszałek et al., 2009; Patron-Perez et al., 2012; Bojanowski et al., 2013; Hoai and Zisserman, 2014). The recognition accuracy of such complex videos can also be improved by relating textual descriptions and visual context to a unified framework (Ramanathan et al., 2013). An alternative approach is a system that takes a video clip as its input and generates short textual descriptions, which may correspond to an activity label, which was unseen during training (Guadarrama et al., 2013). However, natural video sequences may contain irrelevant scenes or scenes with multiple actions. As a result, Bandla and Grauman (2013) proposed a method for recognizing human activities from unsegmented videos using a voting-based classification scheme to find the most frequently used action label.

Marín-Jiménez et al. (2014) used a bag of visual-audio words scheme along with late fusion for recognizing human interactions in TV shows. Even though their method performs well in recognizing human interaction, the lack of an intrinsic audio-visual relationship estimation limits the recognition problem. Bousmalis et al. (2011) considered a system based on HCRFs for spontaneous agreement and disagreement recognition using audio and visual features. Although both methods yielded promising results, they did not consider any kind of explicit correlation and/or association between the different modalities. Hoai and Zisserman (2014) proposed a learning based method based on the context and the properties of a scene for detecting upper body positions and understanding the interaction of the participants in TV shows. An audio-visual analysis for recognizing dyadic interactions was presented by Yang et al. (2014). The author combined a GMM with a Fisher kernel to model multimodal dyadic interactions and predict the body language of each subject according to the behavioral state of his/her interlocutor. Escalera et al. (2012) represented the concept of social interactions as an oriented graph using an influence model to identify human interactions. Audio and visual detection and segmentation were performed to extract the exact segments of interest in a video sequence, and then the influence model was employed. Each link measured the influence of a person over another.

Many works on human activity recognition based on deep learning techniques have been proposed in the literature. In fact, deep learning methods have had a large impact on a plethora of research areas including image/video understanding, speech recognition, and biomedical image analysis. Kim et al. (2013) used deep belief networks (DBN) (Hinton et al., 2006) in both supervised and unsupervised manners to learn the most informative audio-visual features and classify human emotions in dyadic interactions. Their system was able to preserve non-linear relationships between multimodal features and showed that unsupervised learning can be used efficiently for feature selection. Shao et al. (2015) mixed appearance and motion features for recognizing group activities in crowded scenes collected from the web. For the combination of the different modalities, the authors applied multitask deep learning. By these means, they were able to capture the intraclass correlations between the learned attributes while they proposed a novel dataset of crowed scene understanding, called WWW crowd dataset.

Deep learning has also been used by Gan et al. (2015) for detecting and recognizing complex events in video sequences. The proposed approach followed a sequential framework. First, saliency maps were used for detecting and localizing events, and then deep learning was applied to the pretrained features for identifying the most important frames that correspond to the underlying event. Although much of the existing work on event understanding relies on video representation, significant work has been done on recognizing complex events from static images. Xiong et al. (2015) utilized CNNs to hierarchically combine information from different visual channels. The new representation of fused features was used to recognize complex social events. To assess their method, the authors introduced a large dataset with >60,000 static images obtained from the web, called web image dataset for event recognition (WIDER).

Karpathy et al. (2014) performed an experimental evaluation of CNNs to classify events from large-scale video datasets, using one million videos with 487 categories (Sports-1M dataset) obtained from YouTube videos. Chen et al. (2013a) exploited different types of features, such as static and motion features, for recognizing unlabeled events from heterogenous web data (e.g., YouTube and Google/Bing image search engines). A separate classifier for each source is learned and a multidomain adaptation approach was followed to infer the labels for each data source. Tang et al. (2013) studied the problem of heterogenous feature combination for recognizing complex events. They considered the problem as two different tasks. At first, they estimated which were the most informative features for recognizing social events, and then combined the different features using an AND/OR graph structure.

Modeling crowded scenes has been a difficult task due to partial occlusions, interacting motion patterns, and sparsely distributed cameras in outdoor environments (Alahi et al., 2014). Most of the existing approaches for modeling group activities and social interactions between different persons usually exploit contextual information from the scenes. However, such information is not sufficient to fully understand the underlying activity as it does not capture the variations in human poses when interacting with other persons. When attempting to recognize social interactions with a fixed number of participants, the problem may become more or less trivial. When the number of interacting people dynamically changes over time, the complexity of the problem increases and becomes more challenging. Moreover, social interactions are usually decomposed into smaller subsets that contain individual person activities or interaction between individuals. The individual motion patterns are analyzed separately and are then combined to estimate the event. A person adapts his/her behavior according to the person with whom s/he interacts. Thus, such an approach is limited by the fact that only specific interaction patterns can be successfully modeled and is sensitive in modeling complex social events.

5.4. Multimodal Feature Fusion

Consider the scenario where several people have a specific activity/behavior and some of them may emit sounds. In the simple case, a human activity recognition system may recognize the underlying activity by taking into account only the visual information. However, the recognition accuracy may be enhanced from audio-visual analysis, as different people may exhibit different activities with similar body movements, but with different sound intensity values. The audio information may help to understand who is the person of interest in a test video sequence and distinguish between different behavioral states.

A great difficulty in multimodal feature analysis is the dimensionality of the data from different modalities. For example, video features are much more complex with higher dimensions than audio, and thus techniques for dimensionality reduction are useful. In the literature, there are two main fusion strategies that can be used to tackle this problem (Atrey et al., 2010; Shivappa et al., 2010).

Early fusion, or fusion at the feature level, combines features of different modalities, usually by reducing the dimensionality in each modality and creating a new feature vector that represents an individual. Canonical correlation analysis (CCA) (Hardoon et al., 2004) was widely studied in the literature as an effective way for fusing data at the feature level (Sun et al., 2005; Wang et al., 2011c; Rudovic et al., 2013). The advantage of early fusion is that it yields good recognition results when the different modalities are highly correlated, since only one learning phase is required. On the other hand, the difficulty of combining the different modalities may lead to the domination of one modality over the others. A novel method for fusing verbal (i.e., textual information) and non-verbal (i.e., visual signals) cues was proposed by Evangelopoulos et al. (2013). Each modality is separately analyzed and saliency scores are used for linear and non-linear fusing schemes.

The second category of methods, which is known as *late fusion* or fusion at the decision level, combines several probabilistic models to learn the parameters of each modality separately. Then all scores are combined together in a supervised framework yielding a final decision score (Westerveld et al., 2003; Jiang et al., 2014). The individual strength of each modality may lead to better recognition results. However, this strategy is time-consuming and requires more complex supervised learning schemes, which may cause a potential loss of inter-modality correlation. A comparison of early versus late fusion methods for video analysis was reported by Snoek et al. (2005).

Recently, a third approach for fusing multimodal data has come to the foreground (Karpathy et al., 2014). This approach, called *slow fusion*, is a combination of the previous approaches and can be seen as a hierarchical fusion technique that slowly fuses data by successively passing information through early and late fusion levels. Although this approach seems to have the advantages of both early and late fusion techniques, it also has a large computational burden due to the different levels of information processing. **Figure 12** illustrates the graphical models of different fusion approaches.

6. DISCUSSION

Human activity understanding has become one of the most active research topics in computer vision. The type and amount of data that each approach uses depends on the ability of the underlying algorithm to deal with heterogeneous and/or large scale data. The development of a fully automated human activity recognition system is a non-trivial task due to cluttered backgrounds, complex camera motion, large intraclass variations, and data acquisition issues. **Tables 2** and **3** provide a comprehensive comparison of unimodal and multimodal methods, respectively, and list the benefits and limitations of each family of methods.

The first step in developing a human activity recognition system is to acquire an adequate human activity database. This database may be used for training and testing purposes. A complete survey, which covers important aspects of human activity recognition datasets, was introduced by Chaquet et al. (2013). An appropriate human activity dataset is required for the development of a human activity recognition system. This dataset should be sufficiently rich in a variety of human actions. Moreover, the creation of such a dataset should correspond to real world scenarios. The quality of the input media that forms the dataset is one of the most important things one should take into account. These

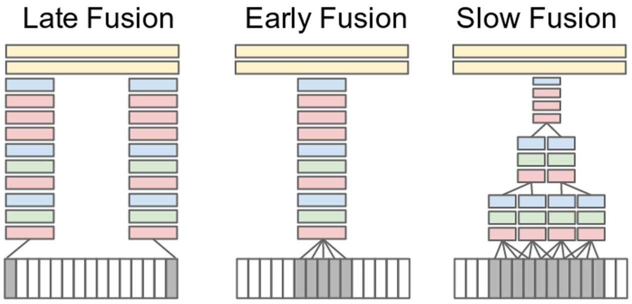

FIGURE 12 | Graphical representation of different fusion approaches (Karpathy et al., 2014).

TABLE 2 | Comparison of unimodal methods.

Type of method	Pros	Cons
Space-time	- Localization of actions - 3D body representation - Good representation of low-level features - Detailed analysis of human movements - Unsupervised learning	- Sensitivity to noise and occlusions - Recognizing complex activities may be tricky - Feature sparsity leads to low repeatability - Gap between low-level features and high-level events - Human body detection is often a prerequisite
Stochastic	- Complex activity recognition - Modeling of human interactions - Recognition from very short clips - Partial occlusion, background clutter, and camera motion handling - High generalization ability - Non-periodic activity recognition	- Learning and inference may be difficult - Learning a large number of parameters - Label bias problem - Prone to overfitting - Approximate solutions - Large number of training data required
Rule-based	- High-level representation of human actions - Sequential activity recognition - Context-free grammar classification - Knowledge transfer between actions - Learning of multiple tasks simultaneously	- Decomposition of complex activities into smaller tasks - Only atomic actions are recognized - Rule/attribute generation is difficult - Problems with long video sequences
Shape-based	- 2D and 3D body representation - Independent modeling of human body parts - Recognition from still images - Upper body action recognition - Existence of low-cost devices for pose estimation	- Large number of degrees of freedom - Skeleton tracking inaccuracies - View-point and self occlusions dependent - Sensitivity to illumination changes and human clothing - Difficulties in mapping image feature space to pose space

TABLE 3 | Comparison of multimodal methods.

Type of method	Pros	Cons
Affective	- Association of human emotions and actions - Better understanding of human activities - Complex activity recognition - Incorporation of well known classification models	- Affective data annotation is difficult - Problems in handling continuous actions - Dimensionality of the different modalities - Gap between low-level features and high-level concepts
Behavioral	- Personalized action recognition - Improve human-computer interaction - Complex activity recognition - Recognizes human interactions - Psychological attributes improve recognition	- Emotional attribute specification is difficult - Mainly frontal view emotion recognition - Complex classification models - Proper feature selection is difficult - Visual feature descriptors cannot capture human emotions - Dimensionality of the different modalities
Social networking	- Recognizes social human interactions - Easy access to data though social platforms - Reliable recognition of human-to-human or human-to-object interactions - Abnormal activity recognition	- Limited by the number of interacting persons - Dimensionality of the different modalities - Decomposition of complex actions into smaller tasks is necessary - Difficulties in crowded scene modeling due to occlusions

input media can be static images or video sequences, colored or gray-scaled. An ideal human activity dataset should address the following issues: (i) the input media should include either still images and/or video sequences, (ii) the amount of data should be sufficient, (iii) input media quality (resolution, grayscale or color), (iv) large number of subjects performing an action, (v) large number of action classes, (vi) changes in illuminations, (vii) large intraclass variations (i.e., variations in subjects' poses), (viii) photo shooting under partial occlusion of human structure, and (ix) complex backgrounds.

Although there exists a plethora of benchmark activity recognition datasets in the literature, we have focused on the most widely used ones with respect to the database size, resolution, and usability. **Table 4** summarizes human activity recognition datasets, categorizing them into seven different categories. All datasets are grouped by their associated category and by chronological order for each group. We also present the number of classes, actors, and video clips along with their frame resolution.

Many of the existing datasets for human activity recognition were recorded in controlled environments, with participant actors performing specific actions. Furthermore, several datasets are not generic, but rather cover a specific set of activities, such as sports and simple actions, which are usually performed by one actor. However, these limitations constitute an unrealistic scenario that does not cover real-world situations and does not address the specifications for an ideal human activity dataset as presented earlier. Nevertheless, several activity recognition datasets that take into account these requirements have been proposed.

Several existing datasets have reached their expected life cycle (i.e., methods on Weizmann and KTH datasets achieved 100% recognition rate). These datasets were captured in controlled environments and the performed actions were obtained from

TABLE 4 | Human activity recognition datasets.

Dataset name and category	Year	# classes	# actors	# videos	Resolution
Single action recognition					
KTH (Schuldt et al., 2004)	2004	6	25	2,391	160 × 120
Weizman (Blank et al., 2005)	2005	10	9	90	180 × 144
UCF Sports (Rodriguez et al., 2008)	2008	9		200	720 × 480
MuHAVi (Singh et al., 2010)	2010	17	14		720 × 576
UCF50 (Reddy and Shah, 2013)	2013	50		6,676	
UCF101 (Soomro et al., 2012)	2012	101		13,320	320 × 240
Movie					
UCF YouTube (Liu et al., 2009)	2009	11		>1.100	720 × 480
Hollywood2 (Marszałek et al., 2009)	2009	12		3,669	
HMDB51 (Kuehne et al., 2011)	2011	51		6,849	320 × 240
TVHI (Patron-Perez et al., 2012)	2012	4	20	300	320 × 240
Surveillance					
PETS 2004 (CAVIAR) (Fisher, 2004)	2004	6		28	384 × 288
PETS 2007 (Fisher, 2007b)	2007	3		7	768 × 576
VIRAT (Oh et al., 2011)	2011	23		17	1,920 × 1,080
Pose					
TUM Kitchen (Tenorth et al., 2009)	2009	10	4	20	324 × 288
Two-person interaction (Yun et al., 2012)	2012	8	7	≈300	640 × 480
MSRC-12 Kinect gesture (Fothergill et al., 2012)	2012	12	30	594	
J-HMDB (Jhuang et al., 2013)	2013	21	1	928	240 × 320
UPCV Action (Theodorakopoulos et al., 2014)	2014	10	20	≈200	
Daily living					
URADL (Messing et al., 2009)	2009	17	5	150	1,280 × 720
ADL (Pirsiavash and Ramanan, 2012)	2012	18	20	≈10 h	1,280 × 960
MPII Cooking (Rohrbach et al., 2012)	2012	65	12	44	1,624 × 1,224
Breakfast (Kuehne et al., 2014)	2014	10	52	≈77 h	320 × 240
Social networking					
CCV (Jiang et al., 2011)	2001	20		9.317	
FPSI (Fathi et al., 2012)	2012	6	8	≈42 h	1,280 × 720
Broadcast field hockey (Lan et al., 2012b)	2012	11		58	
USAA (Fu et al., 2012)	2012	8		≈200	
Sports-1M (Karpathy et al., 2014)	2014	487		1 M	
ActivityNet (Heilbron et al., 2015)	2015	203		27,801	1,280 × 720
WWW Crowd (Shao et al., 2015)	2015	94		10,000	640 × 360
Behavior					
BEHAVE (Fisher, 2007a)	2007	8		321	640 × 480
Canal9 (Vinciarelli et al., 2009)	2009	2	190	≈42 h	720 × 576
USC Creative IT (Metallinou et al., 2010)	2010	50	16	100	
Parliament (Vrigkas et al., 2014b)	2014	3	20	228	320 × 240

a frontal view camera. The non-complex backgrounds and the non-intraclass variations in human movements make these datasets non-applicable for real world applications. However, these datasets still remain popular for human activity classification, as they provide a good evaluation criterion for many new methods. A significant problem in constructing a proper human activity recognition dataset is the annotation of each action, which is generally performed manually by the user, making the task biased.

Understanding human activities is a part of interpersonal relationships. Humans have the ability to understand another human's actions by interpreting stimuli from the surroundings. On the other hand, machines need a learning phase to be able to perform this operation. Thus, some basic questions arise about a human activity classification system:

1. How to determine whether a human activity classification system provides the best performance?
2. In which cases is the system prone to errors when classifying a human activity?
3. In what level can the system reach the human ability of recognizing a human activity?
4. Are the success rates of the system adequate for inferring safe conclusions?

It is necessary for the system to be fully automated. To achieve this, all stages of human activity modeling and analysis are to be performed automatically, namely: (i) human activity detection and localization, where the challenge is to detect and localize a human activity in the scene. Background subtraction (Elgammal et al., 2002) and human tracking (Liu et al., 2010) are usually used as a part of this process; (ii) human activity modeling (e.g., feature extraction; Laptev, 2005) is the step of extracting the necessary information that will help in the recognition step; and (iii) human activity classification is the step where a probe video sequence is classified in one of the classes of the activities that have been defined before building the system.

In addition, the system should work regardless of any external factors. This means that the system should perform robustly despite changes in lighting, pose variations or partially occluded human bodies, and background clutter. Also, the number as well as the type of human activity classes to be recognized is an important factor that plays a crucial role in the robustness of the system. The requirements of an ideal human activity classification system should cover several topics, including automatic human activity classification and localization, lighting and pose variations (e.g., multiview recognition), partially occluded human bodies, and background clutter. Also, all possible activities should be detected during the recognition process, the recognition accuracy should be independent from the number of activity classes, and the activity identification process should be performed in real time and provide a high success rate and low false positive rate.

Besides the vast amount of research in this field, a generalization of the learning framework is crucial toward modeling and understanding real world human activities. Several challenges that correspond to the ability of a classification system to generalize under external factors, such as variations in human poses and different data acquisition, are still open issues. The ability of a human activity classification system to imitate humans' skill in recognizing human actions in real time is a future challenge to be tackled. Machine-learning techniques that incorporate knowledge-driven approaches may be vital for human activity modeling and recognition in unconstrained environments, where data may not be adequate or may suffer from occlusions and changes in illuminations and view point.

Training and validation methods still suffer from limitations, such as slow learning rate, which gets even worse for large scale training data, and low recognition rate. Although much research focuses on leveraging human activity recognition from big data, this problem is still in its infancy. The exact opposite problem (i.e., learning human activities from very little training data or missing data) is also very challenging. Several issues concerning the minimum number of learning examples for modeling the dynamics of each class or safely inferring the performed activity label are still open and need further investigation. More attention should also be put in developing robust methods under the uncertainty of missing data either on training steps or testing steps.

The role of appropriate feature extraction for human activity recognition is a problem that needs to be tackled in future research. The extraction of low-level features that are focused on representing human motion is a very challenging task. To this end, a fundamental question arises are there features that are invariant to scale and viewpoint changes, which can model human motion in a unique manner, for all possible configurations of human pose?

Furthermore, it is evident that there exists a great need for efficiently manipulating training data that may come from heterogeneous sources. The number and type of different modalities that can be used for analyzing human activities is an important question. The combination of multimodal features, such as body motion features, facial expressions, and the intensity level of voice, may produce superior results, when compared to unimodal approaches, On the other hand, such a combination may constitute over-complete examples that can be confusing and misleading. The proposed multimodal feature fusion techniques do not incorporate the special characteristics of each modality and the level of abstraction for fusing. Therefore, a comprehensive evaluation of feature fusion methods that retain the feature coupling is an issue that needs to be assessed.

It is evident that the lack of large and realistic human activity recognition datasets is a significant challenge that needs to be addressed. An ideal action dataset should cover several topics, including diversity in human poses for the same action, a wide range of ground truth labels, and variations in image capturing and quality. Although a list of action datasets that correspond to most of these specifications has been introduced in the literature, the question of how many actions we can actually learn is a task for further exploration. Most of the existing datasets contain very few classes (15 on average). However, there exist datasets with more activities that reach 203 or 487 classes. In such large datasets, the ability to distinguish between easy and difficult examples for representing the different classes and recognizing the underlying activity is difficult. This fact opens a promising research area that should be further studied.

Another challenge worthy of further exploration is the exploitation of unsegmented sequences, where one activity may succeed another. Frequent changes in human motion and actions performed by groups of interacting persons make the problem amply challenging. More sophisticated high-level activity recognition methods need to be developed, which should be able to localize and recognize simultaneously occurring actions by different persons.

7. CONCLUSION

In this survey, we carried out a comprehensive study of state-of-the-art methods of human activity recognition and proposed a hierarchical taxonomy for classifying these methods. We surveyed different approaches, which were classified into two broad categories (unimodal and multimodal) according to the source channel each of these approaches employ to recognize human activities. We discussed unimodal approaches and provided an internal categorization of these methods, which were developed for analyzing gesture, atomic actions, and more complex activities, either directly or employing activity decomposition into simpler actions. We also presented multimodal approaches for the analysis of human social behaviors and interactions. We discussed the different levels of representation of feature modalities and reported the limitations and advantages for each representation. A comprehensive review of existing human activity classification benchmarks was also presented and we examined the challenges of data acquisition to the problem of understanding human activity. Finally, we provided the characteristics of building an ideal human activity recognition system.

Most of the existing studies in this field failed to efficiently describe human activities in a concise and informative way as they introduce limitations concerning computational issues. The gap of a complete representation of human activities and the corresponding data collection and annotation is still a challenging and unbridged problem. In particular, we may conclude that despite the tremendous increase of human understanding methods, many problems still remain open, including modeling of human poses, handling occlusions, and annotating data.

REFERENCES

Aggarwal, J. K., and Cai, Q. (1999). Human motion analysis: a review. *Comput. Vis. Image Understand.* 73, 428–440. doi:10.1006/cviu.1998.0744

Aggarwal, J. K., and Ryoo, M. S. (2011). Human activity analysis: a review. *ACM Comput. Surv.* 43, 1–43. doi:10.1145/1922649.1922653

Aggarwal, J. K., and Xia, L. (2014). Human activity recognition from 3D data: a review. *Pattern Recognit. Lett.* 48, 70–80. doi:10.1016/j.patrec.2014.04.011

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2013). "Label-embedding for attribute-based classification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 819–826.

Alahi, A., Ramanathan, V., and Fei-Fei, L. (2014). "Socially-aware large-scale crowd forecasting," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 2211–2218.

AlZoubi, O., Fossati, D., D'Mello, S. K., and Calvo, R. A. (2013). "Affect detection and classification from the non-stationary physiological data," in *Proc. International Conference on Machine Learning and Applications* (Portland, OR), 240–245.

Amer, M. R., and Todorovic, S. (2012). "Sum-product networks for modeling activities with stochastic structure," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1314–1321.

Amin, S., Andriluka, M., Rohrbach, M., and Schiele, B. (2013). "Multi-view pictorial structures for 3D human pose estimation," in *Proc. British Machine Vision Conference* (Bristol), 1–12.

Andriluka, M., Pishchulin, L., Gehler, P. V., and Schiele, B. (2014). "2D human pose estimation: new benchmark and state of the art analysis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 3686–3693.

Andriluka, M., and Sigal, L. (2012). "Human context: modeling human-human interactions for monocular 3D pose estimation," in *Proc. International Conference on Articulated Motion and Deformable Objects* (Mallorca: Springer-Verlag), 260–272.

Anirudh, R., Turaga, P., Su, J., and Srivastava, A. (2015). "Elastic functional coding of human actions: from vector-fields to latent variables," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3147–3155.

Atrey, P. K., Hossain, M. A., El-Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimed. Syst.* 16, 345–379. doi:10.1007/s00530-010-0182-0

Bandla, S., and Grauman, K. (2013). "Active learning of an action detector from untrimmed videos," in *Proc. IEEE International Conference on Computer Vision* (Sydney, NSW), 1833–1840.

Baxter, R. H., Robertson, N. M., and Lane, D. M. (2015). Human behaviour recognition in data-scarce domains. *Pattern Recognit.* 48, 2377–2393. doi:10.1016/j.patcog.2015.02.019

Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., and Ilic, S. (2014). "3D pictorial structures for multiple human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 1669–1676.

Bilakhia, S., Petridis, S., and Pantic, M. (2013). "Audiovisual detection of behavioural mimicry," in *Proc. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (Geneva), 123–128.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Secaucus, NJ: Springer.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). "Actions as space-time shapes," in *Proc. IEEE International Conference on Computer Vision* (Beijing), 1395–1402.

ACKNOWLEDGMENTS

This research was funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. (2013). "Finding actors and actions in movies," in *Proc. IEEE International Conference on Computer Vision* (Sydney), 2280–2287.

Bousmalis, K., Mehu, M., and Pantic, M. (2013a). Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: a survey of related cues, databases, and tools. *Image Vis. Comput.* 31, 203–221. doi:10.1016/j.imavis.2012.07.003

Bousmalis, K., Zafeiriou, S., Morency, L. P., and Pantic, M. (2013b). Infinite hidden conditional random fields for human behavior analysis. *IEEE Trans. Neural Networks Learn. Syst.* 24, 170–177. doi:10.1109/TNNLS.2012.2224882

Bousmalis, K., Morency, L., and Pantic, M. (2011). "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition* (Santa Barbara, CA), 746–752.

Burenius, M., Sullivan, J., and Carlsson, S. (2013). "3D pictorial structures for multiple view articulated pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 3618–3625.

Burgos-Artizzu, X. P., Dollár, P., Lin, D., Anderson, D. J., and Perona, P. (2012). "Social behavior recognition in continuous video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1322–1329.

Candamo, J., Shreve, M., Goldgof, D. B., Sapper, D. B., and Kasturi, R. (2010). Understanding transit scenes: a survey on human behavior-recognition algorithms. *IEEE Trans. Intell. Transp. Syst.* 11, 206–224. doi:10.1109/TITS.2009.2030963

Castellano, G., Villalba, S. D., and Camurri, A. (2007). "Recognising human emotions from body movement and gesture dynamics," in *Proc. Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, Vol. 4738 (Lisbon), 71–82.

Chakraborty, B., Holte, M. B., Moeslund, T. B., and Gonzàlez, J. (2012). Selective spatio-temporal interest points. *Comput. Vis. Image Understand.* 116, 396–410. doi:10.1016/j.cviu.2011.09.010

Chaquet, J. M., Carmona, E. J., and Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Comput. Vis. Image Understand.* 117, 633–659. doi:10.1016/j.cviu.2013.01.013

Chaudhry, R., Ravichandran, A., Hager, G. D., and Vidal, R. (2009). "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 1932–1939.

Chen, C. Y., and Grauman, K. (2012). "Efficient activity detection with max-subgraph search," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1274–1281.

Chen, H., Li, J., Zhang, F., Li, Y., and Wang, H. (2015). "3D model-based continuous emotion recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1836–1845.

Chen, L., Duan, L., and Xu, D. (2013a). "Event recognition in videos by learning from heterogeneous web sources," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 2666–2673.

Chen, L., Wei, H., and Ferryman, J. (2013b). A survey of human motion analysis using depth imagery. *Pattern Recognit. Lett.* 34, 1995–2006. doi:10.1016/j.patrec.2013.02.006

Chen, W., Xiong, C., Xu, R., and Corso, J. J. (2014). "Actionness ranking with lattice conditional ordinal random fields," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 748–755.

Cherian, A., Mairal, J., Alahari, K., and Schmid, C. (2014). "Mixing body-part sequences for human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 2361–2368.

Choi, W., Shahid, K., and Savarese, S. (2011). "Learning context for collective activity recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3273–3280.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). "Flexible, high performance convolutional neural networks for image classification," in *Proc. International Joint Conference on Artificial Intelligence* (Barcelona), 1237–1242.

Ciresan, D. C., Meier, U., and Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 3642–3649.

Cui, X., Liu, Q., Gao, M., and Metaxas, D. N. (2011). "Abnormal detection using interaction energy potentials," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3161–3167.

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 886–893.

Dalal, N., Triggs, B., and Schmid, C. (2006). "Human detection using oriented histograms of flow and appearance," in *Proc. European Conference on Computer Vision* (Graz), 428–441.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). "Behavior recognition via sparse spatio-temporal features," in *Proc. International Conference on Computer Communications and Networks* (Beijing), 65–72.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., et al. (2015). "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 2625–2634.

Du, Y., Wang, W., and Wang, L. (2015). "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1110–1118.

Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). "Recognizing action at a distance," in *Proc. IEEE International Conference on Computer Vision*, Vol. 2 (Nice), 726–733.

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System (FACS): Manual*. Salt Lake City: A Human Face.

Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. (2002). Background and foreground modeling using nonparametric kernel density for visual surveillance. *Proc. IEEE* 90, 1151–1163. doi:10.1109/JPROC.2002.801448

Escalera, S., Baró, X., Vitrià, J., Radeva, P., and Raducanu, B. (2012). Social network extraction and analysis based on multimodal dyadic interaction. *Sensors* 12, 1702–1719. doi:10.3390/s120201702

Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., et al. (2013). Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimedia* 15, 1553–1568. doi:10.1109/TMM.2013.2267205

Evgeniou, T., and Pontil, M. (2004). "Regularized multi-task learning," in *Proc. ACM International Conference on Knowledge Discovery and Data Mining* (Seattle, WA), 109–117.

Eweiwi, A., Cheema, M. S., Bauckhage, C., and Gall, J. (2014). "Efficient pose-based action recognition," in *Proc. Asian Conference on Computer Vision* (Singapore), 428–443.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. A. (2009). "Describing objects by their attributes," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 1778–1785.

Fathi, A., Hodgins, J. K., and Rehg, J. M. (2012). "Social interactions: a first-person perspective," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1226–1233.

Fathi, A., and Mori, G. (2008). "Action recognition by learning mid-level motion features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Fergie, M., and Galata, A. (2013). Mixtures of Gaussian process models for human pose estimation. *Image Vis. Comput.* 31, 949–957. doi:10.1016/j.imavis.2013.09.007

Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., and Tuytelaars, T. (2015). "Modeling video evolution for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 5378–5387.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2009). "Pose search: retrieving people using their pose," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 1–8.

Fisher, R. B. (2004). *PETS04 Surveillance Ground Truth Dataset*. Available at: http://www-prima.inrialpes.fr/PETS04/

Fisher, R. B. (2007a). *Behave: Computer-Assisted Prescreening of Video Streams for Unusual Activities*. Available at: http://homepages.inf.ed.ac.uk/rbf/BEHAVE/

Fisher, R. B. (2007b). *PETS07 Benchmark Dataset*. Available at: http://www.cvg.reading.ac.uk/PETS2007/data.html

Fogel, I., and Sagi, D. (1989). Gabor filters as texture discriminator. *Biol. Cybern.* 61, 103–113. doi:10.1007/BF00204594

Fothergill, S., Mentis, H. M., Kohli, P., and Nowozin, S. (2012). "Instructing people for training gestural interactive systems," in *Proc. Conference on Human Factors in Computing Systems* (Austin, TX), 1737–1746.

Fouhey, D. F., Delaitre, V., Gupta, A., Efros, A. A., Laptev, I., and Sivic, J. (2014). People watching: human actions as a cue for single view geometry. *Int. J. Comput. Vis.* 110, 259–274. doi:10.1007/s11263-014-0710-z

Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. (2012). "Attribute learning for understanding unstructured social activity," in *Proc. European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 7575 (Florence), 530–543.

Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. (2014). Learning multimodal latent attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 303–316. doi:10.1109/TPAMI.2013.128

Gaidon, A., Harchaoui, Z., and Schmid, C. (2014). Activity representation with motion hierarchies. *Int. J. Comput. Vis.* 107, 219–238. doi:10.1007/s11263-013-0677-1

Gan, C., Wang, N., Yang, Y., Yeung, D. Y., and Hauptmann, A. G. (2015). "DevNet: a deep event network for multimedia event detection and evidence recounting," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 2568–2577.

Gao, Z., Zhang, H., Xu, G. P., and Xue, Y. B. (2015). Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition. *Neurocomputing* 151, 554–564. doi:10.1016/j.neucom.2014.06.085

Gavrila, D. M. (1999). The visual analysis of human movement: a survey. *Comput. Vis. Image Understand.* 73, 82–98. doi:10.1006/cviu.1998.0716

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 2247–2253. doi:10.1109/TPAMI.2007.70711

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R. J., Darrell, T., et al. (2013). "Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE International Conference on Computer Vision* (Sydney, NSW), 2712–2719.

Guha, T., and Ward, R. K. (2012). Learning sparse representations for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1576–1588. doi:10.1109/TPAMI.2011.253

Guo, G., and Lai, A. (2014). A survey on still image based human action recognition. *Pattern Recognit.* 47, 3343–3361. doi:10.1016/j.patcog.2014.04.018

Gupta, A., and Davis, L. S. (2007). "Objects in action: an approach for combining action understanding and object perception," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN), 1–8.

Gupta, A., Kembhavi, A., and Davis, L. S. (2009). Observing human-object interactions: using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 1775–1789. doi:10.1109/TPAMI.2009.83

Haralick, R. M., and Watson, L. (1981). A facet model for image data. *Comput. Graph. Image Process.* 15, 113–129. doi:10.1016/0146-664X(81)90073-3

Hardoon, D. R., Szedmak, S. R., and Shawe-Taylor, J. R. (2004). Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* 16, 2639–2664. doi:10.1162/0899766042321814

Healey, J. (2011). "Recording affect in the field: towards methods and metrics for improving ground truth labels," in *Proc. International Conference on Affective Computing and Intelligent Interaction* (Memphis, TN), 107–116.

Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. (2015). "ActivityNet: a large-scale video benchmark for human activity understanding," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 961–970.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi:10.1162/neco.2006.18.7.1527

Ho, T. K. (1995). "Random decision forests," in *Proc. International Conference on Document Analysis and Recognition*, Vol. 1 (Washington, DC: IEEE Computer Society), 278–282.

Hoai, M., Lan, Z. Z., and Torre, F. (2011). "Joint segmentation and classification of human actions in video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3265–3272.

Hoai, M., and Zisserman, A. (2014). "Talking heads: detecting humans and recognizing their interactions," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 875–882.

Holte, M. B., Chakraborty, B., Gonzàlez, J., and Moeslund, T. B. (2012a). A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. *IEEE J. Sel. Top. Signal Process.* 6, 553–565. doi:10.1109/JSTSP.2012.2193556

Holte, M. B., Tran, C., Trivedi, M. M., and Moeslund, T. B. (2012b). Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. *IEEE J. Sel. Top. Signal Process.* 6, 538–552. doi:10.1109/JSTSP.2012.2196975

Huang, Z. F., Yang, W., Wang, Y., and Mori, G. (2011). "Latent boosting for action recognition," in *Proc. British Machine Vision Conference* (Dundee), 1–11.

Hussain, M. S., Calvo, R. A., and Pour, P. A. (2011). "Hybrid fusion approach for detecting affects from multichannel physiology," in *Proc. International Conference on Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, Vol. 6974 (Memphis, TN), 568–577.

Ikizler, N., and Duygulu, P. (2007). "Human action recognition using distribution of oriented rectangular patches," in *Proc. Conference on Human Motion: Understanding, Modeling, Capture and Animation* (Rio de Janeiro), 271–284.

Ikizler-Cinbis, N., and Sclaroff, S. (2010). "Object, scene and actions: combining multiple features for human action recognition," in *Proc. European Conference on Computer Vision, Lecture Notes in Computer Science*, Vol. 6311 (Hersonissos, Heraclion, Crete, greece: Springer), 494–507.

Iosifidis, A., Tefas, A., and Pitas, I. (2012a). Activity-based person identification using fuzzy representation and discriminant learning. *IEEE Trans. Inform. Forensics Secur.* 7, 530–542. doi:10.1109/TIFS.2011.2175921

Iosifidis, A., Tefas, A., and Pitas, I. (2012b). View-invariant action recognition based on artificial neural networks. *IEEE Trans. Neural Networks Learn. Syst.* 23, 412–424. doi:10.1109/TNNLS.2011.2181865

Jaimes, A., and Sebe, N. (2007). "Multimodal human-computer interaction: a survey," in *Computer Vision and Image Understanding*, Vol. 108 (Special Issue on Vision for Human-Computer Interaction), 116–134.

Jain, M., Gemert, J., Jégou, H., Bouthemy, P., and Snoek, C. G. M. (2014). "Action localization with tubelets from motion," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 740–747.

Jain, M., Jegou, H., and Bouthemy, P. (2013). "Better exploiting motion for better action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 2555–2562.

Jainy, M., Gemerty, J. C., and Snoek, C. G. M. (2015). "What do 15,000 object categories tell us about classifying and localizing actions?," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 46–55.

Jayaraman, D., and Grauman, K. (2014). "Zero-shot recognition with unreliable attributes," in *Proc. Annual Conference on Neural Information Processing Systems* (Montreal, QC), 3464–3472.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. J. (2013). "Towards understanding action recognition," in *Proc. IEEE International Conference on Computer Vision* (Sydney, NSW), 3192–3199.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). "A biologically inspired system for action recognition," in *Proc. IEEE International Conference on Computer Vision* (Rio de Janeiro), 1–8.

Jiang, B., Martínez, B., Valstar, M. F., and Pantic, M. (2014). "Decision level fusion of domain specific regions for facial action recognition," in *Proc. International Conference on Pattern Recognition* (Stockholm), 1776–1781.

Jiang, Y. G., Ye, G., Chang, S. F., Ellis, D. P. W., and Loui, A. C. (2011). "Consumer video understanding: a benchmark database and an evaluation of human and machine performance," in *Proc. International Conference on Multimedia Retrieval* (Trento), 29–36.

Jiang, Z., Lin, Z., and Davis, L. S. (2013). A unified tree-based framework for joint action localization, recognition and segmentation. *Comput. Vis. Image Understand.* 117, 1345–1355. doi:10.1016/j.cviu.2012.09.008

Jung, H. Y., Lee, S., Heo, Y. S., and Yun, I. D. (2015). "Random treewalk toward instantaneous 3D human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 2467–2474.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 1725–1732.

Khamis, S., Morariu, V. I., and Davis, L. S. (2012). "A flow model for joint action recognition and identity maintenance," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1218–1225.

Kim, Y., Lee, H., and Provost, E. M. (2013). "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC), 3687–3691.

Klami, A., and Kaski, S. (2008). Probabilistic approach to detecting dependencies between data sets. *Neurocomputing* 72, 39–46. doi:10.1016/j.neucom.2007.12.044

Kläser, A., Marszałek, M., and Schmid, C. (2008). "A spatio-temporal descriptor based on 3D-gradients," in *Proc. British Machine Vision Conference* (Leeds: University of Leeds), 995–1004.

Kohonen, T., Schroeder, M. R., and Huang, T. S. (eds) (2001). *Self-Organizing Maps*, Third Edn. New York, NY.: Springer-Verlag Inc.

Kong, Y., and Fu, Y. (2014). "Modeling supporting regions for close human interaction recognition," in *Proc. European Conference on Computer Vision* (Zurich), 29–44.

Kong, Y., Jia, Y., and Fu, Y. (2014a). Interactive phrases: semantic descriptions for human interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1775–1788. doi:10.1109/TPAMI.2014.2303090

Kong, Y., Kit, D., and Fu, Y. (2014b). "A discriminative model with multiple temporal scales for action prediction," in *Proc. European Conference on Computer Vision* (Zurich), 596–611.

Kovashka, A., and Grauman, K. (2010). "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, CA), 2046–2053.

Kuehne, H., Arslan, A., and Serre, T. (2014). "The language of actions: recovering the syntax and semantics of goal-directed human activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 780–787.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). "HMDB: a large video database for human motion recognition," in *Proc. IEEE International Conference on Computer Vision* (Barcelona), 2556–2563.

Kulkarni, K., Evangelidis, G., Cech, J., and Horaud, R. (2015). Continuous action recognition based on sequence alignment. *Int. J. Comput. Vis.* 112, 90–114. doi:10.1007/s11263-014-0758-9

Kulkarni, P., Sharma, G., Zepeda, J., and Chevallier, L. (2014). "Transfer learning via attributes for improved on-the-fly classification," in *Proc. IEEE Winter Conference on Applications of Computer Vision* (Steamboat Springs, CO), 220–226.

Kviatkovsky, I., Rivlin, E., and Shimshoni, I. (2014). Online action recognition using covariance of shape and motion. *Comput. Vis. Image Understand.* 129, 15–26. doi:10.1016/j.cviu.2014.08.001

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning* (Williamstown, MA: Williams College), 282–289.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 951–958.

Lan, T., Chen, T. C., and Savarese, S. (2014). "A hierarchical representation for future action prediction," in *Proc. European Conference on Computer Vision* (Zurich), 689–704.

Lan, T., Sigal, L., and Mori, G. (2012a). "Social roles in hierarchical models for human activity recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1354–1361.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., and Mori, G. (2012b). Discriminative latent models for recognizing contextual group activities. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1549–1562. doi:10.1109/TPAMI.2011.228

Lan, T., Wang, Y., and Mori, G. (2011). "Discriminative figure-centric models for joint action localization and recognition," in *Proc. IEEE International Conference on Computer Vision* (Barcelona), 2003–2010.

Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vis.* 64, 107–123. doi:10.1007/s11263-005-1838-7

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). "Learning realistic human actions from movies," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. (2011). "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3361–3368.

Li, B., Ayazoglu, M., Mao, T., Camps, O. I., and Sznaier, M. (2011). "Activity recognition using dynamic subspace angles," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3193–3200.

Li, B., Camps, O. I., and Sznaier, M. (2012). "Cross-view activity recognition using hankelets," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1362–1369.

Li, R., and Zickler, T. (2012). "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 2855–2862.

Lichtenauer, J., Valstar, J. S. M., and Pantic, M. (2011). Cost-effective solution to synchronised audio-visual data capture using multiple sensors. *Image Vis. Comput.* 29, 666–680. doi:10.1016/j.imavis.2011.07.004

Lillo, I., Soto, A., and Niebles, J. C. (2014). "Discriminative hierarchical modeling of spatio-temporally composable human activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 812–819.

Lin, Z., Jiang, Z., and Davis, L. S. (2009). "Recognizing actions by shape-motion prototype trees," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 444–451.

Liu, J., Kuipers, B., and Savarese, S. (2011a). "Recognizing human actions by attributes," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3337–3344.

Liu, N., Dellandréa, E., Tellez, B., and Chen, L. (2011b). "Associating textual features with visual ones to improve affective image classification," in *Proc. International Conference on Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, Vol. 6974 (Memphis, TN), 195–204.

Liu, J., Luo, J., and Shah, M. (2009). "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 1–8.

Liu, J., Yan, J., Tong, M., and Liu, Y. (2010). "A Bayesian framework for 3D human motion tracking from monocular image," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Dallas, TX: IEEE), 1398–1401.

Livne, M., Sigal, L., Troje, N. F., and Fleet, D. J. (2012). Human attributes from 3D pose tracking. *Comput. Vis. Image Understanding* 116, 648–660. doi:10.1016/j.cviu.2012.01.003

Lu, J., Xu, R., and Corso, J. J. (2015). "Human action segmentation with hierarchical supervoxel consistency," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3762–3771.

Lu, W. L., Ting, J. A., Murphy, K. P., and Little, J. J. (2011). "Identifying players in broadcast sports videos using conditional random fields," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3249–3256.

Ma, S., Sigal, L., and Sclaroff, S. (2015). "Space-time tree ensemble for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 5024–5032.

Maji, S., Bourdev, L. D., and Malik, J. (2011). "Action recognition from a distributed representation of pose and appearance," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3177–3184.

Marín-Jiménez, M. J., Noz Salinas, R. M., Yeguas-Bolivar, E., and de la Blanca, N. P. (2014). Human interaction categorization by using audio-visual cues. *Mach. Vis. Appl.* 25, 71–84. doi:10.1007/s00138-013-0521-1

Marszałek, M., Laptev, I., and Schmid, C. (2009). "Actions in context," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 2929–2936.

Martinez, H. P., Bengio, Y., and Yannakakis, G. N. (2013). Learning deep physiological models of affect. *IEEE Comput. Intell. Mag.* 8, 20–33. doi:10.1109/MCI.2013.2247823

Martinez, H. P., Yannakakis, G. N., and Hallam, J. (2014). Don't classify ratings of affect; rank them! *IEEE Trans. Affective Comput.* 5, 314–326. doi:10.1109/TAFFC.2014.2352268

Matikainen, P., Hebert, M., and Sukthankar, R. (2009). "Trajectons: action recognition through the motion analysis of tracked features," in *Workshop on Video-Oriented Object and Event Classification, in Conjunction with ICCV* (Kyoto: IEEE), 514–521.

Messing, R., Pal, C. J., and Kautz, H. A. (2009). "Activity recognition using the velocity histories of tracked keypoints," in *Proc. IEEE International Conference on Computer Vision* (Kyoto), 104–111.

Metallinou, A., Katsamanis, A., and Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image Vis. Comput.* 31, 137–152. doi:10.1016/j.imavis.2012.08.018

Metallinou, A., Lee, C. C., Busso, C., Carnicke, S. M., and Narayanan, S. (2010). "The USC creative IT database: a multimodal database of theatrical improvisation," in *Proc. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality* (Malta: Springer), 1–4.

Metallinou, A., Lee, S., and Narayanan, S. (2008). "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in *Proc. IEEE International Symposium on Multimedia* (Berkeley, CA), 250–257.

Metallinou, A., and Narayanan, S. (2013). "Annotation and processing of continuous emotional attributes: challenges and opportunities," in *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (Shanghai), 1–8.

Metallinou, A., Wollmer, M., Katsamani, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affective Comput.* 3, 184–198. doi:10.1109/T-AFFC.2011.40

Mikolajczyk, K., and Uemura, H. (2008). "Action recognition with motion-appearance vocabulary forest," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Understand.* 104, 90–126. doi:10.1016/j.cviu.2006.08.002

Morariu, V. I., and Davis, L. S. (2011). "Multi-agent event recognition in structured scenarios," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3289–3296.

Morris, B. T., and Trivedi, M. M. (2011). Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 2287–2301. doi:10.1109/TPAMI.2011.64

Moutzouris, A., del Rincon, J. M., Nebel, J. C., and Makris, D. (2015). Efficient tracking of human poses using a manifold hierarchy. *Comput. Vis. Image Understand.* 132, 75–86. doi:10.1016/j.cviu.2014.10.005

Mumtaz, A., Zhang, W., and Chan, A. B. (2014). "Joint motion segmentation and background estimation in dynamic scenes," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 368–375.

Murray, R. M., Sastry, S. S., and Zexiang, L. (1994). *A Mathematical Introduction to Robotic Manipulation*, first Edn. Boca Raton, FL: CRC Press, Inc.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). "Multimodal deep learning," in *Proc. International Conference on Machine Learning* (Bellevue, WA), 689–696.

Ni, B., Moulin, P., Yang, X., and Yan, S. (2015). "Motion part regularization: improving action recognition via trajectory group selection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3698–3706.

Ni, B., Paramathayalan, V. R., and Moulin, P. (2014). "Multiple granularity analysis for fine-grained action detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 756–763.

Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affective Comput.* 2, 92–105. doi:10.1109/T-AFFC.2011.9

Nicolaou, M. A., Pavlovic, V., and Pantic, M. (2014). Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1299–1311. doi:10.1109/TPAMI.2014.16

Nie, B. X., Xiong, C., and Zhu, S. C. (2015). "Joint action recognition and pose estimation from video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1293–1301.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.* 79, 299–318. doi:10.1007/s11263-007-0122-4

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C., Lee, J. T., et al. (2011). "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3153–3160.

Oikonomopoulos, A., Pantic, M., and Patras, I. (2009). Sparse B-spline polynomial descriptors for human activity recognition. *Image Vis. Comput.* 27, 1814–1825. doi:10.1016/j.imavis.2009.05.010

Oliver, N. M., Rosario, B., and Pentland, A. P. (2000). A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 831–843. doi:10.1109/34.868684

Ouyang, W., Chu, X., and Wang, X. (2014). "Multi-source deep learning for human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 2337–2344.

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). "Zero-shot learning with semantic output codes," in *Proc. Annual Conference on Neural Information Processing Systems* (Vancouver, BC), 1410–1418.

Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2006). "Human computing and machine understanding of human behavior: a survey," in *Proc. International Conference on Multimodal Interfaces* (New York, NY), 239–248.

Pantic, M., and Rothkrantz, L. (2003). "Towards an affect-sensitive multimodal human-computer interaction," in *Proc. IEEE, Special Issue on Multimodal Human-Computer Interaction, Invited Paper*, Vol. 91 (IEEE), 1370–1390.

Park, H. S., and Shi, J. (2015). "Social saliency prediction," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4777–4785.

Patron-Perez, A., Marszalek, M., Reid, I., and Zisserman, A. (2012). Structured learning of human interactions in TV shows. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 2441–2453. doi:10.1109/TPAMI.2012.24

Perez, P., Vermaak, J., and Blake, A. (2004). Data fusion for visual tracking with particles. *Proc. IEEE* 92, 495–513. doi:10.1109/JPROC.2003.823147

Perronnin, F., and Dance, C. R. (2007). "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN), 1–8.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Pirsiavash, H., and Ramanan, D. (2012). "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 2847–2854.

Pirsiavash, H., and Ramanan, D. (2014). "Parsing videos of actions with segmental grammars," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 612–619.

Pishchulin, L., Andriluka, M., Gehler, P. V., and Schiele, B. (2013). "Strong appearance and expressive spatial models for human pose estimation," in *Proc. IEEE International Conference on Computer Vision* (Sydney, NSW), 3487–3494.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vis. Comput.* 28, 976–990. doi:10.1016/j.imavis.2009.11.014

Prince, S. J. D. (2012). *Computer Vision: Models Learning and Inference*. New York, NY: Cambridge University Press.

Quattoni, A., Wang, S., Morency, L. P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 1848–1852. doi:10.1109/TPAMI.2007.1124

Rahmani, H., Mahmood, A., Huynh, D. Q., and Mian, A. S. (2014). "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. IEEE Winter Conference on Applications of Computer Vision* (Steamboat Springs, CO), 626–633.

Rahmani, H., and Mian, A. (2015). "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 2458–2466.

Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., et al. (2015). "Learning semantic relationships for better action retrieval in images," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1100–1109.

Ramanathan, V., Liang, P., and Fei-Fei, L. (2013). "Video event understanding using natural language descriptions," in *Proc. IEEE International Conference on Computer Vision* (Sydney, NSW), 905–912.

Raptis, M., Kokkinos, I., and Soatto, S. (2012). "Discovering discriminative action parts from mid-level video representations," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1242–1249.

Rawlinson, G. (2007). The significance of letter position in word recognition. *IEEE Aerosp. Electron. Syst. Mag.* 22, 26–27. doi:10.1109/MAES.2007.327521

Reddy, K. K., and Shah, M. (2013). Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* 24, 971–981. doi:10.1007/s00138-012-0450-4

Robertson, N., and Reid, I. (2006). A general method for human activity recognition in video. *Comput. Vis. Image Understand.* 104, 232–248. doi:10.1016/j.cviu.2006.07.006

Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). "Action MACH: a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Rodríguez, N. D., Cuéllar, M. P., Lilius, J., and Calvo-Flores, M. D. (2014). A survey on ontologies for human behavior recognition. *ACM Comput. Surv.* 46, 1–33. doi:10.1145/2523819

Rohrbach, M., Amin, S., Mykhaylo, A., and Schiele, B. (2012). "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1194–1201.

Roshtkhari, M. J., and Levine, M. D. (2013). Human activity recognition in videos using a single example. *Image Vis. Comput.* 31, 864–876. doi:10.1016/j.imavis.2013.08.005

Rudovic, O., Petridis, S., and Pantic, M. (2013). "Bimodal log-linear regression for fusion of audio and visual features," in *Proc. ACM Multimedia Conference* (Barcelona), 789–792.

Sadanand, S., and Corso, J. J. (2012). "Action bank: a high-level representation of activity in video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1234–1241.

Salakhutdinov, R., Torralba, A., and Tenenbaum, J. B. (2011). "Learning to share visual appearance for multiclass object detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 1481–1488.

Samanta, S., and Chanda, B. (2014). Space-time facet model for human activity classification. *IEEE Trans. Multimedia* 16, 1525–1535. doi:10.1109/TMM.2014.2326734

Sanchez-Riera, J., Cech, J., and Horaud, R. (2012). "Action recognition robust to background clutter by using stereo vision," in *Proc. European Conference on Computer Vision* (Firenze), 332–341.

Sapienza, M., Cuzzolin, F., and Torr, P. H. S. (2014). Learning discriminative space-time action parts from weakly labelled videos. *Int. J. Comput. Vis.* 110, 30–47. doi:10.1007/s11263-013-0662-8

Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. Multimedia* 9, 1396–1403. doi:10.1109/TMM.2007.906583

Satkin, S., and Hebert, M. (2010). "Modeling the temporal extent of actions," in *Proc. European Conference on Computer Vision* (Heraklion), 536–548.

Schindler, K., and Gool, L. V. (2008). "Action snippets: how many frames does human action recognition require?," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Schuldt, C., Laptev, I., and Caputo, B. (2004). "Recognizing human actions: a local SVM approach," in *Proc. International Conference on Pattern Recognition* (Cambridge), 32–36.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011). "Avec 2011 -the first international audio visual emotion challenge," in *Proc. International Audio/Visual Emotion Challenge and Workshop, Lecture Notes in Computer Science*, Vol. 6975 (Memphis, TN), 415–424.

Sedai, S., Bennamoun, M., and Huynh, D. Q. (2013a). Discriminative fusion of shape and appearance features for human pose estimation. *Pattern Recognit.* 46, 3223–3237. doi:10.1016/j.patcog.2013.05.019

Sedai, S., Bennamoun, M., and Huynh, D. Q. (2013b). A Gaussian process guided particle filter for tracking 3D human pose in video. *IEEE Trans. Image Process.* 22, 4286–4300. doi:10.1109/TIP.2013.2271850

Seo, H. J., and Milanfar, P. (2011). Action recognition from one example. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 867–882. doi:10.1109/TPAMI.2010.156

Shabani, A. H., Clausi, D., and Zelek, J. S. (2011). "Improved spatio-temporal salient feature detection for action recognition," in *Proc. British Machine Vision Conference* (Dundee), 1–12.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton NJ: Princeton University Press.

Shao, J., Kang, K., Loy, C. C., and Wang, X. (2015). "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4657–4666.

Shivappa, S., Trivedi, M. M., and Rao, B. D. (2010). Audiovisual information fusion in human-computer interfaces and intelligent environments: a survey. *Proc. IEEE* 98, 1692–1715. doi:10.1109/JPROC.2010.2057231

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., et al. (2011). "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 1297–1304.

Shu, T., Xie, D., Rothrock, B., Todorovic, S., and Zhu, S. C. (2015). "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4576–4584.

Siddiquie, B., Khan, S. M., Divakaran, A., and Sawhney, H. S. (2013). "Affect analysis in natural human interaction using joint hidden conditional random fields," in *Proc. IEEE International Conference on Multimedia and Expo* (San Jose, CA), 1–6.

Sigal, L., Isard, M., Haussecker, H. W., and Black, M. J. (2012a). Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation. *Int. J. Comput. Vis.* 98, 15–48. doi:10.1007/s11263-011-0493-4

Sigal, L., Isard, M., Haussecker, H., and Black, M. J. (2012b). Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation. *Int. J. Comput. Vis.* 98, 15–48. doi:10.1007/s11263-011-0493-4

Singh, S., Velastin, S. A., and Ragheb, H. (2010). "Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods," in *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance* (Boston, MA), 48–55.

Singh, V. K., and Nevatia, R. (2011). "Action recognition in cluttered dynamic scenes using pose-specific part models," in *Proc. IEEE International Conference on Computer Vision* (Barcelona), 113–120.

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. doi:10.1023/B:STCO.0000035301.49549.88

Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. (2005). "Early versus late fusion in semantic video analysis," in *Proc. Annual ACM International Conference on Multimedia* (Singapore), 399–402.

Soleymani, M., Pantic, M., and Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Trans. Affective Comput.* 3, 211–223. doi:10.1109/T-AFFC.2011.37

Song, Y., Morency, L. P., and Davis, R. (2012a). "Multimodal human behavior analysis: learning correlation and interaction across modalities," in *Proc. ACM International Conference on Multimodal Interaction* (Santa Monica, CA), 27–30.

Song, Y., Morency, L. P., and Davis, R. (2012b). "Multi-view latent variable discriminative models for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 2120–2127.

Song, Y., Morency, L. P., and Davis, R. (2013). "Action recognition by hierarchical sequence summarization," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 3562–3569.

Soomro, K., Zamir, A. R., and Shah, M. (2012). *UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild*. Cornell University Library. CoRR, abs/1212.0402.

Sun, C., and Nevatia, R. (2013). "ACTIVE: activity concept transitions in video event classification," in *Proc. IEEE International Conference on Computer Vision* (Sydney, NSW), 913–920.

Sun, Q. S., Zeng, S. G., Liu, Y., Heng, P. A., and Xia, D. S. (2005). A new method of feature fusion and its application in image recognition. *Pattern Recognit.* 38, 2437–2448. doi:10.1016/j.patcog.2004.12.013

Sun, X., Chen, M., and Hauptmann, A. (2009). "Action recognition via local descriptors and holistic features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (Los Alamitos, CA), 58–65.

Tang, K. D., Yao, B., Fei-Fei, L., and Koller, D. (2013). "Combining the right features for complex event recognition," in *Proc. IEEE International Conference on Computer Vision*, pages (Sydney, NSW), 2696–2703.

Tenorth, M., Bandouch, J., and Beetz, M. (2009). "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Proc. IEEE International Workshop on Tracking Humans for the Evaluation of Their Motion in Image Sequences (THEMIS)* (Kyoto), 1089–1096.

Theodorakopoulos, I., Kastaniotis, D., Economou, G., and Fotopoulos, S. (2014). Pose-based human action recognition via sparse representation in dissimilarity space. *J. Vis. Commun. Image Represent.* 25, 12–23. doi:10.1016/j.jvcir.2013.03.008

Theodoridis, S., and Koutroumbas, K. (2008). *Pattern Recognition*, Fourth Edn. Boston: Academic Press.

Thurau, C., and Hlavac, V. (2008). "Pose primitive based human action recognition in videos or still images," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Tian, Y., Sukthankar, R., and Shah, M. (2013). "Spatiotemporal deformable part models for action detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Portland, OR), 2642–2649.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 1, 211–244. doi:10.1162/15324430152748236

Toshev, A., and Szegedy, C. (2014). "Deeppose: human pose estimation via deep neural networks," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 1653–1660.

Tran, D., Yuan, J., and Forsyth, D. (2014a). Video event detection: from subvolume localization to spatiotemporal path search. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 404–416. doi:10.1109/TPAMI.2013.137

Tran, K. N., Gala, A., Kakadiaris, I. A., and Shah, S. K. (2014b). Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognit. Lett.* 44, 49–57. doi:10.1016/j.patrec.2013.09.015

Tran, K. N., Kakadiaris, I. A., and Shah, S. K. (2012). Part-based motion descriptor image for human action recognition. *Pattern Recognit.* 45, 2562–2572. doi:10.1016/j.patcog.2011.12.028

Turaga, P. K., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: a survey. *Proc. IEEE Trans. Circuits Syst. Video Technol.* 18, 1473–1488. doi:10.1109/TCSVT.2008.2005594

Urtasun, R., and Darrell, T. (2008). "Sparse probabilistic regression for activity-independent human pose inference," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH), 588–595.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). "Canal9: a database of political debates for analysis of social interactions," in *Proc. International Conference on Affective Computing and Intelligent Interaction and Workshops* (Amsterdam: De Rode Hoed), 1–4.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). "Show and tell: a neural image caption generator," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3156–3164.

Vrigkas, M., Karavasilis, V., Nikou, C., and Kakadiaris, I. A. (2013). "Action recognition by matching clustered trajectories of motion vectors," in *Proc. International Conference on Computer Vision Theory and Applications* (Barcelona), 112–117.

Vrigkas, M., Karavasilis, V., Nikou, C., and Kakadiaris, I. A. (2014a). Matching mixtures of curves for human action recognition. *Comput. Vis. Image Understand.* 119, 27–40. doi:10.1016/j.cviu.2013.11.007

Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2014b). "Classifying behavioral attributes using conditional random fields," in *Proc. 8th Hellenic Conference on Artificial Intelligence, Lecture Notes in Computer Science*, Vol. 8445 (Ioannina), 95–104.

Wang, H., Kläser, A., Schmid, C., and Liu, C. L. (2011a). "Action recognition by dense trajectories," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3169–3176.

Wang, J., Chen, Z., and Wu, Y. (2011b). "Action recognition with multiscale spatio-temporal contexts," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 3185–3192.

Wang, Y., Guan, L., and Venetsanopoulos, A. N. (2011c). "Kernel cross-modal factor analysis for multimodal information fusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (Prague), 2384–2387.

Wang, H., Kläser, A., Schmid, C., and Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* 103, 60–79. doi:10.1007/s11263-012-0594-8

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012a). "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1290–1297.

Wang, S., Yang, Y., Ma, Z., Li, X., Pang, C., and Hauptmann, A. G. (2012b). "Action recognition by exploring data distribution and feature correlation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1370–1377.

Wang, Z., Wang, J., Xiao, J., Lin, K. H., and Huang, T. S. (2012c). "Substructure and boundary modeling for continuous action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1330–1337.

Wang, L., Hu, W., and Tan, T. (2003). Recent developments in human motion analysis. *Pattern Recognit.* 36, 585–601. doi:10.1016/S0031-3203(02)00100-0

Wang, S., Ma, Z., Yang, Y., Li, X., Pang, C., and Hauptmann, A. G. (2014). Semi-supervised multiple feature analysis for action recognition. *IEEE Trans. Multimedia* 16, 289–298. doi:10.1109/TMM.2013.2293060

Wang, Y., and Mori, G. (2008). "Learning a discriminative hidden part model for human action recognition," in *Proc. Annual Conference on Neural Information Processing Systems* (Vancouver, BC), 1721–1728.

Wang, Y., and Mori, G. (2010). "A discriminative latent model of object classes and attributes," in *Proc. European Conference on Computer Vision* (Heraklion), 155–168.

Wang, Y., and Mori, G. (2011). Hidden part models for human action recognition: probabilistic versus max margin. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1310–1323. doi:10.1109/TPAMI.2010.214

Westerveld, T., de Vries, A. P., van Ballegooij, A., de Jong, F., and Hiemstra, D. (2003). A probabilistic multimedia retrieval model and its evaluation. *EURASIP J. Appl. Signal Process.* 2003, 186–198. doi:10.1155/S111086570321101X

Wu, C., Zhang, J., Savarese, S., and Saxena, A. (2015). "Watch-n-patch: unsupervised understanding of actions and relations," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 4362–4370.

Wu, Q., Wang, Z., Deng, F., Chi, Z., and Feng, D. D. (2013). Realistic human action recognition with multimodal feature selection and fusion. *IEEE Trans. Syst. Man Cybern. Syst.* 43, 875–885. doi:10.1109/TSMCA.2012.2226575

Wu, Q., Wang, Z., Deng, F., and Feng, D. D. (2010). "Realistic human action recognition with audio context," in *Proc. International Conference on Digital Image Computing: Techniques and Applications* (Sydney, NSW), 288–293.

Wu, X., Xu, D., Duan, L., and Luo, J. (2011). "Action recognition using context and appearance distribution features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO), 489–496.

Xiong, Y., Zhu, K., Lin, D., and Tang, X. (2015). "Recognize complex events from static images by fusing deep channels," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1600–1609.

Xu, C., Hsieh, S. H., Xiong, C., and Corso, J. J. (2015). "Can humans fly? Action understanding with multiple classes of actors," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 2264–2273.

Xu, R., Agarwal, P., Kumar, S., Krovi, V. N., and Corso, J. J. (2012). "Combining skeletal pose with local motion for human activity recognition," in *Proc. International Conference on Articulated Motion and Deformable Objects* (Mallorca), 114–123.

Yan, X., Kakadiaris, I. A., and Shah, S. K. (2014). Modeling local behavior for predicting social interactions towards human tracking. *Pattern Recognit.* 47, 1626–1641. doi:10.1016/j.patcog.2013.10.019

Yan, X., and Luo, Y. (2012). Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier. *Neurocomputing* 87, 51–61. doi:10.1016/j.neucom.2012.02.002

Yang, W., Wang, Y., and Mori, G. (2010). "Recognizing human actions from still images with latent poses," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, CA), 2030–2037.

Yang, Y., Saleemi, I., and Shah, M. (2013). Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1635–1648. doi:10.1109/TPAMI.2012.253

Yang, Z., Metallinou, A., and Narayanan, S. (2014). Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues. *IEEE Trans. Multimedia* 16, 1766–1778. doi:10.1109/TMM.2014.2328311

Yao, A., Gall, J., and Gool, L. V. (2010). "A Hough transform-based voting framework for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, CA), 2061–2068.

Yao, B., and Fei-Fei, L. (2010). "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, CA), 17–24.

Yao, B., and Fei-Fei, L. (2012). Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1691–1703. doi:10.1109/TPAMI.2012.67

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L. J., and Fei-Fei, L. (2011). "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE International Conference on Computer Vision* (Barcelona), 1331–1338.

Ye, M., Zhang, Q., Wang, L., Zhu, J., Yangg, R., and Gall, J. (2013). "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging, Lecture Notes in Computer Science*, Vol. 8200. eds M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb (Berlin Heidelberg: Springer), 149–187.

Yi, S., Krim, H., and Norris, L. K. (2012). Human activity as a manifold-valued random process. *IEEE Trans. Image Process.* 21, 3416–3428. doi:10.1109/TIP.2012.2197008

Yu, G., and Yuan, J. (2015). "Fast action proposals for human action detection and search," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1302–1311.

Yu, G., Yuan, J., and Liu, Z. (2012). "Propagative Hough voting for human activity recognition," in *Proc. European Conference on Computer Vision* (Florence), 693–706.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D. (2012). "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (Providence, RI), 28–35.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 39–58. doi:10.1109/TPAMI.2008.52

Zhang, Z., Wang, C., Xiao, B., Zhou, W., and Liu, S. (2013). Attribute regularization based human action recognition. *IEEE Trans. Inform. Forensics Secur.* 8, 1600–1609. doi:10.1109/TIFS.2013.2258152

Zhang, Z., Wang, C., Xiao, B., Zhou, W., and Liu, S. (2015). Robust relative attributes for human action recognition. *Pattern Anal. Appl.* 18, 157–171. doi:10.1007/s10044-013-0349-3

Zhou, Q., and Wang, G. (2012). "Atomic action features: a new feature for action recognition," in *Proc. European Conference on Computer Vision* (Firenze), 291–300.

Zhou, W., and Zhang, Z. (2014). Human action recognition with multiple-instance Markov model. *IEEE Trans. Inform. Forensics Secur* 9, 1581–1591. doi:10.1109/TIFS.2014.2344448

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.