Action Recognition in Real Homes using Low Resolution Depth Video Data

Flávia Dias Casagrande*, Oda Olsen Nedrejord*, Wonho Lee, and Evi Zouganeli
Dept. of Mechanical, Electronics and Chemical Engineering, OsloMet – Oslo Metropolitan University
Oslo, Norway
Email: *flacas@oslomet.no, *oda.nedrejord@gmail.com, defender.of.kimchi@gmail.com, evizou@oslomet.no

***Abstract*—We report work in progress from interdisciplinary research on Assisted Living Technology in smart homes for older adults with mild cognitive impairments or dementia. We present our field trial, the set-up for collecting and storing data from real homes, and preliminary results on action recognition using low resolution depth video cameras. The data have been collected from seven apartments with one resident each over a period of two weeks. We propose a pre-processing of the depth videos by applying an Infinite Response Filter (IIR) for extracting the movements in the frames prior to classification. In this work we classify four actions: TV interaction (turn it on/ off and switch over), standing up, sitting down, and no movement. Our first results indicate that using the IIR filter for movement information extraction improves accuracy and can be an efficient method for recognizing actions. Our current implementation uses a convolutional long short-term memory (ConvLSTM) neural network, and achieved an average peak accuracy of 86%.**

***Index Terms*—depth video, neural networks, smart homes, low resolution, action recognition**

I. Introduction

The Assisted Living project is an interdisciplinary project with expertise in the fields of smart-home technology, machine learning, nursing and occupational therapy, and ethics. The aim is to develop assisted living technology (ALT) to support older adults with mild cognitive impairment or dementia (MCI/D) live a safe and independent life at home [1]. MCI and dementia involve a cognitive decline that can affect attention, concentration, memory, comprehension, reasoning, and problem solving [2]. A number of research studies have investigated functions in smart-home environments to support older adults in general, and those with MCI/D in particular, in their everyday life. These include assisting functions such as prompting with reminders or encouragement, diagnosis tools, as well as alarm creation, prediction, anticipation, and prevention of hazardous situations. The majority of these functions requires reliable activity/ action recognition and prediction algorithms to work properly. This field is at a quite early stage at the moment. With the exception of fall detection, there are currently no commercial systems with such functionality nor are there any complete prototypes available at research and development level. In addition, most of the research that is published in the literature has been carried out in the lab based on scripted actions/ activities.

The aim of our work is to use activity prediction to realize support functions for older adults with MCI/D. In this paper we present work in progress on computer vision based action recognition using data from real homes, seven apartments, each with one older adult resident over 65 years old – the majority over 80 years old. We use a low resolution depth video camera that is in fact a commercial fall detection system called RoomMate [3]. We report preliminary results on action recognition based on video frames that contain movement information. Movement classification can allow inference of the future pose of the person (e.g. after a sitting down movement has been classified, the person is sitting), activities (e.g. drinking water), and intention (e.g. person would like to turn on the TV) and can be a stepping stone towards activity prediction. After a brief summary of related work, the paper gives an account of the field trial, the set-up in each apartment, and the way the data have been collected and stored. Subsequently we present the method we have used for processing the videos for the classification of actions in the homes where we propose a pre-processing of the depth videos by applying an Infinite Response Filter (IIR) for extracting the movements in the frames prior to classification. The use of the IIR filter ought to facilitate movement classification as different movements should involve different frequency components. We compare the results with and without the IIR filter to evaluate its efficacy. We subsequently summarize the results and conclude the paper by discussing how to improve our results as well as other future work.

II. Related Work

There is strong evidence that technology can support aging at home [4] and a large number of studies have implemented assistive technology to support older adults live a safe and independent life at home [5]–[8].

Human activity recognition (HAR) has been well studied in the past years [9], [10]. Motion History Images (MHI) is a method widely used in depth images [11], based on encoding a sequence of moving silhouettes. Another representation is the Histogram of Oriented 4D Normals (HON4D) [12], which captures the distribution of the surface normal orientation in the 4D space. Depth Motion Maps (DMM)] generate three 2D maps from 3D frames (front, side, and top views) [13]. For each map, a motion energy is calculated from the difference of two consecutive mappings. The same authors

Financed by the Norwegian Research Council under the SAMANSVAR programme (247620/O70).

have also developed a framework called Super Normal Vector (SNV) [14]. Several studies extract skeleton data from the depth images, and this has also proved to be useful [15], [16].

A number of algorithms have been used for HAR. Dynamic time warping has led to high accuracy of 96% on a dataset with four gestures and RGB data [17]. Generative models such as Hidden Markov Models (HMM) achieved a maximum accuracy of 97% with the MSR Action3D dataset with skeleton data histograms fed to a HMM [18]. Other models include support vector machines (SVM) and neural networks. Yang et al. [13] reached 97% accuracy on the same dataset by using DMM-HOG features and SVM. Among neural networks, convolutional neural networks have achieved best results for HAR from depth data. Wang et al. [19] achieved 100% accuracy on the same action dataset with a deep convolutional network by using weighted hierarchical DMMs of the video sequences. The same method was applied to other public datasets, achieving remarkable results. These were improved in a later work, where the authors used an improved pseudo-coloring on the obtained weighted hierarchical DMMs and attained better results in other datasets (2-9% higher) [20].

III. Field Trial

Our field trial involves seven independent one-bedroom apartments within a community care facility for people over 65 years old. Each apartment comprises a bedroom, a living room, open kitchen area, a bathroom, and an entrance hall (Figure 1).

The trial and the deployed sensor system have been decided in close collaboration with the residents after a series of Dialogue Cafés [1]. We have two RoomMate depth cameras (Figure 2) in each apartment. One of them monitors the living room and kitchen area, while the other monitors the bedroom area, as shown in Figure 1. The RoomMate is an infra-red (IR)-based depth sensor and measures the distance of surfaces to the camera by time-of-flight (TOF) technology with pulses at 15MHz. The resolution is 160x120 pixels, with a rate of 25 frames per second. This is rather low resolution – a fact that is advantageous with respect to privacy, but makes data processing quite challenging.

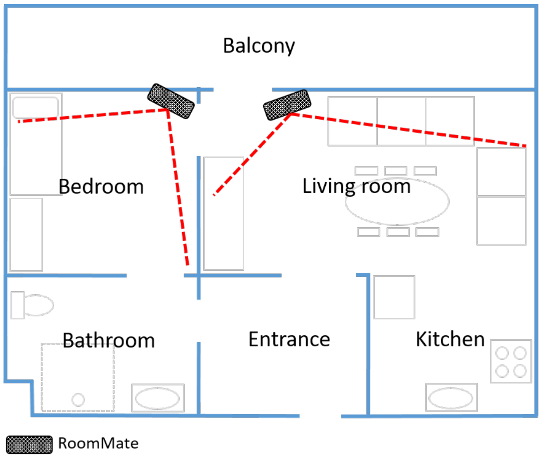

Fig. 1. Schematic of the apartment with the depth sensors.

Fig. 2. RoomMate depth sensor deployed in our field trial [3].

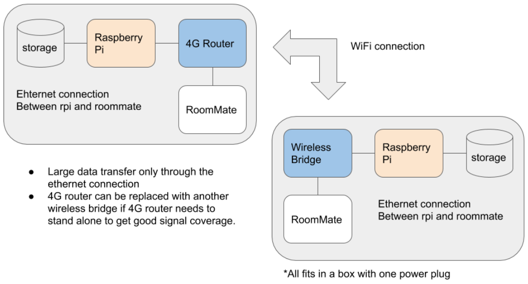
storage
Raspberry Pi
4G Router
Ehternet connection Between rpi and roommate
RoomMate
WiFi connection
Wireless Bridge
Raspberry Pi
storage
RoomMate
Ethernet connection Between rpi and roommate
• Large data transfer only through the ethernet connection
• 4G router can be replaced with another wireless bridge if 4G router needs to stand alone to get good signal coverage.
*All fits in a box with one power plug

Fig. 3. RoomMate set-up in each of the apartments.

A. Data Collection and Storage

Each RoomMate is connected through TCP/IP to a Raspberry PI and the data are stored on a local hard disk. The RoomMate needs to be able to connect to the Internet and send an alarm in case of a fall. On the other hand, in order to safeguard privacy, the video data we collect need to be saved only locally with no connection to the internet. Hence, we opted for the configuration shown in Figure 3 for data collection from each apartment. 4G ensures good connectivity to the internet at all times. We record only occasionally and for a maximum of two weeks at a time. The residents are informed prior to each recording period and no recording takes place otherwise. The local hard disks are collected at the end of each recording period and physically transported for storage in a secure server (TSD) [21].

IV. Method – Data Processing

A. Pre-processing

Median filtering is applied to the raw depth video data to remove noise. The process consists of removing very low and very high pixel values in the image and replacing them with the median value of the nearest neighbors. A 5x5 filter was applied to each frame, as a compromise between image sharpness (quality) and its high frequency background noise. Figure 5 shows the result of a median filtered image, applied to the raw image in Figure 4.

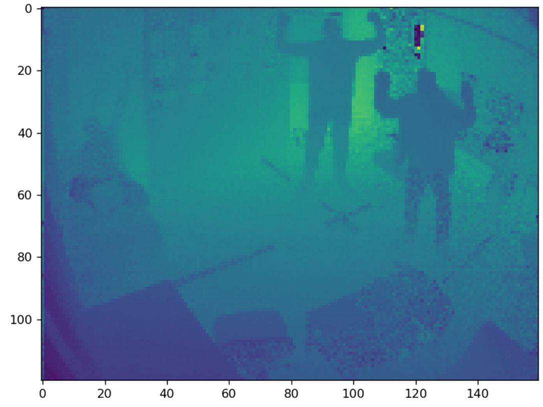

Fig. 4. Raw image from depth sensor.

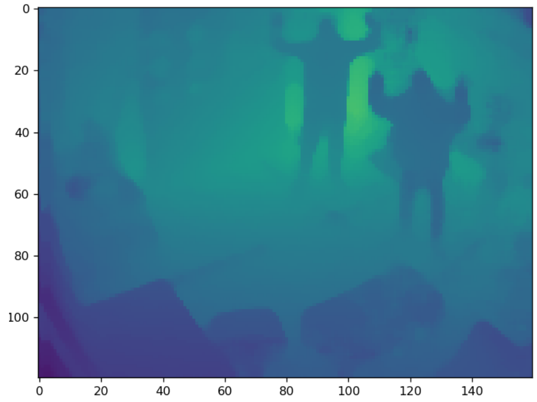

Fig. 5. Median filtering applied to raw image in Figure 4.

B. Infinite Impulse Response Filter

After this first step, we apply an IIR filter. This is a recursive filter, which means that its output is dependent on past output values. The filter is configured as a first-order high pass in this work, which leads to capturing any movement. The filter parameters were tuned empirically to allow the capture of both slow and fast movements within the home. The transfer function of the filter and its coefficients are shown in Equation 1. Figure 6 shows the frequency response of the IIR filter that attenuates frequencies below 0.09. Figure 7 shows the sum of all pixel values in each frame which can be thought of as a measure of the energy. The energy is computed with (blue line) and without (orange line) filtering. The filtered graph clearly indicates when there is rapid changing in the image, in this case around frames 50, 200 and 350. These cannot be seen from the unfiltered image. By manually inspecting the video, it has been confirmed that there are movements in the frame numbers that correspond to the peaks in the graph. Hence, based on these findings, the motion graph verifies that the filter has been successful in extracting motion from the images. Figure 8 shows the results of a frame after the application of both the median filter and the IIR filter.

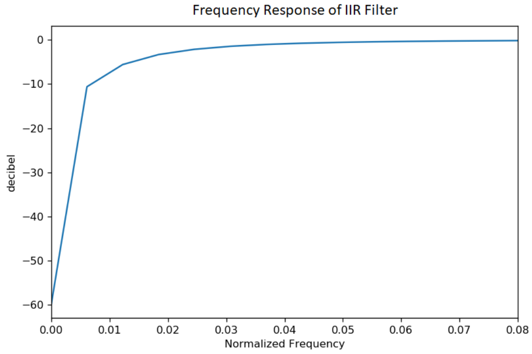

Fig. 6. Frequency response of implemented IIR filter.

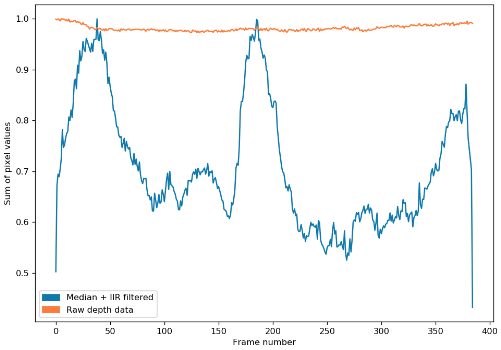

Fig. 7. Movement extraction in a sample video.

$$A_0 = 1 \quad B_0 = 1 \quad A_1 = -(1-\alpha) \quad B_1 = -1 \quad \alpha = 0.02$$

$$H(z) = \frac{B_0 + (B_1 \times z)}{A_0 + (A_1 \times z)} \tag{1}$$

C. Frame Length Normalization

Finally, a last processing step is performed in order to normalize the length n_i of frames of the video samples to a fixed length N, as this is a prerequisite for the convLSTM model. We use a method that we refer to as median insert. Frames are deleted if the sequence is shorter than N, or inserted if the sequence is longer than N. In both cases this takes place in equally spaced positions in the sequence (in accordance with the number of frames that need to be deleted/ inserted), until $n_i = N$. The value of each pixel in the inserted frames is equal to the mean between the preceding and succeeding frames.

D. Classification

We use a convolutional long short-term memory network (convLSTM) for the classification. Convolutional neural networks (CNNs) have been widely used to process multiple arrays of data, including color or depth images [22], [23].

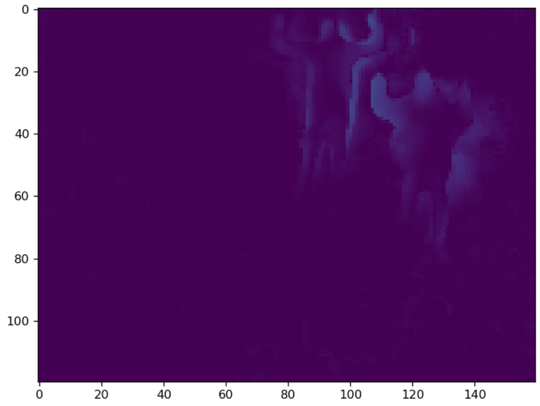

Fig. 8. IIR filtering applied to image in Figure 8.

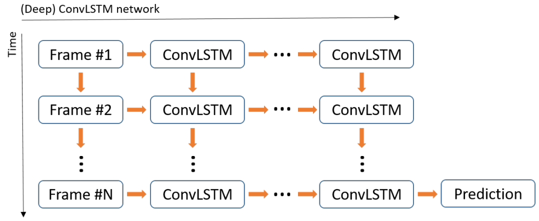

Fig. 9. ConvLSTM model.

Recurrent neural networks (RNNs) have been extensively applied to sequence prediction tasks because of the property of keeping an internal memory [24], [25]. The long short-term memory (LSTM) network is an RNN architecture designed to be better at storing and accessing information than the standard RNN [26]. By combining both types of neural networks, the model is able to learn both spatial and temporal features from a sequence of frames [27]. A schematic of the convLSTM network process is shown in Figure 9.

In this work, the pre-processed and IIR-filtered frame sequences were labelled into four movement categories: TV-related movements (turn it on/ off and switch over channels), standing up, sitting down and no movement. After the pre-processing, the sequences were fed to a convLSTM.

E. Implementation

The first two steps were implemented by using the median and IIR filters from the SciPy library. The convLSTM was implemented using the Keras library, based on TensorFlow. Both libraries are Python-based open source software. A number of parameters were tuned in the network. The trained model comprises one convLSTM layer with three 3x3 filters and hyperbolic tangent activation, followed by a dense layer with softmax activation. The batch size was 16 and learning rate 0.01. Optimization function Adam and loss function categorical cross-entropy were optimal. We set the dropout ratio to 0.5 in order to avoid overfitting, as well as early stopping (stop the network training when there was no improvement in the validation loss for 5 consecutive epochs).

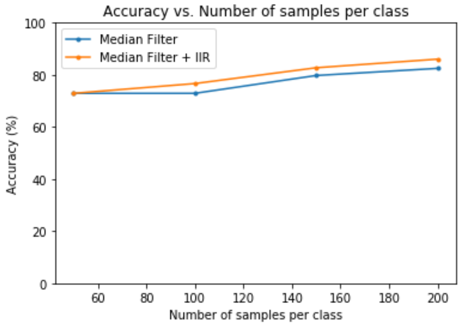

Fig. 10. Accuracy vs dataset sizes, for median filter, and both median and IIR filter.

V. Results and Discussion

A total of 800 video sequences (200 of each category) were extracted from recordings acquired from real homes, from seven different residents. Each video sequence is length normalized to a size of 100 frames. We split our dataset into training (80%) and testing (20%) sets. They are both balanced for all classes (i.e. equal percentage of samples per class in each set).

We first analize the test accuracy attained for different sizes of datasets for two cases: only median filter, and both median and IIR filter – Figure 10. The obtained results correspond to an average of the three best accuracies achieved by different trained models – shuffling the training and testing data. The use of the IIR filter resulted in a best average peak accuracy of 86.04%, whereas by comparison without the IIR filter the best average peak accuracy achieved was 82.50%. Using the IIR filter improves the accuracy by approximately 4%, for all data sizes. We can notice that the accuracy improves slowly as more samples are added, in both cases. The model has not yet reached stability as the accuracy keeps increasing with data size, hence better accuracy should be possible to achieve with additional samples.

Figures 11 and 12 present the train/ test accuracy and loss of the model that achieved the best test accuracy (91.25%). The test curves show a quite unstable behavior, indicating that even though a dropout layer was used together with early stopping, the model is still overfitting. This is yet another indication that that additional data is required for the training.

The confusion matrix derived from the best model when training with 800 samples is shown in Figure 13. The best predicted class was the TV interaction, with an accuracy of 97.5%. This was mis-classified once as sitting down. The classes sitting down, standing up, and no movement had comparable accuracies, and they cause most of the confusion. In addition to the innate accuracy of the method, the confusion may be in part due to residual noise that was not completely filtered out in all the video sequences with the parameters we used in the median filter, as well as any erroneous labelling.

We compare the confusion matrix when using the IIR filter in Figure 13 with the confusion matrix for the same data but

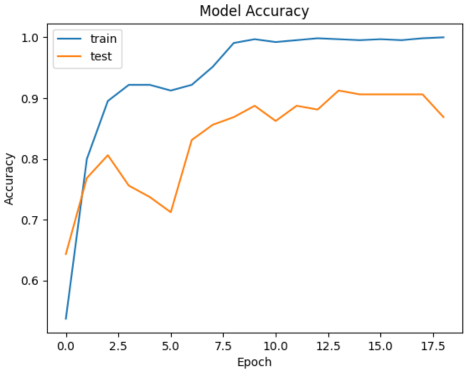

Fig. 11. Training and testing accuracy per epoch.

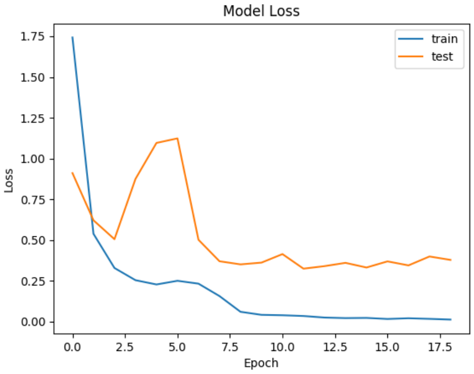

Fig. 12. Training and testing loss per epoch.

without using the IIR filter in Figure 14. The latter leads to lower accuracy for all classes. The most pronounced effect is however that the class no movement has 20% lower accuracy without the IIR. The reason the IIR has this large effect on this particular class is not completely clear. This is however not too surprising considering that the IIR filter has been configured as a high-pass filter, hence it should be effective in distinguishing no movement from movement. Also, since the other three classes involve a movement while this one does not, it may be that the IIR filter removed any low-frequency noise in this class that was mistaken as movement.

VI. Conclusions and Future Work

Activity/ action recognition algorithms in smart home environments using depth cameras are useful for a number of functionalities. Especially low resolution depth cameras, allow remote monitoring without being too intrusive and compromising the privacy of the home. Processing the data is however much more challenging than normal cameras or higher resolution depth videos. Most of the work reported in the literature has been carried out using data collected in lab environments and testbeds, with scripted activities, actors, and using higher resolution videos, e.g. using Kinect.

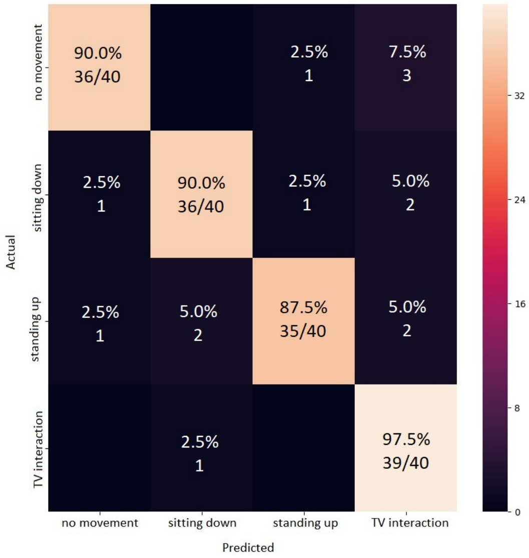

Fig. 13. Confusion matrix of model for 800 samples with median and IIR filter.

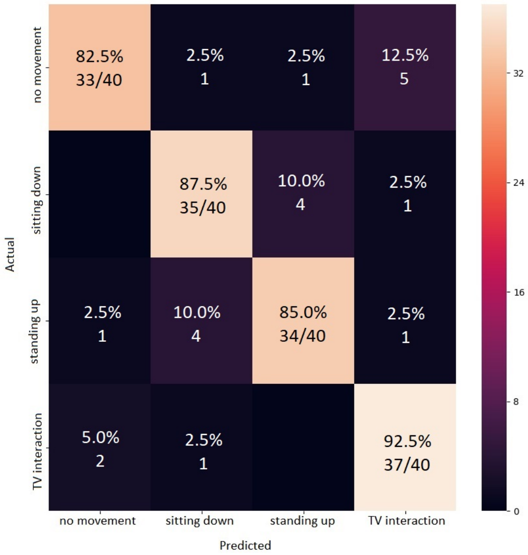

Fig. 14. Confusion matrix of model for 800 samples with only median filter.

In this paper, we present work in progress, our field-trial and set-up for collecting and storing the data, and very preliminary results on action recognition from real homes. The field trial comprises seven apartments with two low resolution depth video cameras each. Data have been collected from these over two weeks and 800 video samples were extracted containing four classes: no movement, standing up, sitting down, and TV interaction. We use a relatively simple processing method where we apply an IIR filter to extract movements from the frames prior to feeding them to a convLSTM network for the classification. We achieved an overall mean peak accuracy of 86%, with the accuracy of all classes reaching at least 85%.

The method managed to identify TV-interaction actions with a peak accuracy of 97.5%. When the IIR filter is not used the accuracy is about 4-5% lower.

State of the art classification methods using deep neural networks [10] and state of the art pre-processing techniques [9] are bound to improve our results dramatically. A better overall accuracy needs to be achieved to allow a robust and reliable implementation of smart functions in a real home. Additional data will be collected to avoid overfitting. We will also improve our testing method by using the k-fold cross-validation method, which would be a better testing approach with all data being used at least once in the training and testing sets. A combination with other information, e.g. location in the room, will enable better activity recognition. Finally, the work will be expanded to classify additional movements/ actions that are relevant for the realization of smart functionalities that are potentially useful for older adults in general, and those with MCI/D in particular.

Acknowledgment

The authors wish to thank the residents and the housekeepers at the seniors' care unit Skøyen Omsorg+; Torhild Holthe and Dr. Anne Lund (OsloMet) for recruiting participants for the trial and communicating with the residents throughout; and the rest of the participants of the Assisted Living Project for a fruitful interdisciplinary collaboration. Finally, the collaboration with RoomMate AS is greatly appreciated. (Flávia Dias Casagrande and Oda Olsen Nedrejord are co-first authors.)

References

[1] E. Zouganeli, F. Casagrande, T. Holthe, A. Lund, L. Halvorsrud, D. Karterud, A. Flakke-Johannessen, H. Lovett, S. Mørk, J. Strøm-Gundersen, E. Thorstensen, R. Norvoll, R. Meulen, M.-R. Kennedy, R. Owen, M. Ladikas, and E.-M. Forsberg, "Responsible development of self-learning assisted living technology for older adults with mild cognitive impairment or dementia," *ICT4AWE 2017 - Proceedings of the 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health*, no. Ict4awe, pp. 204–209, 2017.

[2] B. Winblad, K. Palmer, M. Kivipelto, V. Jelic, L. Fratiglioni, L.-O. Wahlund, a. Nordberg, L. Bäckman, M. Albert, O. Almkvist, H. Arai, H. Basun, K. Blennow, M. de Leon, C. DeCarli, T. Erkinjuntti, E. Giacobini, C. Graff, J. Hardy, C. Jack, a. Jorm, K. Ritchie, C. van Duijn, P. Visser, and R. C. Petersen, "Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment." *Journal of internal medicine*, vol. 256, no. 3, pp. 240–246, 2004.

[3] "RoomMate," https://www.roommate.no/, [Online; accessed 16-November-2018].

[4] M. E. L. A. C. S. T. H. J. . D. G. Reeder, B., "Framing the evidence for health smart homes and home-based consumer health technologies as a public health intervention for independent aging: a systematic review," *International journal of medical informatics*, vol. 7, no. 82, pp. 565–579, 2013.

[5] M. Schwenk, J. Mohler, C. Wendel, K. D'Huyvetter, M. Fain, R. Taylor-Piliae, and B. Najafi, "Wearable sensor-based in-home assessment of gait, balance, and physical activity for discrimination of frailty status: Baseline results of the arizona frailty cohort study," *Gerontology*, vol. 61, 12 2014.

[6] S. Chaudhuri, H. Thompson, and G. Demiris, "Fall detection devices and their use with older adults: A systematic review." *Journal of geriatric physical therapy (2001)*, vol. 37, 12 2013.

[7] R. Chen and P. Schulz, "The effect of information communication technology interventions on reducing social isolation in the elderly: A systematic review," *Journal of Medical Internet Research*, vol. 18, p. e18, 01 2016.

[8] J. Hoey, P. Poupart, A. von Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis, "Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process," *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 503–519, 2010.

[9] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A Review on Human Activity Recognition Using Vision-Based Method," *Journal of Healthcare Engineering*, vol. 2017, 2017.

[10] Z. Zhang, X. Ma, R. Song, X. Rong, X. Tian, G. Tian, and Y. Li, "Deep learning based human action recognition: A survey," *Proceedings - 2017 Chinese Automation Congress, CAC 2017*, vol. 2017-Janua, no. October, pp. 3780–3785, 2017.

[11] J. Davis and A. Bobick, "The representation and recognition of human movement using temporal templates," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, no. September, pp. 928–934, 2015.

[12] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723.

[13] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, 2012, p. 1057.

[14] X. Yang and Y. Tian, "Super Normal Vector for Activity Recognition Using Depth Sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 804–811.

[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[16] X. Yang and Y. L. Tian, "EigenJoints-based action recognition using Naïve-Bayes-Nearest- Neighbor," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2012, pp. 14–19.

[17] T. Darrell and A. Pentland, "Space-time gestures," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Comput. Soc. Press, 1993, pp. 335–340.

[18] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2012, pp. 20–27.

[19] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action Recognition from Depth Maps Using Deep Convolutional Neural Networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, 2016.

[20] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, Aug 2016.

[21] "Services for sensitive data (TSD," //www.uio.no/english/services/it/research/sensitive-data/index.html, [Online; accessed 16-November-2018].

[22] P. Y. Simard, D. Steinkraus, and C. J. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," in *ISeventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 2003, pp. 958–963.

[23] R. Vaillant, C. Monrocq, and Y. L. Cun, "Original approach for the localisation of objects in images," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 141, no. 4, p. 245, 1994.

[24] J. Martens, "Generating Text with Recurrent Neural Networks," *Neural Networks*, vol. 131, no. 1, pp. 1017–1024, 2011.

[25] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech Recognition With Deep Recurrent Neural Networks," *Icassp*, no. 3, pp. 6645–6649, 2013.

[26] A. Graves, "Generating Sequences With Recurrent Neural Networks," *arXiv*, 2014.

[27] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," 2015.