Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

# Autonomous Recognition of Human Activity to Aid in the Development of Robots

Arjun Singh[1], Hemant Nautiyal[2], Viveksheel Yadav[3]

[1]Assistant Professor, Electronics and Communication, School of Engineering, Dev Bhoomi Uttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007

[2, 3]Assistant Professor, Mechanical Engineering, School of Engineering, Dev Bhoomi Uttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007

[1]chancelloroffice@dbuu.ac.in, [2]me.hemant@dbuu.ac.in, [3]me.viveksheel@dbuu.ac.in

**Article Info**
**Page Number: 2543-2552**
**Publication Issue:**
**Vol. 71 No. 4 (2022)**

**Article History**
**Article Received: 25** March 2022
**Revised: 30** April 2022
**Accepted: 15** June 2022
**Publication: 19** August 2022

Abstract

It is a challenging effort due to challenges such as backdrop clutter, partial occlusion, differences in size, perspective, lighting, and appearance to identify human activities captured in video sequences or still images. A wide variety of applications, such as video surveillance systems, human-computer interfaces, and robots for human behavior classification, each need their own unique activity identification system in order to function properly. In this article, we provide a comprehensive review of recent and cutting-edge research accomplishments in the field of classifying human activities. We begin by presenting a taxonomy of human activity research approaches and then proceed to evaluate the benefits and drawbacks associated with each approach. We divide human activity classification algorithms into two primary categories according to whether or not they use data from a wide variety of different modalities. The first group is described as "not using data from a wide variety of modalities," whereas the second group is described as "utilizing data from a wide variety of modalities." After that, each of these categories is further separated into subcategories that represent how they copy human behaviors and the sort of activities in which they are engaged. These subcategories are also based on how they interact with humans. In addition, we provide a detailed analysis of the human activity classification datasets that are already accessible to the general public, as well as an investigation into the characteristics that should be met by an ideal human activity identification dataset. Both of these analyses can be found in the following paragraphs. In conclusion, we address certain issues that still need to be resolved in terms of human activity identification and describe the characteristics of possible future study fields. Humanoid robots often use a dialogue system that is based on pre-programmed templates. This kind of system can react well inside a certain discourse domain; nevertheless, it is unable to respond appropriately to information that falls outside of that discourse domain. The rules for the dialogue system are created by hand rather than being formed automatically. This is due to the fact that the interactive elements do not have a method for detecting emotions. Both a humanoid robot open-domain chat system and a deep neural network emotion analysis model were developed specifically for the aim of achieving this goal. The former is intended to do an investigation on the feelings that may be experienced

by interacting items. Analysis of a person's emotional state is a component of this method, along with studies on Word2vec and language coding. After that, the emotional state of a humanoid robot is taught by using a Training and emotional state analysis paradigm, which is quite specific. Creating templates is at the heart of the conventional approach for humanoid robots to carry on conversations. This approach is able to create appropriate responses while inside the designated discussion zone, but it cannot do so outside of that zone. The rules of the communication system are made by hand, and there is no emotional recognition included in them. This research developed an open-domain communication system for a humanoid robot as well as an emotion analysis model based on a deep neural network. Both of these were accomplished via the same study. The model was used in order to determine how the interacting items felt about one another. Language processing, coding, feature analysis, and Word2vec are all necessary components of an emotional state analysis. The findings of an emotional state analysis training session conducted on a humanoid robot are broken down and discussed in this article, along with the implications of those findings. Robots have gradually permeated every facet of human existence in recent years as a result of advances in science and technology. Robots are used in many different fields, including manufacturing, the military, home healthcare, education, and laboratories [1]. According to the three tenets that serve as the foundation of robotics [2, 3], the ultimate aim of robot development is to attain human-like behavior in robots, to assist people in doing their duties in a more efficient manner, and to realize their ambitions. Individuals need to improve the quality of their communication with the robot in order for the human-robot cooperation project to be successful [4, 5]. A person will typically interact with a computer by inputting data using a keyboard, mouse, and a variety of other manual input devices, while the computer will typically output data to a person via a display and a variety of other peripherals. This is the standard method of human-computer interaction. For this interaction, quite a few more resources are required. There are some people in the actual world who do not have access to computers [6]. Natural ways of communication between people and technology include the use of speech, vision, touch, hearing, proximity, and other human interactions [7]. This form of connection is not only common but also advantageous to both parties involved. [8] In order to facilitate more fruitful collaboration between people and robots. [9] The emotion analysis model of the humanoid robot is able to assess and detect the emotional information of the interacting object when the object is engaging with one another. The language of the item provides a great deal of emotional information when it is touched, and the written content demonstrates a deep comprehension of human thought.

**Keywords:** Dense HOG, depth sensor, feature-level fusion, human action recognition, an inertial sensor, and an RGB.

## 1. INVESTIGATION

Recognition of human activity has a significant impact on the dynamics of human-to-human interactions as well as interpersonal connections. Prints on a person's fingers might provide information about their identity as well as their personality and mental health. Researchers in the fields of computer vision and machine learning are particularly

interested in the ways in which people perceive the actions of other people. As a direct result of this body of work, a multimodal activity detection system is now required for the purpose of characterizing human behavior in a wide range of applications. These applications include video surveillance systems, human-computer interaction, and robotics.

Two of the questions that are asked the most often to classify anything are "What action?" and "What action?" ("Where are you in the video?" and "Where are you in the video?" both refer to the same place in the film. (that is, the issue with the localization) It is necessary to determine the kinetic states of a human being before the computer can accurately detect the activity. In normal life, human activities such as "walking" and "running" may be identified from one another in a straightforward manner. On the other hand, more intricate activities, such as "peeling an apple," are more challenging to comprehend. It is possible to simplify difficult tasks by breaking them up into a series of smaller, more manageable steps. By analyzing the components of a scene and how they connect to the events occurring in the surrounding area, we might potentially get a deeper understanding of human behaviors (Gupta and Davis, 2007).

The vast majority of research on human activity recognition operates on the assumption of a figure-centric scene in which there is no background clutter and the actor is free to execute an activity. Human activity detection is a challenging topic to tackle due to the difficulties presented by issues such as backdrop clutter and partial opacity. Additional issues include shifts in dimensions, shifts in viewpoint, shifts in lighting, and changes in frame resolution. In addition, annotating behavioral roles is a time-consuming process that requires a solid understanding of the individual event in order to be effective. Because of the many ways in which the different classes are similar, finding a solution to the problem is made much more difficult. Because members of various classes often use the same body gestures, it may be difficult to differentiate the actions performed by members of different classes from those performed by members of the same class. Because different people carry out the same task in a variety of different ways, it may be difficult to pin down exactly what it is that they are doing. It is difficult to construct a real-time visual model for learning and judging human movement since there are so few benchmark datasets available.

There are three elements that must be present in order to solve these problems: 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 those that move or change (foreground); (ii) human tracking, in which the system locates human motion over time; and (iii) acquiescence detection. I background subtraction is a technique that was developed by Elgammal and colleagues in 2002 and

Examining people's movements in video or still photographs may be done with the use of a technology called human activity recognition. Because of this reality, human activity

identification algorithms are driven to properly classify input data into the category that best describes the activity that it represents. Activities performed by humans may be broken down into the following categories: gestures; atomic actions; human-to-object or human-to-human interactions; collective behaviors; and events. A classification of human activities according to their level of complexity is shown in Figure 1.

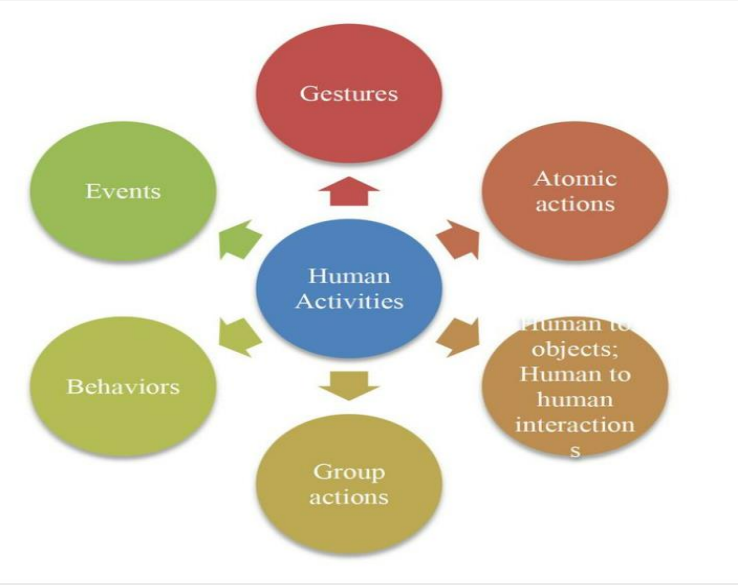


**Figure 1. Decomposition of human activities**

A person will make a basic movement of their body in order to communicate a concept or an action via the use of a gesture (Yang et al., 2013). "Atomic actions" are a person's movements that characterize a specific motion that may be a component of a more complex activity. These motions may be performed alone or as part of a larger activity (Ni et al., 2015). Both "human to human" and "human to object" interactions are terms that may be used to describe interactions between people and other objects (Patron-Perez et al., 2012). Activities that are carried out by a group of people are referred to as "group actions" (Tran et al., 2014b). The term "human behavior" refers to activities that are physically carried out by a person that are related to that individual's feelings, personality, and mental state (Martinez et al., 2014). In a nutshell, events are high-level acts that reveal people's intentions or social positions and describe the behaviors of the individuals who participate in them (Lan et al., 2012a).

The subject of human activity detection has been the subject of a wide range of studies throughout the years. Gavrila (1999) distinguished between 2D research procedures and 3D research methodologies based on whether or not explicit form models were used. Analysis of human motion, tracking from single and multiview cameras, and activity identification were the primary focuses of Aggarwal and Cai's research (1999). It has been decided to design a categorization hierarchy for actions very much like the one that Wang et al. (2003) presented. The study that was carried out by Moeslund et al. (2006) concentrated mostly on posture-based action recognition approaches. They also suggested

a four-tier taxonomy that encompassed the commencement of human motion, tracking, pose estimate, and recognition methods. '

When classifying activity detection systems, Turaga et al. (2008) found that the concepts of "action" and "activity" are very diverse from one another. This difference is noteworthy. In his study, Poppe (2010) compared and contrasted two approaches to identifying human activities: "top-down" and "bottom-up." However, Aggarwal and Ryoo (2011) propose a taxonomy of human activity identification methods. Their taxonomy is based on a tree-structured taxonomy that divides the methods into two main categories: "single layer" approaches and "hierarchical" approaches, both of which include several layers of categorization. This taxonomy divides the methods into two main categories: "single layer" approaches and "hierarchical" approaches.

In scholarly writing, the terms "activity" and "behavior" are often used synonymously with one another (Castellano et al., 2007; Song et al., 2012a). In light of the fact that the term "activity" refers to a collection of actions that correspond to a particular bodily movement, it is essential to distinguish between these two concepts within the context of this study. The term "behavior" may refer to either the activities themselves or the occurrences that are connected to a single person's gestures, emotions, facial expressions, and voice cues. A condensed version of some of the most fundamental human activities may be seen in Figure 2.

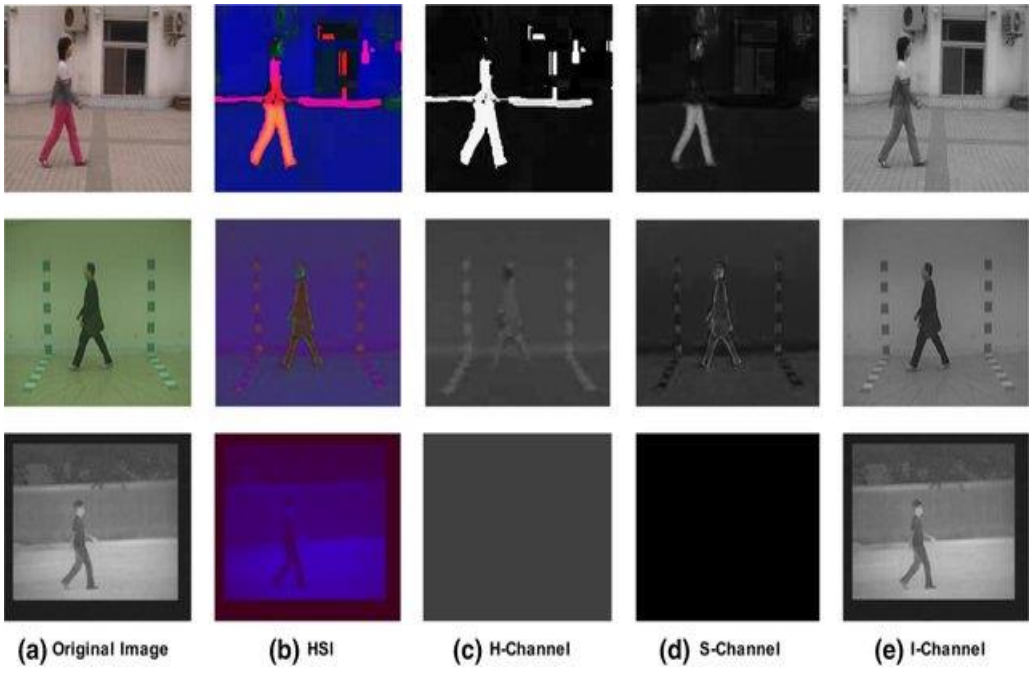


**Figure 2. Representative frames of the main human action classes for various datasets**.

3. Emotional State Recognition and Detection in Images and Videos

3.1.1. Before the voice recognition module can extract text information from the audio file during usage, the interactive object's speech must be captured using a microphone

and transformed into an audio file. The emotional state of the interactive object is generated by the model's emotion analysis using pre-processed text input. In this work [21], machine learning was used to construct a model for analyzing emotions based on acquired data. Once offline training with data sets is finished in a certain manner, the model is reserved. To produce predictions, the previously stored model is employed. Figure 2 shows how a machine learning algorithm may recognize text emoticons.

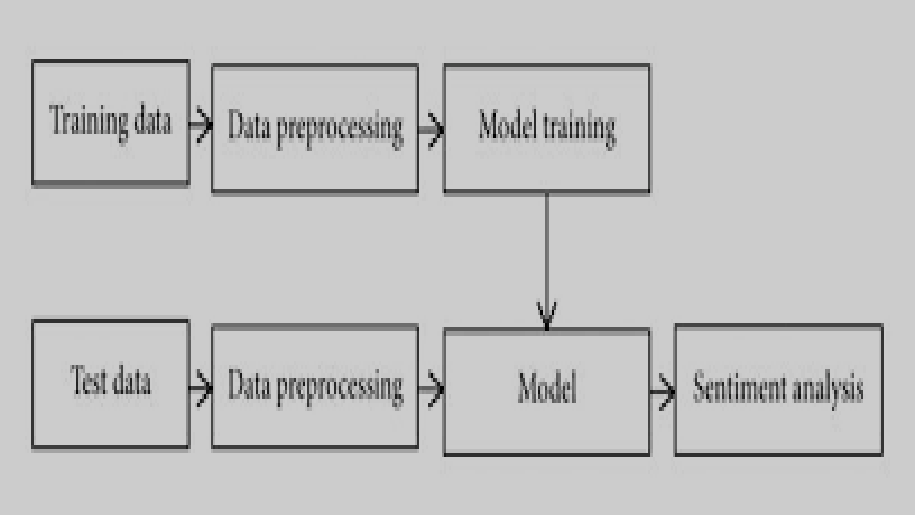


**Figure 3** Text emotion analysis process based on machine learning

Data Collection and Preprocessing 3.2

Data capture and pre-processing are included in the data. The following is a breakdown of the contents.

3.2.1 The Accumulation of Data

With the help of the "Microblog Cross-Language Emotion Recognition Dataset," we were able to develop an emotion analysis model for a humanoid robot. The corpus is separated into the positive and the negative categories. There are 12,153 negative corpus groupings and 12,178 positive corpus groupings included in the data set. Because of its conversational tone, the microblog corpus is an excellent choice for the training of an emotion identification system. Data preparation Following the down sampling process, which included removing 25 negative label items from the negative label corpus, the positive and negative samples were combined into a total of 12,153 items. Because the vast bulk of the corpus was taken from Weibo, it contains a lot of emoticons and punctuation that is used several times. The use of voice recognition eliminates the need for repeating punctuation [22]. Because of the preprocessing, the characteristics of repetitive punctuation and emoticons have been deleted. During the portion of this inquiry devoted to the segmentation of words, Jieba's toolkit was put to use. In text processing, punctuation and facial expressions are compared to one another. A table contrasting the use of punctuation with the interpretation of facial expressions in text processing There are several methods available for encoding a phrase in vector space; however, one of the models that is applied rather often is the word bag model.

## 4. An Inquiry Into the Qualities Involved

The chi-square test, information gain, the mutual information approach, and the text frequency-inverse document frequency measure are some of the typical feature selection processes used in the text vector space model. The absolute term frequency (TF) and inverse document frequency (IDF) are the two components of the TF-IDF combination that the word bag model employs in order to emphasize keywords in documents (IDF). The feature item's spelling may be determined based on its absolute word frequency (TF), which can be found in the training text. Absolute word frequency is a method that may be used to rapidly identify the important phrases included in a document. By using this method, one is able to calculate the inverted document frequency (IDF). For instance, ni denotes the frequency with which each feature item occurs in the training set in comparison to the total number of documents. The IDF emphasizes fewer terms in their list, but those words are better organized. When it comes time to do the actual computation, IDF will make sure that peculiar phrases are not left out of the database. Word2vec was first conceptualized by Google in the year 2013. In this demonstration of dense feature representation or distributed representation, the features of individual words serve as the units of representation. Two training models that are used for Word2vec are known as CBOW (continuous bag of words) and Skip-Gram. Hierarchical SoftMax and Negative Sampling are two enhanced strategies for Word2vec that strive to speed up computation and training. Hierarchical SoftMax was developed by Google while Negative Sampling was developed by Microsoft. The output of the projection layer for the Hierarchical SoftMax model is the mean word vector sum, which is also the output for the Skip-Gram model. This is the same as the previous sentence. The Hierarchical SoftMax technique utilizes Huffman trees rather than a SoftMax mapping of the projection layer to the output layer. This is done so that the probability of each individual word does not have to be calculated. Lack of Representation in the Sample Word2vec is a pretraining approach for neural networks that may improve the training starting point of the neural network and make it simpler to optimize [23]. In addition to this, a word vector may be used as a pretraining method. The dense feature can be calculated more quickly and does not experience dimension explosion as the sparse feature does, which enables more generalization. instead than using separate thermal coding. It is possible to compare features when a dense representation of those characteristics is used. In natural language processing applications such as Chinese word segmentation, sentiment analysis, and reading comprehension, this distributed form of the word vector is employed extensively.

5. The development of a model for the analysis of emotional states

Text may also be categorized with the assistance of the support vector machine (SVM). Finding the hyperplane in the feature space that has the greatest interval in it is the primary objective of a support vector machine (SVM). This method has the benefit of functioning well even in situations in which the number of dimensions is more than the number of samples, such as when dealing with a high-dimensional space. It is possible to use a diverse set of kernel functions with support vector machines if they are properly developed. When the number of samples is more than the number of features, the

performance of an SVM decreases. The most important thing that this research brings to the table is a fresh perspective. Only a basic description is given of the procedures required for each component. The purpose of this study is to examine an SVM-based sentiment analysis model. The performance of a conventional machine learning model is examined with a wide variety of inputs, and the neural network model is shown to be influenced by attention processes. The support vector machine with TF-IDF achieved the best classification results in the trial with a single model (F1 = 0.795, accuracy of 78.94%, AUC = 0.863).

## CONCLUSION

In this work, we presented a hierarchical taxonomy for classifying the various ways to identifying human activities after conducting a comprehensive analysis of the most recent and cutting-edge methods currently available. We investigated a wide array of methods, which we then classified into one of two primary classes—namely, the unimodal or multimodal category—depending on the source channel that was utilized to identify human actions. We investigated the various unimodal approaches and categorized them in-house. These techniques were developed for the purpose of evaluating gestures, atomic actions, and more complex activities, either directly or by dissecting such activities into a series of smaller acts. In addition to this, we discussed the use of multimodal methods for researching human social behaviors and interactions. We spoke about the several levels of feature modality representation and highlighted the merits and drawbacks of each type of the representation. In addition to this, a comprehensive analysis of the existing criteria for classifying human activities was presented, and we evaluated the difficulties associated with data collection in connection to the issue of comprehending human activities. In the last part of this discussion, we went through the characteristics of an effective human activity identification system. Due to the limitations of computers, the majority of the earlier studies conducted in this subject were unable to accurately characterize human activities in a way that was both easy and helpful. A comprehensive picture of human actions and the attendant data collecting and annotation continue to be challenging and contentious issues that have not been satisfactorily addressed. In particular, we may draw the conclusion that, despite the significant advancements that have been made in methods for human comprehension, a great lot of difficulty still exists in areas such as modeling human postures, dealing with occlusions, and annotating data.

## Reference:

1. L. Gabriella, G. Márta, K. Veronika et al., “Emotion attribution to a non-humanoid robot in different social situations,” PLoS One, vol. 9, no. 12, Article ID e114207, 2014.View at: Publisher Site | Google Scholar
2. A. Rozanska and M. Podpora, “Multimodal sentiment analysis applied to interaction between patients and a humanoid robot pepper,” IFAC-Papers Online, vol. 52, no. 27, pp. 411–414, 2019.View at: Publisher Site | Google Scholar

3. M. Viríkova and S. Peter, "Teach your robot how you want it to express emotions," Advances in Intelligent Systems and Computing, vol. 316, pp. 81–92, 2015.View at: Google Scholar
4. X. Ke, Y. Shang, and K. Lu, "Based on hyper works humanoid robot facial expression simulation," Manufacturing Automation, vol. 137, no. 1, pp. 118–121, 2015.View at: Google Scholar
5. F. Azni Jafar, N. Abdullah, N. Blar, M. N. Muhammad, and A. M. Kassim, "Analysis of human emotion state in collaboration with robot," Applied Mechanics and Materials, vol. 465-466, pp. 682–687, 2013.View at: Publisher Site | Google Scholar
6. Z. Shao, R. Chandramouli, K. P. Subbalakshmi, and C. T. Boyadjiev, "An analytical system for user emotion extraction, mental state modeling, and rating," Expert Systems with Applications, vol. 124, no. 7, pp. 82–96, 2019.View at: Publisher Site | Google Scholar
7. J. Hernandez-Vicen, S. Martinez, J. Garcia-Haro, and C. Balaguer, "Correction of visual perception based on neuro-fuzzy learning for the humanoid robot TEO," Sensors, vol. 18, no. 4, pp. 972-973, 2018.View at: Publisher Site | Google Scholar
8. A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," IEEE Transactions on Human-Machine Systems, vol. 44, no. 2, pp. 157–168, 2014.View at: Publisher Site | Google Scholar
9. J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, "Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism," IEEE Transactions on Autonomous Mental Development, vol. 6, no. 3, pp. 183–199, 2014.View at: Publisher Site | Google Scholar
10. L. Tang, Z. Li, X. Yuan, W. Li, and A. Liu, "Analysis of operation behavior of inspection robot in human-machine interaction," Modern Manufacturing Engineering, vol. 3, no. 3, pp. 7-8, 2021.View at: Google Scholar
11. Z. Li and H. Wang, "Design and implementation of mobile robot remote human-computer interaction software platform," Computer Measurement & Control, vol. 25, no. 4, pp. 5-6, 2017.View at: Google Scholar
12. H. Huang, N. Liu, M. Hu, Y. Tao, and L. Kou, "Robot cognitive and affective interaction model based on game," Journal of Electronics and Information Technology, vol. 43, no. p. 1.View at: Google Scholar
13. 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
14. 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.

15. Chen, C. Y., and Grauman, K. (2012). "Efficient activity detection with maxsubgraph 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).
16. "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.
17. 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.
18. 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.
19. Van den Heuvel, H, Huijnen, C, Caleb-Solly, P. Mobiserv: a service robot and intelligent home environment for the Provision of health, nutrition and safety services to older adults. Gerontechnology 2012; 11(2): 373
20. Demiris, G, Hensel, BK, Skubic, M. Senior residents perceived need of and preferences for smart home sensor technologies. Int J Technol Assess Health Care 2008;
21. Ram, R, Furfari, F, Girolami, M. UniversAAL: provisioning platform for AAL services. In: Van Berlo, A, Hallenborg, K, Corchado Rodríguez, JM. (eds) Ambient intelligence-software and applications. Berlin: Springer International
22. Rashidi, P, Cook, DJ. Keeping the resident in the loop: adapting the smart home to the user. IEEE T Syst Man Cy A 2009; 39(5): 949–959
23. Roy, N, Misra, A, Cook, D. Ambient and smartphone sensor assisted ADL recognition in multi-inhabitant smart environments. J Ambient Intell Human Comput 2016; 7(1): 1–19.