RESEARCH ARTICLE OPEN ACCESS

"SEQUENTIAL HUMAN ACTIVITY RECOGNITION": A REVIEW

[1] Mr.Jogi John ,[2] Abhinav Agarwal, [3] Vaishnavi Nasery, [4] Vaishnavi Sahu , [5] Khush Dhande, [6] Niharika Jagne

[1] Assistant Professor,Department of Computer Technology Priyadarshini College of Engineering – Nagpur

[2],[3],[4],[5],[6] Student Department of Computer Technology,Priyadarshini College of Engineering – Nagpur - India

ABSTRACT

Human Activity Recognition is one of the active research areas in computer vision for various contexts like security surveillance, healthcare and human computer interaction. Detecting human activity from video sequences or still images is a challenging task because of problems, such as background clutter, slows shut-off, scale changes, vision, brightness, and appearance. Many applications, including video surveillance systems, human computer interactions, and robots to measure human behavior, require a multi-tasking system. In this work, we provide a detailed review of the latest developments and advanced research in the field of human resource segregation. We propose a breakdown of human resource methods and discuss their benefits and limitations. In particular, we classify methods for classifying human activities into two broad categories depending on whether they use data for different methods or not. Then, each of these categories is re-analyzed into sub-categories, which reflect how they reflect human activities and what kind of activities they are interested in. In addition, we provide a comprehensive analysis of existing, publicly available segregated data sets and assess the requirements for an appropriate database to monitor human activity. Finally, we report features for future research guides and introduce some open-ended issues in the recognition of human activity.

Keywords*:* Human Activity recognition, sensing technology, depth sensor, media pipe, python.

I. INTRODUCTION

Recognition of human activity plays an important role in human interactions and interpersonal relationships. Because it conveys the idea of identity, personality, and attitude, it is difficult to dismiss. The ability to see into another person's work is one of the main areas of computer science research and machine learning. As a result of this study, many applications, including video surveillance systems, human computer interactions, and robots to measure human behavior, require a multidisciplinary monitoring system.

Among the various strategies for separation arise two key questions: "What action?" (i.e., problem recognition) and "Where in the video?" (that is, local performance problem). When attempting to detect human activity, one has to determine the kinetic conditions of the individual, so that the computer can detect this activity effectively. Human activities, such as "walking" and "running," occur naturally in everyday life and are easy to detect.

We live in a digital world where we see everything through digital glass. Technology has taken away some of the easiest and most difficult tasks. The need for someone else to be with us to do simple work is diminishing.

Computer vision and deep learning are the foundation of what we call the digital vigilance of our modern era. Whether it is safety, learning, or daily activities, computer vision helps us to monitor relevant results without human effort. With advanced computer vision technology, we can detect, monitor, and control results in our own way.

One of the most important parts of computer vision is to measure the shape of the human body. Whole body sculpting involves seeing the limbs and limbs and removing light, clothing, and sound from an image. Doing this over a live video stream adds to the challenge. Here MediaPipe and OpenCV start working. We will discuss what MediaPipe and OpenCV are, how it speeds up the process of body measurement, and how it works. This article will consider how to use this technology to create a tool that helps individuals to exercise properly without trainers.

OpenCV (Open-source Computer Vision) is an open source library. Provides a list of tools for image processing, extraction, and decryption. The library is cross-platform and supports Windows, Linux, and macOS and other platforms. Other OpenCV applications that assist in Pose Estimation include movement tracking, touch detection, and structure from Motion.

MediaPipe is also an open source platform that provides integrated, ready-to-use, fast-paced machine learning solutions for live streaming media. Provides machine learning solutions - known as pipelines, for multilingual communication. Other ML solutions in MediaPipe include face recognition, iris movement, fullness, position measurement, and face mesh.

Mediapipe pose measurement: This is a standalone measurement method developed by google researchers and works with a flaming model of how to find a pose. Fast and efficient model with 24FPS mouse and ready to adjust live video format. The Blazepose model returns 33 key points or location symbols from a given image where a person is found. These points are the major combined points of the human body and the returned points are 3-D links with visual value. With an invisible member, it predicts member links using the concept of Leonardo's Vitruvian man and therefore the central area of the human hip, a circular area that includes the human angle and the line angle associated with the shoulder and midpoint of the hip.

Research gaps: Most of the work in the recognition of a person's work takes place in the middle of a pure background, where the character is free to do the work. The development of an automated human recognition system, which is able to differentiate one's tasks with low errors, is a challenging task due to problems, such as background, slow closure, scale changes, vision, brightness and appearance, and frame adjustment. . In addition, defining ethical roles is time consuming and requires knowledge of a particular event. Moreover, the similarity of intra- and interclass makes the problem even more challenging. That is, actions within the same category may be expressed by different people with different body movements, and actions between different categories may be difficult to distinguish as they may be represented by the same information. How people do a particular job depends on their habits, and this makes the problem of identifying basic work more difficult to determine. Also, building a visual model for learning and analyzing human movement in real time with insufficient scale datasets for testing is challenging. To overcome these problems, a three-component work is required, namely: (i) background removal (Elgammal et al, 2002; Mumtaz et al., 2014), in which the system attempts to separate image components. that do not change over time (background) in moving or changing objects (front); (ii) human tracking, in which the system records the movement of a person over time (Liu et al., 2010; Wang et al., 2013; Yan et al., 2014); and (iii) human action and object acquisition (Pirsiavash and Ramanan, 2012; Gan et al., 2015; Jainy et al., 2015), in which the system is able to perform human activity locally in an image [30].

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 [14.]

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) [5].

II. RELATED WORK

There are a few surveys in the literature recognizing human activities. Gavrila (1999) categorized the study into 2D (with a clear and external model) and 3D methods. In Aggarwal and Cai (1999), a new tax was introduced focusing on human movement analysis, single-view tracking and multi-view cameras, and human activity recognition. Similar to air and previous taxonomy, Wang et al. (2003) have proposed a sequence of action sequences of action. A study by Moeslund et al. (2006) focused on location-based recognition methods and proposed a fourfold tax, which included the implementation of human movement, tracking, position measurement, and monitoring methods [1].

A good distinction between the definitions of "action" and "activity" was proposed by Turaga et al. (2008), in which task recognition methods were categorized according to their level of complex function. Poppe (2010) has developed methods for recognizing human activities into two main categories, defining "top-down" and "top-down." On the other hand, Agarwal and Ryoo (2011) introduced tree-based taxonomy, in which the methods of recognizing human activity were divided into two main categories, "single-layer" and "sequential" methods, each of which has several layers [8].

3D modeling is also a new practice, and it was extensively researched by Chen et al. (2013b) and He et al. (2013). Since the human body is made up of interlocking organs, one can model these parts using powerful features, located on deep cameras, and create 3D representations of the human body, much more instructive than the 2D functional simulations of an image plane. Aggarwal and Xia (2014) recently introduced the separation of human recognition methods from 3D stereo and motion pictures with the main focus on methods using 3D depth data. To date, Microsoft Kinect has played a major role in capturing specific skeletal movements using deep sensors [2].

Although much of the research has focused on programs to monitor human activity from video sequencing, the detection of human activity from still images remains an open and challenging task. Most human activity recognition studies are associated with facial recognition and / or posture measurement methods. Guo and Lai (2014) summarize all methods of human activity recognition from vertical images and divide them into two main categories according to the output level and the type of features used in each method [3].

Jaimes and The Department (2007) proposed a multidisciplinary personal computer interaction focused on heart-to-heart communication from gestures, facial expressions, and speech. Pantic and Rothkrantz (2003) conducted a comprehensive study of human behavior recognition techniques that included non-verbal multimodal signals, such as facial expressions and voice. [21] Pantic et al. (2006) studied a number of high-level approaches to human behavior that include sensory and social indicators and included many open computer problems and how they can be effectively integrated into the human-computer interaction system. Zeng et al. (2009) presented a review of high-quality cardiac awareness systems that use visual and audio signals to identify autoimmune conditions and provide a list of related data sets to recognize human expression. [19] Bousmalis et al. (2013a) proposed the analysis of a number of non-verbal (i.e., visual and auditory signals) behavioral recognition methods and data sets of automatic contracts and discrepancies. Such social values may play a key role in analyzing social behavior, which is the key to public participation. Finally, a comprehensive analysis of ontologies for human behavior recognition from a data perspective and information representation was presented by Rodríguez et al. (2014) [9].

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.

2.1 Human Activity Categorization

The problem of segregating human activities remains a challenging task in computer vision for more than two decades. Previous activities to describe human behavior have shown great potential in this area. First, we classify human activity monitoring into two main categories: (i) unimodal and (ii) multimodal activity recognition methods according to the type of sensory data they use. Then, each of these two categories is re-analyzed into sub-categories depending on how they model human activities.

Unimodal methods represent human activities based on one-way data, such as images, and are further subdivided into: (i) space-time, (ii) stochastic, (iii) legal framework, and (iv) shape-based methods. .

Local timing mechanisms include task-monitoring mechanisms, representing human activities such 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 using mathematical models to represent human actions (e.g., hidden Markov models) (Lan et al., 2011; Iosifidis et al., 2012a). Legislative-based approaches use a set of rules to define human activities (Morariu and Davis, 2011; Chen and Grauman, 2012). Shape-based approaches best represent activities with a high degree of cognitive function in modeling the movement of human organs (Sigal et al., 2012b; Tran et al., 2012) [31,26,36,39].

Multimodal approaches incorporate elements collected from different sources (Wu et al., 2013) and are divided into three categories: (i) affective, (ii) behavioral, and (iii) social media [33].

Effective methods represent human activities in terms of emotional communication and the affected human condition (Liu et al., 2011b; Martinez et al., 2014). Behavioral behaviors aim to recognize behavioral traits, symptoms of multiple non-verbal behaviors, such as touch, facial expressions, and auditory symptoms (Song et al., 2012a; Vrigkas et al., 2014b). Finally, social media models model human behavior and behavior in a few aspects of human interaction in social events ranging from gestures, gestures, and speech (Patron-Perez et al., 2012; Marín-Jiménez et al., 2014). [11, 14.24].

Often, the terms "work" and "behavior" are used interchangeably in literature (Castellano et al., 2007; Song et al., 2012a). In this survey, we distinguished between the two terms in the sense that the word “activity” is used to describe the sequence of actions associated with a particular body movement. On the other hand, the term “behavior” is used to describe both activities and events associated with touch, emotional states, facial expressions, and individual sensory signals [25].

III. EXISTING METHODS

2.1 Unimodal Method

Methods of recognizing unusual human activity identify human activities in one-way data. Many existing methods represent human activities such as a set of visual elements extracted from video sequences or still images and recognize the original work label using several classification models (Kong et al., 2014a; Wang et al., 2014) [34]. Consistent approaches are appropriate for identifying human activities based on dynamic factors. However, the ability to see the basic phase only through movement is a challenging task in itself. The big problem is how we can ensure the continuity of the movement over time as the action occurs in the same way or in the middle of the video sequence. Some methods use captions for motion trajectories (Matikainen et al., 2009; Raptis et al., 2012), while others use full-length curve movements by tracking the flow flow characteristics (Vrigkas et al., 2014a).[31]

We divide unimodal approaches into four broad categories: (i) space-time, (ii) stochastic, (iii) legal framework, and (iv) contextual-based approaches. Each of these sub-categories describes specific attributes of how people perceive human activities in terms of the type to represent each method used.

2.1.1 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. (Wang et al., 2013) [33].

A number of methods for recognizing human activities based on representation of local time have been suggested in the literature (Efros et al., 2003; Schuldt et al., 2004; Jhuang et al., 2007; Fathi and Mori, 2008; Niebles et al., 2008). The large family of methods depends on the optical flow, which has proved to be an important indicator.[45] Efros and others. (2003) visual actions from low-resolution sports video tracking using a nearby neighbor separator, in which people are represented by windows up to 30 pixels high. Fathi and Mori's (2008) approach was based on intermediate dynamics, which were also built directly into optical flow characteristics. In addition, Wang and Mori (2011) used dynamic features such as inputs for random conditional compounds (HCRFs) (Quattoni et al., 2007) and vector machine class (SVM) divisions for support (Bishop, 2006) [16]. Real-time segregation and predicting future actions were proposed by Morris and Trivedi (2011), in which job vocabulary is learned through a three-step process. Other methods based on visual flow that received thunder were introduced by Dalal et al. (2006), Chaudhry et al. (2009), and Lin et al. (2009). Flexibility in interpreting and interpreting the scale presented by Oikonomopoulos et al. (2009). Spatiotemporal features based on B-splines are extracted from the optical flow field. To model this definition, the Bag-of-Words (BoW) method is used, and the division of labor is performed using appropriate vector equipment (RVM) (Tipping, 2001). [25]

Sequential video classification using local features in the spatiotemporal area is also very focused. Schultt et al. (2004) represent local events in the video using space-time features, while the SVM separator was used to detect action. Gorelick et al. (2007) [25] viewed actions as silhouettes of 3D time-lapse travelers. They used the Poisson equation solution to accurately describe the action by using a visual correlation between the sequence of elements and using the adjacent neighbor division to illustrate the action. Niebles et al. (2008) addressed the problem of action recognition by creating a codebook of points of interest for space time. The management approach was followed by Jhuang et al. (2007), in which a featured video was analyzed into descriptions of several features in terms of complexity. The final division is done by the SVM divider of many categories. Dollar et al. (2005) proposed spatiotemporal features based on cuboid definitions. Instead of coding human movements to separate the action, Jainy et al. (2015) propose to incorporate information from human interactions to objects and compile several databases to transfer information from one database to another.[43,44].

An action dictionary for histograms of interesting points, based on the work of Schuldt et al. (2004), presented by Yan and Luo (2012). The informal forest of action representation has also attracted the widespread interest in action recognition Mikolajczyk and Uemura (2008) and Yao et al. (2010). In addition, the important issue of how many frameworks are needed to detect the action was handled by Schindler and Gool (2008). Shabani et al. (2011) proposed a temporary unequal screening to determine the character and perception of a task. The extracted elements were significantly stronger under geometric modification than the features defined by the Gabor filter (Fogel and Sagi, 1989). Sapienza et al. (2014) used a spatiotemporal volume feature bag to identify and personalize local actions from video labels with a weak label using multiple learning example. [50, 11,24]

2.1.2 Stochastic Methods

In recent years, there has been a significant increase in the number of computer-assisted research studies aimed at understanding human activities. There has been an emphasis on jobs, where the business to be recognized can be considered a sequence of statistically predictable conditions. Researchers have conceived and used a number of stochastic techniques, such as the hidden Markov model (HMMs) (Bishop, 2006) and random conditional (HCRFs) (Quattoni et al., 2007), to obtain useful results for human activity recognition.[16]

Robertson and Reid (2006) likened human behavior to a sequence of actions. Each action is defined by a feature vector, which includes information about location, speed, and location descriptions. HMM was hired to record human actions, while recognition was made by searching for image elements representing the action [41]. Opening this work, Wang and Mori (2008) were among the first to raise HCRFs with the issue of job recognition. The human action is performed as the configuration of parts of the image view. Moving features were developed to create a BoW model. Activity and local recognition by central model presented by Lan et al. (2011). The human environment is treated as a subtle variable, extracted from a subtle variable model with a simultaneous detection of action. A real-time algorithm showing human interaction was proposed by Oliver et al. (2000). The algorithm was able to detect and track a person's movement, creating a feature vector that describes movement. This vector is given input to HMM, which is used for action separation. Ingoma et al. (2013) considered that the sequence of human action of different interim decisions. At each level of smoking, they learned a hierarchical model with hidden variables to collect the same semantic attributes for each layer.[16, 33, 41]

Sun and Nevatia (2013) treat video sequences as sets of short clips instead of representing the whole action. Each clip corresponds to a subtle variation in the HMM model, while the Fisher kernel strategy (Perronnin and Dance, 2007) was developed to represent each clip with a vertical vector length. Ngi et al. (2014) break down the problem of identifying complex work into two consecutive sub-tasks with increasing levels of granularity. First, the authors used interpersonal-to-object interaction techniques to identify the area of interest, and then used this information based on context to train the conditional random field model (CRF) (Lafferty et al., 2001) and to identify basic action. Lan et al. (2014) have proposed a hierarchical approach to predicting future human actions, which may be considered a response to a previous action. They introduced a new representation of human kinematic states, called "hierarchical movements," calculated from different degrees of roughness to fine granularity. Predicting future events from partial video clips with partial performance has also been studied by Kong et al. (2014b). A series of previously seen features were used as a universal representation of actions and a CRF model was used to capture the occurrence of actions over time in each action phase [3, 7.14.35].

The method of classifying group activities was introduced by Choi et al. (2011). Authors have been able to identify works such as a group of people speaking or standing in line. The proposed system was based on random forests, which can select spatiotemporal volume samples from the video showing the action. The Markov (MRF) random field framework (MRF) (Prince, 2012) may be used to classify and locate activities at the scene. Lu et al. (2015) also used the hierarchical MRF model to represent parts of human action by extracting super-voxels from different scales and automatically estimating forward movements using the remarkable features of neighboring super-voxels [43].

2.1.3 Rule-Based Methods

Law-based methods determine ongoing events by modeling the work using rules or sets of attributes that define an event. Each function is regarded as a set of ancient rules / attributes, which allow for the development of a descriptive model for the recognition of human activity.

Activity recognition of complex multi-topic scenes was proposed by Morariu and Davis (2011). Each topic should follow a set of specific rules while performing the action. The recognition process is made up of basketball game videos, in which players are first discovered and tracked, producing a set of trajectories used to create a set of spatiotemporal events. Based on the concept of the first order and possible methods, such as the Markov networks, the authors were able to determine which event occurred. Figure 6 summarizes their approach using the ancient rules of observing human actions. Liu et al. (2011a) faced the problem of recognizing actions through a set of descriptive and discriminatory attributes. Each attribute was associated with features that defined the spatiotemporal status of functions. These attributes are regarded as subtle variables, which capture the level of importance of each attribute for each action in the form of a hidden SVM [39].

A combination of job recognition and local practice was introduced by Chen and Grauman (2012). The whole approach was based on the construction of a space graph using a high-level adjective, in which the algorithm seeks to find a suitable subgraph that maximizes the classification function (i.e., obtain a low-weight graph below, which is a common phenomenon. Is a complete NP problem). Kuehne et al. (2014) proposed a systematic systematic approach to the daily recognition of human activity. The author used HMMs to model human actions as action units and then applied the program rules to create complex sequences of actions by combining different action units. When a temporary language system is used for action planning, the main problem involves treating long video sequences due to the complexity of the models. One way to deal with this limit is to divide video sequences into smaller clips containing smaller actions, using the sequence method (Pirsiavash and Ramanan, 2014). A brief descriptive generation from video follow-up (Vinyals et al., 2015) based on convolutional neural networks (CNN) (Ciresan et al., 2011) [37] has also been used for job recognition (Donahue et al., 2015).

Central semantic features for detecting invisible actions during training were proposed (Wang and Mori, 2010). These intermediate features were learned during training, while the parameter sharing between classes was enhanced by capturing the correlation between the most common low-level features (Akata et al., 2013). Learning how to identify new classes that have not been seen during training, by combining intermediate features with class labels, is a necessary element of transferring information between training and test samples. This problem is commonly referred to as zero-shot learning (Palatucci et al., 2009). Therefore, instead of studying one phase for each attribute, a two-step classification method has been proposed by Lampert et al. (2009). Specific attributes are predicated on class compilers that have been studied and are designed for class level schooling.

2.1.4 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. t 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.

Greater focus on visual perception from visual images or videos has been made in the context of visual scene (Thurau and Hlavac, 2008; Yang et al., 2010; Maji et al., 2011) [37]. Clearly, Thurau and Hlavac (2008) represent actions with histograms of pose primitives, and n-gram expressions were used to classify action. Also, Yang et al. (2010) integrated actions and personal stereotypes, treating posture as subtle differences, specifying the action label on vertical images. Water et al. (2011) [36] introduced a representation of the posture, called the “poselet activation vector,” which defined the 3D shape of the head and body and provided a strong representation of the posture and appearance. In addition, a classification of actions based on modeling the movement of parts of the human body was presented by Tran et al. (2012), in which less representation was used to model and recognize complex actions. In terms of template matching techniques, Rodriguez et al. (2008) introduced a high-resolution medium-length filter (MACH), which was a method for capturing intraclass variables by combining a single MACH filter action for a given action phase. Sedai et al. (2013a) proposed a combination of shapes and adjectives for appearance to represent the spatial features of measuring human shapes. Different types of definitions were integrated at the decision level using a discriminatory learning model. However, identifying which organs are most important in identifying human complex functions remains a challenging task (Lillo et al., 2014). The differentiation model and the specific examples representing the individual position estimation are shown.

Ikizler and Duygulu (2007) modeled the human body as a sequence of rectangular pads. The authors describe a variation of the BoW method called bag-of-rectangles. Area-based histograms are designed to describe human action, while action classification is performed using four different methods, such as independent voting, global histogramming, SVM classification, and dynamic time warping (DTW) (Theodoridis and Koutroumbas, 2008). Yao and Fei-Fei's (2012) study of human modeling of human interaction by presenting a similar contextual model. The types of posture, as well as the geographical relationships between different segments of the population, are modeled. Automatic maps (SOM) (Kohonen et al., 2001) presented by Iosifidis et al. (2012b) to study the shape of the human body, in terms of ambiguous distances, in order to achieve a consistent representation of action. The proposed algorithm was based on multi-layer perceptrons, in which each layer was fed to an associated camera, in order to set a consistent viewing action. Human interaction was discussed by Andriluka and Sigal (2012). First, the 2D human dimensions were estimated from image structures from groups of people and then each measured structure was placed in a 3D space. To date, several 2D posture benchmarks have been proposed to test methods for measuring posture (Andriluka et al., 2014) [3,7,14,30,31]. Activity recognition using deep cameras presented by Wang et al. (2012a), in which a new type of feature called “habitat pattern” was also proposed. This feature was variable in translation and was able to capture the correlation between the parts of the human body. The authors also proposed a new human action model called the “actionlet ensemble model,” which captures intraclass variability and is strong in error made by deep cameras. 3D human shapes have been considered in recent years and several algorithms for recognizing human activities have been developed. A recent review of 3D image stabilization and work visibility proposed by Holte et al. (2012b). [29]

2.2 Affective Methods

A major problem with affective computing data is accurately defined. Ratings are one of the most popular annotation tools. However, this is a challenge to find real world conditions, as emotional events are presented differently by different people or occur simultaneously with other activities and emotions. Pre-processing the annotations may be detrimental to the production of accurate and abstract models due to the biased representation of the annotation. To date, research on how to produce informative educational labels has been proposed by Healey (2011). Soleymani et al. (2012) investigated features of an independent emotional alert system that can detect educational tags from electroencephalogram (EEG) signals, pupillary reflex, and body responses associated with video stimulation. Nicolau et al. (2014) proposed a novel approach based on the possible analysis of canonical correlation (PCCA) (Klami and Kaski, 2008) and DTW to integrate multimodal emotional

annotations and to facilitate temporal sequencing [30,34,35].
Liu et al. (2011b) multimodal correlated (i.e., textual and visual) features for distinguishing affected conditions in vertical images. The authors assert that visual aids are not sufficient to comprehend human emotions, and thus additional information that explains the image is needed. Dempster-Shafer theory (Shafer, 1976) was employed to integrate various methods, while SVM was used for fragmentation. Hussain et al. (2011) proposed a framework for integrating multimodal psychological features, such as facial muscle function, skin reactions, and breathing, in order to detect and identify cardiac conditions. AlZoubi et al. (2013) examined the effect of variability on the affected factor over time in classifying the affected conditions [4].
Siddiquie et al. (2013) analyzed four different variables involved, such as performance, expected duration, intensity, and valence (Schuller et al., 2011). To date, they have proposed Randomized Combined Spaces (JHCRF) as a new segregation scheme for the benefit of multimodal data. Moreover, their method uses late compounding to combine audio information and visual information together. This can lead to a significant loss of intermodality dependence, while there is the problem of spreading the distortion error at different levels of dividers. Although their method was able to accurately identify the affected human condition, the calculation load was higher as JHCRFs required twice as many hidden alternatives than conventional HCRFs when the features represented two different modes [8, 21, and 26].

2.2.1 Multimodal Feature Fusion:

Imagine a situation where a few people have a particular activity / behavior and some of them may make noises. In the simplest case, the human activity recognition system may detect the basic function by considering only visual information. However, the accuracy of perception may be improved in sound and visual analysis, as different people may exhibit different functions associated with the same body movements, but with different sound levels. Audio information can help you understand who is interested in exploring video sequences and distinguishing between different behavioral situations.
The main difficulty in analyzing the multimodal aspect is the size of the data from the different types. For example, video features are more complex with higher magnitude than audio, so size reduction techniques are helpful. In the literature, there are two major integration strategies that can be used to address this problem (Atrey et al., 2010; Shivappa et al., 2010) [24].
Pre-integration, or integration at the element level, combines the features of a variety of methods, usually by reducing the size in each mode and creating a new feature vector representing each person. Canonical correlation analysis (CCA) (Hardoon et al., 2004) was widely studied in literature as an effective way to integrate data at the element level (Sun et al., 2005; Wang et al., 2011c; Rudovic et al., 2013). The advantage of early integration is that it produces good recognition results when different approaches are closely related, as only one learning phase is required. On the other hand, the difficulty of combining different approaches may lead to the domination of one method over another. The novel's method of combining words (i.e., textual information) with non-verbal (i.e., visual signals) was proposed by Evangelopoulos et al. (2013). Each method is analyzed separately and key points are used for straight and non-line integration schemes.[49]
The second category of methods, known as late merging or merging at the decision level, incorporates a few possible models to study the parameters of each method separately. Then all the points are grouped together in a supervised framework that provides the final decision score (Westerveld et al., 2003; Jiang et al., 2014). The individual strengths of each approach may lead to better recognition results. However, this strategy is time consuming and requires sophisticated supervised learning schemes, which can create potential losses of the methods used. Comparisons of early and recent compilation methods for video analysis were reported by Snoek et al. (2005) [17].

Recently, a third method of multimodal data integration has emerged earlier (Karpathy et al., 2014). This method, called slow fusion, is a combination of previous methods and can be seen as a composite method that combines data by transferring information sequentially to pre- and end-level integration rates. Although this method seems to have the advantages of both early and later integration methods, it also has a significant computational burden due to different levels of processing information [35].

IV. CONCLUSION

In this survey, we conducted a comprehensive study of high-quality methods for human activity recognition and proposed a hierarchical tax to differentiate these methods. We explored different methods, which were divided into two broad categories (unimodal and multimodal) depending on the source channel for each of these methods used to monitor human activities. We discussed the same methods and provided the internal division of these methods, which are designed to analyze touch, atomic actions, and more complex functions, directly or indirectly using function decay into simple actions. We also introduce multimodal methods for social behavior analysis and interaction. We discussed the different levels of feature representation and reported the limitations and benefits of each representation. A comprehensive benchmark review of human resource segregation was also presented and we examined the challenges of data acquisition on the problem of understanding human activity. Finally, we have provided the features to build a system to monitor human activities.
Most of the studies in this field have failed to accurately describe human activities in a concise and informative way as they introduce limitations on numeracy problems. The gap in the overall representation of human activities and the data collection associated with the annotation remains a challenging and unresolved issue. In particular, we may conclude that in spite of the rapid growth of human perceptions, many problems remain open, including human modeling, gripping, and annotation data.

REFERENCES

1. 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.

2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. Aggarwal, J. K., and Ryoo, M. S. (2011). Human activity analysis: a review. ACM Comput. Surv. 43, 1–43. doi:10.1145/1922649.1922653
9. 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.
10. nirudh, 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.
11. 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
12. 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.
13. 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
14. 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.
15. 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.
16. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Secaucus, NJ: Springer.
17. 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.
18. 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.
19. 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
20. 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
21. 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.
22. 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.
23. 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.
24. 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
25. 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.

26. 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

27. 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

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

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

47. 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

48. 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

49. 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

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