Video-based Activity Level Recognition for Assisted Living using Motion Features

Sandipan Pal
Dept. of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD
s.pal@sheffield.ac.uk

Charith Abhayaratne
Dept. of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD
c.abhayaratne@sheffield.ac.uk

ABSTRACT

Activities of daily living of the elderly is often monitored using passive sensor networks. With the reduction of camera prices, there is a growing interest of video-based approaches to provide a smart, safe and independent living environment for the elderly. In this paper, activity level in context of tracking the movement pattern of an individual as a metric to monitor the daily living of the elderly is explored. Activity levels can be an effective indicator that would denote the amount of busyness of an individual by modelling motion features over time. The novel framework uses two different variants of the motion features captured from two camera angles and classifies them into different activity levels using neural networks. A new dataset for assisted living research called the Sheffield Activities of Daily Living (SADL) dataset is used where each activity is simulated by 6 subjects and is captured under two different illumination conditions within a simulated assisted living environment. The experiments show that the overall detection rate using a single camera setup and a dual camera setup is above 80%.

CCS Concepts

•Social and professional topics → Assistive technologies; •Computing methodologies → Activity recognition and understanding; *Neural networks;*

Keywords

activity level recognition; activities of daily living; video-based monitoring; assisted living; motion analysis; neural networks

1. INTRODUCTION

Ageing is a global phenomenon with the elderly population of the world increasing at an alarming rate. In the United Kingdom alone the number of people between the age of 60 to 75 years would rise by 2 million, whereas the above 75 years category, the rise would be nearly 4 million [1] by 2035. The older and vulnerable people can be provided with a secure environment thereby improving their quality of life with the use of information and communication technology based products, services and systems. In order to facilitate healthy and independent living for the elderly population the need for smarter homes embedded with the latest technology for assisted living is imperative. Smarter homes and the effective use of technology would also reduce the overall costs of health and social care [5]. One of the earliest studies of behavioural monitoring of individuals done by Celler *et al.* [6] found out that the health status of an individual can be determined by monitoring a number of simple parameters (mobility, sleep patterns, utilisation of cooking, washing and toilet facilities) which are representative of the interaction of the individual with his environment. Hine *et al.* in [12] states that the activities of daily living (ADLs) of an individual are indicative of ones' well-being precursors as well as predictive of ones' well-being outcomes and that monitoring of ones daily living should be about visualizing activities at different levels of granularity.

Currently most of these parameters and interactions are captured using passive sensor networks to detect behavioural patterns or deviant activities [4]. Another approach to monitor these interactions is by using a body sensor network [2]. Body sensor networks are obtrusive and often elderly people might forget to wear them. In [5], Cardinaux *et al.* notes that to effectively track a person within a room, multiple passive sensors are required. Compared to complex and obtrusive sensor networks, video-based solutions offers more contextual information about the individual and environment. In recent years, video-based human activity analysis is gaining much recognition. In this paper, monitoring of individuals for an independent living is explored using the concept of activity levels. The key highlights of this paper are -

- a novel video-based framework for a smart assisted living environment
- exploring different motion features and classifying them into different activity levels using neural networks
- comparing the performance of the motion features with a single camera setup and an orthogonally positioned dual camera setup
- Sheffield Activities of Daily Living Dataset(SADL)[1] - a new dataset for video-based assisted living research.

[1]Sheffield Activities of Daily Living Dataset (SADL) is available to download in http://svc.group.shef.ac.uk/SADL.html

ICDSC '15, September 08 - 11, 2015, Seville, Spain
© 2015 ACM. ISBN 978-1-4503-3681-9/15/09…$15.00
DOI: http://dx.doi.org/10.1145/2789116.2789140

Activity Levels can be an effective indicator that would denote the amount of busyness of an individual. Experiments show that the overall classification performance of the neural networks with both the camera setups using the proposed features is well over 80%; which is the threshold accuracy needed in recognizing the activities of daily living (ADLs) to detect for long term conditions as noted by Storf *et al.* in [19].

2. BACKGROUND

Video-based human activity recognition and analysis has been a topic of research for a very long time. Inspite of many research efforts, activity recognition for indoor monitoring is still a challenging task and is dependent on a number of factors like camera angles, occlusions *etc.* It is also dependent on the type of features one uses for the classification and the type of dataset the algorithm is used on. The most common approach is to is to detect local image features like Histogram of Gradient (HOG) [9] and use the distribution of such features over time to locate spatio-temporal points like Spatial Temporal Interest Points (STIP) [15]. These points are then modelled to represent a spatio-temporal pattern for recognition. Computing local features and then modelling it over time is resource intensive. Moreover these approaches have seldom been tested in context of a home monitoring system or assisted living as this is an emerging domain of application and also due to the unavailability of benchmark datasets. Recently there has been a keen interest to use visual systems for home monitoring systems. In [16], Messing *et al.* have used the velocity histories of tracked key points for daily activity recognition. In [8], Cheng *et al.* have also used the local HOG features over time to model different actions. Along with the local features, Wang *et al.* in [21] have used depth information to model different actions into an ensemble of *actionlet* to label each activity. All these approaches have been evaluated on their own created datasets where the subjects in the video is facing the camera.

Although visual data can be considered as intrusive, in a study by Boise *et al.*, around 60% of the participants in their study have expressed a willingness for a video-based monitoring system [3]. Recognizing the activities of an individual would be as good as recording their activities which can have privacy concerns among the end users. Hence recognition of a more abstracted information in the form of activity levels would form an effective metric to denote the way an individual interacts with the environment thereby giving a measure of the amount of activity an individual has undertaken.

3. FRAMEWORK AND IMPLEMENTATION

The fundamental essence of an assisted living environment is providing a safe, independent and non-intrusive living condition. Excessive sedentary behaviours are not good for ones health. The inter-relation between physical activity and health is well supported by over 60 years of scientific inquiry, and the beneficial effects of moderate-to-vigorous physical activity have been more clearly defined in recent years [14]. For a healthy living, one should undertake a few activities which would involve either movement of your body parts or physically moving from one location to another.

3.1 Activity Levels

The relationship between mobility and health status is well recognized. Increased mobility improves ones psychological well-being and quality of life [11]. One of the common behavioural patterns which can be monitored using a video-based system [5] in an ambient assisted living environment is mobility pattern. Activity levels is a representation of the amount of movement undertaken by an individual in carrying out the activities of daily living. In this work, 9 daily activities have been identified that an elderly person is expected to undertake in day to day life. These activities are selected such that it involves varying movement patterns. 3 classes of activity levels are proposed and are defined as:

- High Activity Level - denotes physically moving from one location to another, *e.g.* walking
- Low Activity Level - denotes minor or no change in ones location but involved in some sort of an activity which involves movement, *e.g.* taking things out of a cupboard or washing plates *etc.*
- No Activity Level - denotes extremely low or no movement at all, *e.g.* watching television or sitting idle *etc.*

Table 1: Activity Table

List of Activities	Location	Activity Level
Walking around	Kitchen, Living	HIGH
Using Fridge, Pour water, Using Oven (Plate In and out), Using cupboard	Kitchen	LOW
Sitting idle, Reading book, Watching Television	Living	NO

The list of activities and their corresponding Activity Level is listed out in Table 1. The above definitions form the basis for the generation of the ground-truth of the dataset.

3.2 Datasets

Within assisted living research one of the most extensive and real life dataset is the Senior Home Monitoring dataset [8] in which 6 elderly individuals are video-recorded over a period of 4 months within their dwelling performing 9 different activities sometimes assisted by a carer. The motivation for our research is to promote independent living and focuses on the monitoring of single occupancy dwellings where there is only one person living independently. The University of Rochester Activities of Daily Living Dataset [16] and the Kitchen Activity Dataset [18] are a few datasets where different activities are carried under a simulated home-environment. The MSRDailyActivity3D dataset [21] created in a lab environment has depth information along with the RGB video and has all the activities carried out in the the sitting as well as the standing posture. However all these datasets are restricted to a single location and a single fixed camera angle. To get a more distributed capture of visual information about the different activities, the new SADL dataset was created for this work. This dataset is a first of its kind where all the activities are simulated in two different locations within a simulated assisted living environment and captured by two orthogonally positioned cameras. Also unlike the other datasets, the activities are not always captured with the subject facing the camera.

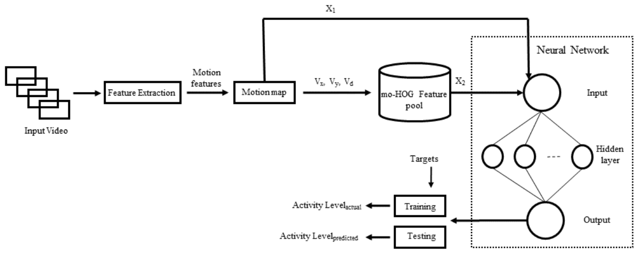


Figure 1: System Architecture

In the SADL dataset 9 basic activities of daily living are simulated within Home Lab. The Home Lab is a simulated indoor environment developed by CATCH[2]. The Home Lab mimics the home or care environment and acts as a bridge between the laboratory-based research environment and the homes of people in the community. Within the Home Lab, two locations, namely the kitchen and living room a shown in Figure 5 to simulate the 9 different basic activities of daily living. The activities are performed by 6 subjects and are captured from two orthogonally positioned different camera angles. As shown in Table 1, the activities are grouped into 3 classes of Activity Levels based on the amount of movement undertaken while carrying out the activities. In the dataset, there are 36 high activity level videos, 60 low activity level videos and 36 no activity level videos for each camera angle making a total of 264 videos. The activities are simulated under two illumination conditions, one being with the indoor lighting switched on and the other being with indoor lighting switched off. Each video is captured at a frame-rate of 50 frames per second (fps) and is in the .avi format. The resolution of the videos are 1920x1080 pixels. The different activities in the dataset is shown in Figure 5.

The dataset is called Sheffield Activities of Daily Living Dataset (SADL) and will be made freely available online for the research community.

3.3 Implementation

The framework as shown in Figure 1 has two main components to it. First features are extracted to compute a feature vector and subsequently the vector is fed into a neural network for the classification task. One of the assumptions, while conducting the experiments were that the activity levels of a single person would be recognized. There are two sets of experiments conducted, one with a single camera setup and the other with the dual camera setup. The first task was to extract the motion features from the video. This was done using the optical flow method given by the equation:

$$u = v + \delta, \tag{1}$$

where u is a pixel pixel location in the frame F_i, v is the location of the same pixel in frame F_j, i and j denotes the frame numbers and δ the displacement. For an elderly monitoring environment, one can assume that movement undertaken within 1 second is minimal. Hence to reduce computation, instead of every successive pair of frames, the motion vectors are computed between F_i and F_j where the difference between i and j is given by:

$$j - i = \Phi, \tag{2}$$

and Φ denotes the frames per second (fps). As mentioned before, two sets of experiments were carried out. For each of the two sets of experiments, the two input feature vectors, X_1 and X_2 were computed. It must be noted that the dimensions for X_1 and X_2 differs in the two sets of experiments.

3.3.1 Motion Map

The motion map is a representation of of the motion patterns within a video over N frames. For monitoring daily activities, motion vectors of magnitude less than 1 offer negligible information of the motion pattern. Hence the motion map is refined by a threshold of $\delta < 1$ to filter out the sensor noise and the motion of negligible magnitude. The refined motion features are then normalized by a factor α where α is given by:

$$\alpha = \frac{N}{\Phi}. \tag{3}$$

The normalized motion features are then used to create the feature vector X_1, the tuples of which are defined as follows -

- Normalized sum total of all the x-components of the motion vectors
- Normalized sum total of all the y-components of the motion vectors
- Normalized sum total of the number of motion vectors.

The aggregate of the x-components, the y-components and the density of the motion vectors represented by V_x, V_y and V_d respectively are forwarded to the the mo-HOG feature pool.

3.3.2 mo-HOG feature pool

The HOG features have been extensively used for object detection which is like detecting similar patterns within an image which defines an object. Activity levels are a representation of the motion pattern over time. To detect similar motion patterns invariant to the subject, mo-HOG features were designed. In the mo-HOG feature pool, the HOG features are computed on the motion magnitude. Unlike, the HOOF features as proposed in [7], the histogram is not computed on the euclidean distance of the x and y components of the motion magnitude. Instead, the features are computed separately on the two different components and all the features were considered instead of using the 'winner takes all' strategy as mentioned in [10]. The mo-HOG feature vector was computed with the win size of 32x32, which gave 5832 features for each component. Since the histograms were computed on the motion features, there are regions in the motion map which did not represent any motion and so the features computed on those regions can be ignored. To quantize the dimensions of the vector, principal component analysis using the single value decomposition algorithm was carried out making a resultant vector of 131 dimensions.

Feature Vector X_2 is an extension of vector X_1 fused with the mo-HOG features. The tuples of the the vector X_2 is given by:

- Feature vector X_1

[2]CATCH is the Centre for Assistive Technology and Connected Healthcare within The University of Sheffield. Website : http://www.catch.org.uk/

Table 2: Number of videos in training set and testing set for each Neural Network

	Training Set	Testing Set
HIGH	24	12
LOW	40	20
NO	24	12

- principal components mo-HOG features of the x-components of the motion vectors
- principal components mo-HOG features of the y-components of the motion vectors.

The extracted feature vectors serve as the input to the neural network for classification.

3.3.3 Neural Network

Artificial neural networks are statistical learning algorithms that are inspired by the understanding of how the human brain learns. Neural Networks find their application in practical applications such as speech recognition [13], object recognition, hand written digit recognition, temperature prediction and so on. Neural networks also show promising predictions even for safety critical applications like medical research [20]. Though the training of a neural network is computationally expensive, it requires comparatively less statistical data to predict the interaction between the predictor variables. More recently neural networks have also found applications in the activity monitoring domain [22] [17].

In this work a different two layer feed-forward back propagation neural network is trained for each of the experiments. For the first set of experiments the feature vector X_1 has 3 dimensions while vector X_2 has 265 dimensions. For the second set of experiments the feature vector X_1 has 6 dimensions while vector X_2 has 550 dimensions. Owing to three activity level classes, the network has a 3 neuron output layer. There are 10 neurons in the hidden layer and a log-sigmoid transfer function is used. Each of the network is trained using the Levenberg-Marquardt algorithm. Each of the 132 videos is split into a mutually exclusive training set and testing set. Out of the 6 subjects, videos of 4 subjects were chosen for the training set while videos of the remaining 2 subjects were used as the testing set. This was done so that, for each activity level, the number of videos in the testing set for each activity level is exactly 50% of the number of videos in the training set as shown in Table 2. The training is continued iteratively till a minimum gradient is reached. The training performance for the two sets of experiments are shown in Table 3 and Table 4. The classification results of each of the neural network is discussed in following section.

4. CLASSIFICATION RESULTS

Different neural networks were trained and used for each of the experiments. The training performance for each of the neural network is discussed in subsection 3.3.3. In this section, only the testing results of the neural network is discussed. In the initial experiments (subsection 4.1), features from a single camera was considered individually. The second set of experiments were considered, to investigate further into the initial results (subsection 4.2).

Table 3: Training performance of Neural Networks for the single camera setup

	Feature Vector X_1		Feature Vector X_2	
	Camera 1	Camera 2	Camera 1	Camera 2
Min. gradient	0.04	0.06	9.9e-07	2.0e-05
No. of iterations	25	17	24	19

Table 4: Training performance of Neural Networks for the dual camera setup

	Feature Vector X_1	Feature Vector X_2
Minimum gradient	0.08	9.2e-05
Number of iterations	30	16

Table 5: Confusion Matrix for Camera 1 and Camera 2 using feature vector X_1

		HIGH	LOW	NO
Camera 1	HIGH	**100%**	20%	0
	LOW	0	**80%**	0
	NO	0	0	**100%**
Camera 2	HIGH	**58%**	0	0
	LOW	42%	**100%**	0
	NO	0	0	**100%**

Table 6: Confusion Matrix for Camera 1 and Camera 2 using feature vector X_2

		HIGH	LOW	NO
Camera 1	HIGH	**75%**	35%	0
	LOW	25	**65%**	0
	NO	0	0	**100%**
Camera 2	HIGH	**75%**	5%	0
	LOW	25%	**95%**	0
	NO	0	0	**100%**

4.1 Single Camera Setup

In the single camera setup, features from only a single camera was considered at a time. Table 5 shows that using simple motion features like magnitude and density of the motion vectors gives promising result to determine the amount of movement undertaken by an individual for a particular camera angle. However if the angle of the camera is changed, the performance drops considerably, specially of the High Activity Level class. This is mainly due to the difference in the motion magnitude when captured from different perspectives. To solve this problem, the experiments were repeated with a feature vector of a higher dimension; the results of which are shown in Table 6.

4.2 Synchronized Dual Camera Setup

In the initial set of experiments, it was worth noting that the overall positive detections for feature vector X_1 from

Table 7: Confusion Matrix for Feature Vector X_3 and Feature Vector X_4 with a synchronized dual camera setup

		HIGH	LOW	NO
Feature Vector X_1	HIGH	**83%**	25%	0
	LOW	17%	**75%**	0
	NO	0	0	**100%**
Feature Vector X_2	HIGH	**67%**	0	0
	LOW	33%	**100%**	8%
	NO	0	0	**92%**

Camera 1 is the same as the feature vector X_2 from camera angle 2. Though X_2 appeared to be the more consistent feature vector irrespective of the camera position, a further set of experiments were needed to establish, that the better performance of X_2 was not only because of the high dimensionality of the input vector. Hence, features from both the cameras were fused in this set of experiments; the results of which are shown in Table 7. It must be noted that both the camera were synchronized, so essentially it is the same activity captured from two different camera angles between a fixed time interval.

4.3 Analysis

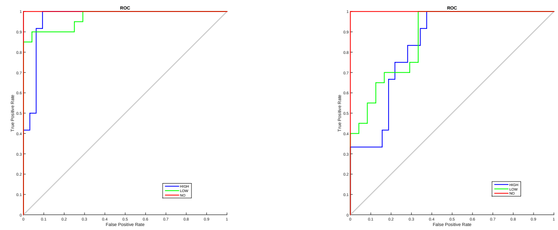
In the first set set of experiments, for feature vector X_1 the overall positive classification drops from 90.9% for Camera 1 to about 77.3% for Camera 2. On the other hand, the overall positive classifications for feature vector X_2 improves from 88.6% for Camera 1 to 90.9%. The difference in the performance between the two camera angles for X_2 is much smaller than X_1 which clearly establishes that a fusion of the motion magnitude and the mo-HOG features provides a more invariant feature for motion analysis. The Receiver Operation Curves (ROC) for each of the experiments is shown in Figure 2, Figure 3 and Figure 4.

In the second set of experiments, the fused features of both the cameras gave a higher dimensionality to the input vectors, but that does not affect the performance of the system much. The performance of the fused features is still better than only the motion magnitude. However even with a higher dimension input vector, the performance of X_2 for the dual setup is the same as that of the Camera Angle 1.

This shows that only a higher dimension input vector is not enough for effective classification and that the positioning of the camera plays a vital role for motion analysis. It must also be noted that the subjects in the testing videos were different from the subjects in the training videos. Different people have different body structures, different movement patterns and different ways to interact with their immediate environment.

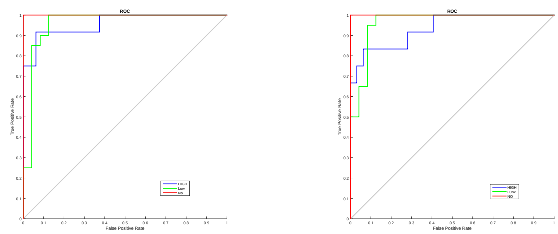
5. CONCLUSIONS

In this work, a novel framework to recognize activity levels within a smart home environment is proposed. Activity levels are a metric to denote the amount of activity that has been undertaken in ones daily living. Different classes of activity levels have been defined based on the amount of movement one needs to make to carry out different activities by an individual. Two different feature vectors with non-intrusive motion features are used for a performance comparison under a single camera as well as the orthogonally positioned dual camera setup. The performance of



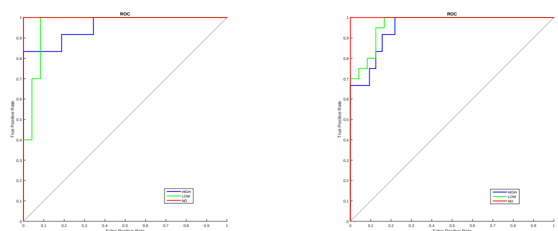
(a) (b)

Figure 2: ROC curve for feature vector X_1 from (a) Camera 1 and (b) Camera 2



(a) (b)

Figure 3: ROC curve for feature vector X_2 from (a) Camera 1 and (b) Camera 2



(a) (b)

Figure 4: ROC curve for a dual camera setup with feature vector (a) X_1 and (b) X_2

the neural networks show that motion features can be effectively used to measure activity levels of individuals. The new SADL dataset would provide a benchmark dataset for further video-based assisted living research.

6. REFERENCES

[1] *National Population Projections, 2010-Based Statistical Bulletin*. Office for National Statistics, October 2011.

[2] L. Atallah, B. Lo, R. Ali, R. King, and G.-Z. Yang. Real-time activity classification using ambient and wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):1031–1039, November 2009.

[3] L. Boise, K. Wild, N. Mattek, M. Ruhl, H. H. Dodge, and J. Kaye. Willingness of older adults to share data and privacy concerns after exposure to unobtrusive in-home monitoring. *Gerontechnology : International*

Figure 5: Different activities of the SADL dataset under two different illumination conditions in two different locations with two different camera angles

journal on the fundamental aspects of technology to serve the ageing society, 11(3):428–435, 2013.

[4] S. Brownsell, S. Blackburn, and M. S. Hawley. An evaluation of second and third generation telecare services in older people's housing. *Journal of Telemedicine and Telecare*, 14(1):8–12, 2008.

[5] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. S. Hawley. Video based technology for ambient assisted living: A review of the literature. *Journal of Ambient Intelligence and Smart Environments*, 3(3):253–269, August 2011.

[6] B. Celler, W. Earnshaw, E. Ilsar, L. Betbeder-Matibet, M. Harris, R. Clark, T. Hesketh, and N. Lovell. Remote monitoring of health status of the elderly at home. A multidisciplinary project on aging at the University of New South Wales. *International journal of bio-medical computing*, 40(2):147–155, 1995.

[7] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 1932–1939, June 2009.

[8] H. Cheng, Z. Liu, Y. Zhao, and G. Ye. Real world activity summary for senior home monitoring. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–4, July 2011.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR)*, volume 1, pages 886–893, June 2005.

[10] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 428–441, 2006.

[11] H. Gokalp and M. Clarke. Monitoring activities of daily living of the elderly and the potential for its use in telecare and telehealth: a review. *TELEMEDICINE and e-HEALTH*, 19(12):910–923, 2013.

[12] N. Hine, A. Judson, S. Ashraf, J. Arnott, A. Sixsmith, S. Brown, and P. Garner. Modelling the behaviour of elderly people as a means of monitoring well being. In *Proceedings of the 10th International Conference on User Modeling (UM)*, pages 241–250, Edinburgh, Scotland, UK, 2005.

[13] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[14] P. T. Katzmarzyk. Physical activity, sedentary behavior, and health: Paradigm paralysis or paradigm shift? *Diabetes*, 59(11):2717–2725, 2010.

[15] I. Laptev and T. Lindeberg. Space-time interest points. In *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV)*, pages 432–439 vol.1, October 2003.

[16] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV)*, Washington, DC, USA, 2009.

[17] S. Oniga and J. Suto. Human activity recognition using neural networks. In *Proceedings of the 15th International Carpathian Control Conference (ICCC)*, pages 403–406, May 2014.

[18] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1194–1201, June 2012.

[19] H. Storf, T. Kleinberger, M. Becker, M. Schmitt, F. Bomarius, and S. Prueckner. *An event-driven approach to activity recognition in ambient assisted living*. Springer, 2009.

[20] J. V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225 – 1231, 1996.

[21] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, June 2012.

[22] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3D human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the ACM International Conference on Multimedia*, pages 97–106, New York, USA, 2014.