3D Convolutional Neural network for Home Monitoring using Low Resolution Thermal-sensor Array

Lili Tao[1], Timothy Volonakis[1], Bo Tan[2], Ziqi Zhang[1,3], Yanguo Jing[2], Melvyn Smith[1]

[1] University of the West of England
{lili.tao, tim.volonakis, melvyn.smith}@uwe.ac.uk
[2] Coventry University
{bo.tan, ac2716}@coventry.ac.uk
[3] University of Bristol
zz17417@mybristol.ac.uk

Keywords: Low resolution sensor, home monitoring, daily action recognition, 3D convolutional neural network

Abstract

The recognition of daily actions, such as walking, sitting or standing, in the home is informative for assisted living, smart homes and general health care. A variety of actions in complex scenes can be recognised using visual information. However cameras succumb to privacy concerns. In this paper, we present a home action recognition system using an 8×8 infared sensor array. This low spatial resolution retains user visual privacy, but is still a powerful representation of actions in a scene. Actions are recognised using a 3D convolutional neural network, extracting not only spatial but temporal information from video sequences. Experimental results obtained from a publicly available dataset *Infra-ADL2018* demonstrate a better performance of the proposed approach compared to the state-of-the-art. We show that the sensor is considered better at detecting the occurrence of falls and actions of daily living. Our method achieves an overall accuracy of 97.22% across 7 actions with a fall detection sensitivity of 100% and specificity of 99.31%.

1 Introduction

The health systems of many countries are faced with an ever increasing number of long term health conditions given a rising older population. In England alone, 15.4 million people have at least one chronic medical condition, such as, dementia, stroke, cardiovascular or musculoskeletal disease [4]. In such cases, continuous management and medical treatment may be required for many years outside of hospital. Currently this requires 70% of the total National Health Services (NHS) budget [8]. The economic burden of chronic diseases and conditions requires new solutions not only in traditional clinical in-patient treatments, but also out-patient long-term monitoring using innovative technologies.

For these reasons, developing a reliable home monitoring system has drawn much attention in recent years. It can provide clinicians and patients with many solutions, such as early risk detection, early diagnostic, rehabilitation monitoring, treatment suggestions and detection of changes in daily behaviours [6]. Current home monitoring systems often include environmental sensors, wearable inertial sensors and visual sensors. Such systems can enable various types of applications in healthcare provision, such as to help diagnose and manage health and well-being conditions [23].

Wearable sensor based techniques have emerged over recent years that coarsely categorise daily actions, but offer low cost, low energy consumption, and data simplicity. Among these, tri-axial accelerometers are the most broadly used inertial sensors to recognise actions [15]. Wearable sensors are often compared with other types of sensors in terms of accuracy, intrusiveness and efficiency. The issues surrounding missed communications, limited battery life, irregular wearing, and poor comfort remain problematic in wearable devices.

Contactless sensors, such as wireless passive sensing systems and visual sensors, have the potential to address several limitations of the wearable sensors. WiFi is ubiquitous in a home enviornment and wireless signals, such as WiFi, have been exploited to detect human movement [14] - that is, when a person engages in an action, the body movement affects the wireless signals. This technology has shown capabilities in assisted living and residential care [13]. However, Wi-Fi sensing systems still suffer from low accuracy, single-user capability and signal source dependency problems. On the other hand, visual sensors can capture rich data and multiple events simultaneously. There exists a significant body of literature describing the inference of actions from 2D colour intensity imagery [2] and 3D data [3]. Recent advances in computer vision allow for a fine-grained analysis of human action that can be used in more healthcare related applications, such as estimation of calorific expenditure at home [16] and assess the correctness of movement for rehabilitation monitoring [17]. These have now opened up the possibility of integrating these devices seamlessly into home monitoring systems [22]. However, visual sensors have not been widely integrated, and this is because cameras compromise user privacy.

Visual privacy protection has been discussed recently [11]. To overcome the privacy limitations with cameras, we propose a home action monitoring system using an 8×8 infrared sensor array. The sensor provides 64 pixels of thermal data and can be used to offer coarse classification of action, whilst importantly preserving privacy for the users. We proposed to use the state-of-the-art 3D convolutional neural network based method to recognise daily actions. This architecture can extract both the spatial and temporal features from video sequences automatically without any prior knowledge, thus capture motion information from dynamic postures. The sensor applied in the home monitoring system context is new, where publicly available datasets that use this sensor are relatively limited. We evaluate our method on the dataset, *Infra-ADL2018*, for monitoring actions of daily living and to detect occurrence of falls. We also compare the proposed method against the baseline method [18]. The dataset contains 7 daily actions performed by 8 subjects. The infrared sensor itself is mounted onto the ceiling to give an overhead view. In summary, the major contributions of this paper are, (a) exploration of a low resolution thermal sensor in a healthcare scenario to preserve user privacy, and (b) demonstration that the proposed method is able to achieve high recognition results, especially for detecting fall events.

The remainder of this paper is organised as follows: Section 2 discusses the relevant works to our study. Section 3 describes the proposed framework for recognising human actions. The experimental setup and the results are presented in Section 4.

2 Related work

In recent years, much effort has been made on applying computer vision techniques to help with increasing personal safety and reducing risks at home. However, studies that address privacy concerns in vision-based home monitoring systems have been relatively limited. Our work therefore explores this field further and builds on several relevant subject areas in computer vision.

Infrared sensor array - An infrared sensor array is a device composed of a small number of discrete infrared sensors. It represents the spatial distribution of temperature as a low-resolution image. Unlike colour cameras, infrared sensor arrays only capture the shape of the human body, therefore making individual identification harder. Additionally, the low spatial resolution also makes identification of individuals difficult. As this is more comfortable for users, it becomes more acceptable for installation in residential environments. Such infrared sensor arrays can be applied in many scenarios. A 4×4 sensor array has been used to recognise hand motion directions [21], although the extremely low resolution of this sensor renders it unsuitable for more complex visual tasks. A 8×8 pixel sensor array has been successfully used to detect, count and track people indoors [20]. Human movements has also be inferred by using the subject's location and moving trajectory using a 16x16 sensor array [7]. Most recently, a system uses a 8×8 pixel sensor array has been designed for human action recognition [18]. A temporal and spatial Discrete Cosine Transform is used to construct features, which our method will be compared against.

Hand-crafted feature extraction - The visual trace of human action in video forms a spatio-temporal pattern. Here the salient features are well-developed for images captured by conventional visible-light RGB cameras [2]. Traditional approaches that rely on the construction of hand-crafted features for effective video analysis have been introduced in the area of action recognition [3].

Several features have been investigated specifically for low resolution infrared sensors, most notably [5]. Connected component analysis was used to evaluate the number of individuals present in the scene, which subsequently led to motion tracking of the individuals; however this method was sensitive to background noise. A thermo-spatial sensitive histogram feature approach was able to reduce the noise from background pixels [7]. Although counting and tracking of individuals is a non-trivial task, here we are concerned with the action of each individual. Intuitively, this would appear to require finer detail, and this poses a difficult task given the low spatial resolution of the image.

However, the majority of well-developed features, such as histogram of oriented gradients or optical flow, are not appropriate and applicable for very low resolution images such as those captured in this study.

Deep neural network based action recognition - In contrast to extract hand-crafted features, there is a growing trend toward image features learned by deep neural network [10]. Deep neural network models can learn a hierarchy of features from low-level feature to build high-level ones, and therefore automating process of feature extraction. Modelling the temporal information in a video has been the key difference between video and image models, and this has provided an additional and important clue for recognition. The work in [12] learnt spatial and temporal features consecutively. It captured the complementary information on appearance from still frames and also motion between frames. 3D convolution neural network (ConvNet) has been applied to video stream analysis and recognition tasks in recent years. A 3D ConvNet approach was firstly introduced for action recognition [9]. The method requires segmentation of videos based on human detector and head tracking, thus it is difficult to train on large scale dataset. Later, a work in [19] exploited 3D ConvNet in the context of large-scale supervised training datasets. Their proposed feature descriptor is generic and compact, and it works well even with a simple linear classifier. In order to take full use of the advantages of deep neural network, we apply 3D convolutional deep network to action recognition from a low resolution video sequences captured by an infrared sensor array. We also use $3 \times 3 \times 3$ convolution kernel for all layers that has been suggested to produce the best performance in [19].

3 Proposed method

3.1 Pre-processing

Images may be captured on different days and variation in room temperature may add unwanted and uninformative variation.

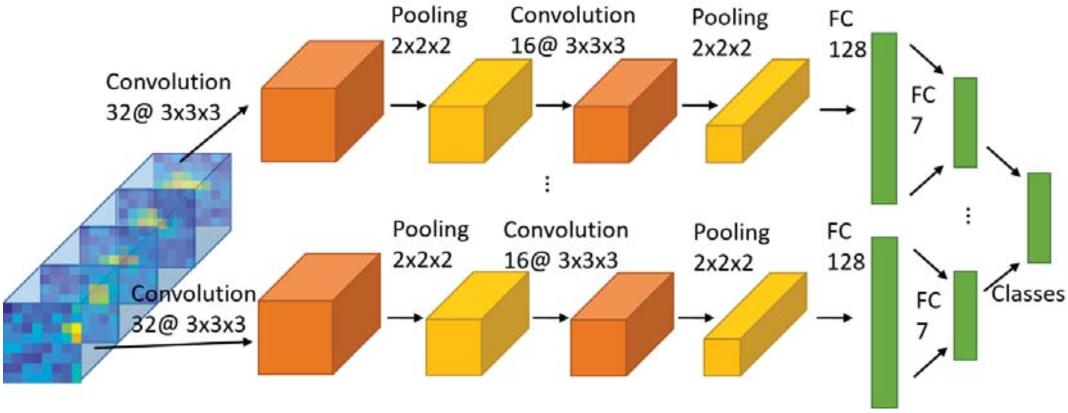

Figure 1. Overview of the proposed 3D ConvNet architecture for action recognition.

The signal to noise ratio was increased by estimating the background noise using sequence of backgrounds and then subtracting this from any frame to be evaluated. Firstly F frames without human subjects in them are identified as background images. The background image B_i is formed by averaging of all i^{th} pixels along the sequence $B_i = \frac{1}{F}\sum_{f=1}^{F} B_i^f$. The processed background subtracted image is extracted by subtracting the corresponding pixel at the background image, $P_i = \hat{P}_i - B_i$, where P_i and $\hat{P}_i$ are the processed image and raw image, respectively.

3.2 3D-ConvNet architecture

We refer video clips with a size of $h \times w \times f$, where h and w are the height and width of the image, respectively, and f is the number of frames in video clip. The kernel size for 3D convolution and pooling layers are referred as $k \times k \times d$, where k and d are spatial and temporal size of kernel, respectively.

The 3D-ConvNet architecture is depicted in Figure 1. 20 Frames are evaluated each time, i.e. to evaluate the action at the present time, the current and previous 19 video frames are evaluated. The network architecture comprises two stacks of sequential convolution-pooling, followed by a fully connected layer.

The input layer is an $8 \times 8 \times 20$ tensor with 32 feature maps. The first convolution layer consists of 32 feature maps produced by $3 \times 3 \times 3$ spatial and temporal 3D convolutional kernels. The convolutional layer consists of convolution, batch normalisation, and activation layers. A 3D max-pooling with kernel size $2 \times 2 \times 2$ is then applied. The second convolution layer used 16 feature maps with $3 \times 3 \times 3$ 3D kernels followed by 3D max-pooling with kernel size $2 \times 2 \times 2$. All of these convolution layers are applied with padding on both spatial and temporal dimensions with stride length of 1. These were further followed by a fully connected softmax classifier with 128 neurons in their hidden layer, which produced class probabilities for actions.

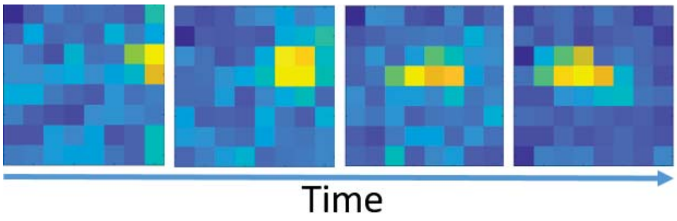

Figure 2. Selected frames from action *fall* sequence.

4 Experimental results

Dataset The dataset is collected from the Grid-EYE 8×8 infrared sensor array developed by Panasonic [1]. The I*nfra-ADL2018* dataset is introduced in [18] for monitoring home actions and detect occurrence of falls. The dataset includes 8 subjects performing 7 daily actions: *fall, sit still, stand still, sit to stand (sit2stand), stand to sit (stand2sit), walking from left to right (walk L2R)*, and *walking from right to left (walk R2L)*. Each action is performed 3 times by the same subject. Figure 2 shows the selected frames from an *fall* sequence.

Figure 3 is an overview of data that shows 7 actions performed by three subjects. It represents the pixel intensity changes along the time. It is observed that subjects may perform same actions differently, e.g. actions may performed in different position within an image. For example, in Walking actions, subject 2 and subject 3 tended to start in different positions that led the different pixels triggered in images.

Implementation details
The proposed network was implemented and trained in Keras using Tensorflow as a backend. The activation function adopted was a rectified linear unit (ReLU). A mean squared error is applied to form a loss function. Optimal parameters were found by training each network for 500 epochs and selecting the model with the minimum validation loss of training.

Quantitative evaluation
We test the proposed method on the dataset *Infra-ADL2018*. We perform a leave-one-subject-out cross validation. The confusion matrix of the action recognition is shown in Figure

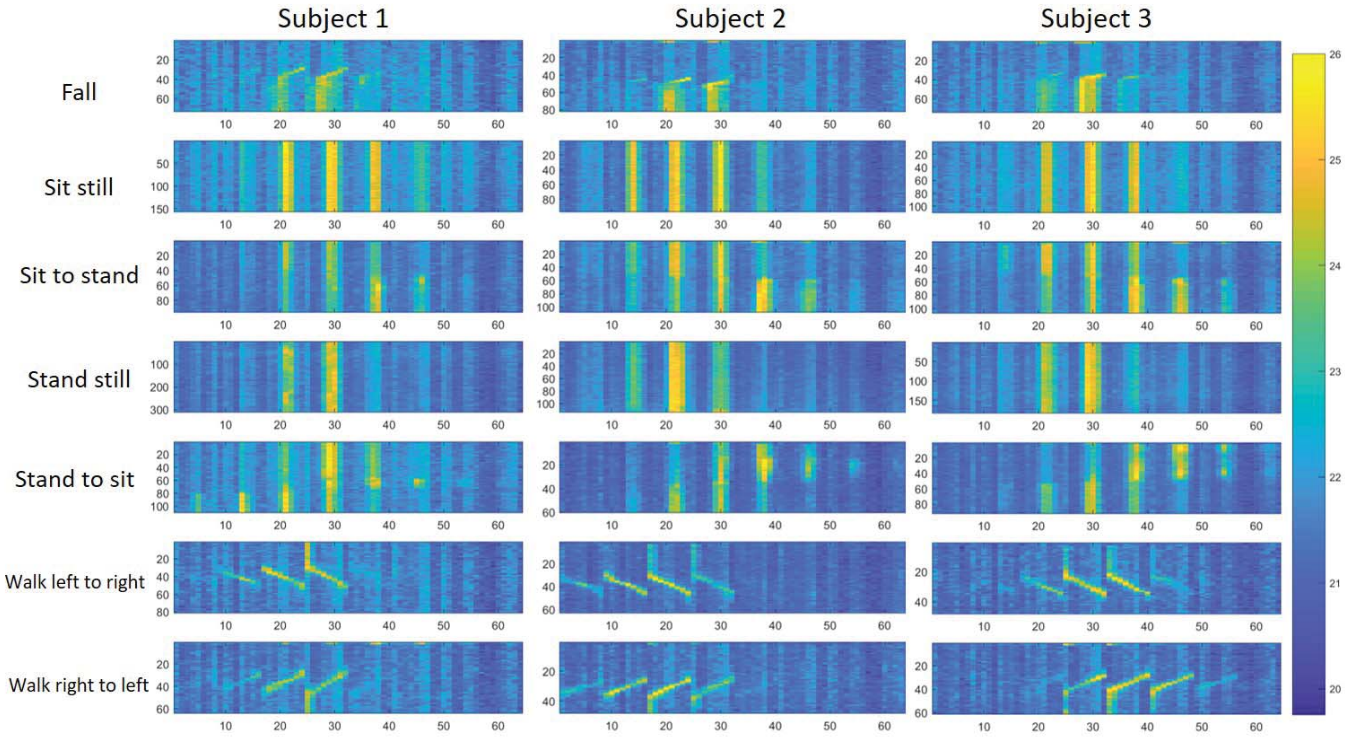

Figure 3. Example sequences of seven action performed by three subjects. For each plot, each row is a video frame, and each column is an element in the 8 × 8 array. Pixel intensity corresponds to temperature.

	Fall	Sit still	Sit2stand	Stand still	Stand2sit	Walk L2R	Walk R2L	Overall
Method in [18]	100	71.43	90.48	61.11	90.47	95.24	100	87.50
Proposed	100	100	90.48	100	95.24	100	95.24	97.22

Table 1. The average recognition accuracy for each action. The proposed method compares with the exiting method using hand-crafted features on *Infra-ADL2018* dataset.

4. The average recognition rate of the proposed method is 97.22%. More precisely, it can reach 100% sensitivity for detecting the occurrence of fall, and 99.31% specificity, indicating very few false fall alarms. The only action that is misclassified as a fall is walking right to left. This is because when a subject falls over, they tend to walk first and then fall. Figure 2 shows a subject walking from right to left and then fall in the middle. *Sit to stand* and *stand to sit* are likely to misclassified as *sit still*. This is because *sit still* is part of these two actions.

Figure 5 presents a visual depiction of the confidence of prediction to certain action. Each horizontal bar corresponds to one action performing by all subjects. It shows that the prediction is less confident at the beginning or at the end of the sequence. This intuitive, since the time sequence at the middle of an action is most discriminative. Actions *stand to sit* and *sit to stand* are both mistaken for sitting. This is reflected in the confidence score where confidence is low at the beginning and at the end for *sit to stand stand to sit* respectively; this is intuitive because you must be sitting to stand up, and you must be standing to sit down.

5 Conclusion and Future work

In this paper, we proposed a home action monitoring method using an infrared sensor array. The proposed method uses a 3D convolutional neural network to extract retain spatial and temporal information that is suitable for representing human action in very low-resolution thermal images. Given that the sensor would be used in home environments, potential future directions include a multi-sensor system that comprises multiple viewing angles that can deal with view-invariance and occlusion.

References

[1] Sensors for automotive and industrial applications: Grid-eye infrared array senor. `https://na.industrial.panasonic.com/products/sensors/sensors-automotive-industrial-applications/grid-eye-infrared-array-sensor`.

[2] Jake K Aggarwal and Michael S Ryoo. Human activity

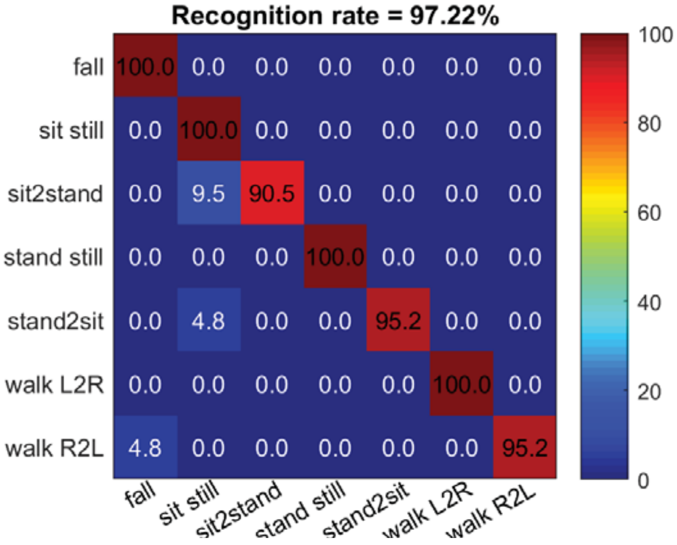

Figure 4. The confusion matrix of the action recognition, corresponding to an average recognition rate of 97.22%

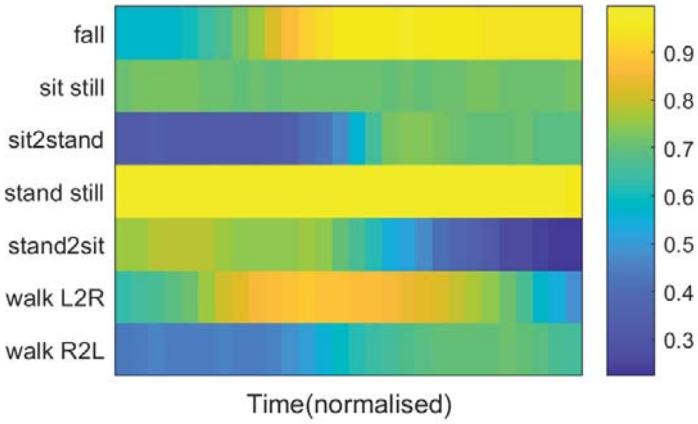

Figure 5. Confidence of prediction to each action. Colour indicates the confidence level.

analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[3] Jake K Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[4] Zofia Bajorek, Anthony Hind, and Stephen Bevan. *The impact of long term conditions on employment and the wider UK economy*. Work Foundation, 2016.

[5] Chandrayee Basu and Anthony Rowe. Tracking motion and proxemics using thermal-sensor array. *arXiv preprint arXiv:1511.08166*, 2015.

[6] Gianluca Castelnuovo, Giancarlo Mauri, Susan Simpson, Angela Colantonio, and Stephen Goss. New technologies for the management and rehabilitation of chronic diseases and conditions. *BioMed research international*, 2015, 2015.

[7] Takashi Hosono, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Tomoyoshi Aizawa, and Masato Kawade. Human tracking using a far-infrared sensor array and a thermo-spatial sensitive histogram. In *Asian Conference on Computer Vision*, pages 262–274. Springer, 2014.

[8] Gareth Iacobucci. Nhs in 2017: Keeping pace with society. *BMJ: British Medical Journal (Online)*, 356, 2017.

[9] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[10] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[11] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177–4195, 2015.

[12] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[13] Bo Tan, Qingchao Chen, Kevin Chetty, Karl Woodbridge, Wenda Li, and Robert Piechocki. Exploiting wifi channel state information for residential healthcare informatics. *IEEE Communications Magazine*, 56(5):130–137, 2018.

[14] Bo Tan, Karl Woodbridge, and Kevin Chetty. Awireless passive radar system for real-time through-wall movement detection. *IEEE Transactions on Aerospace and Electronic Systems*, 52(5):2596–2603, 2016.

[15] Lili Tao, Tilo Burghardt, Sion Hannuna, Massimo Camplani, Adeline Paiement, Dima Damen, Majid Mirmehdi, and Ian Craddock. A comparative home activity monitoring study using visual and inertial sensors. In *E-health Networking, Application & Services (HealthCom), 2015 17th International Conference on*, pages 644–647. IEEE, 2015.

[16] Lili Tao, Tilo Burghardt, Majid Mirmehdi, Dima Damen, Ashley Cooper, Sion Hannuna, Massimo Camplani, Adeline Paiement, and Ian Craddock. Calorie counter: Rgb-depth visual estimation of energy expenditure at home. In *Asian Conference on Computer Vision*, pages 239–251. Springer, 2016.

[17] Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi, Sion Hannuna, Massimo Camplani, Tilo Burghardt, and Ian Craddock. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Computer vision and image understanding*, 148:136–152, 2016.

[18] Lili Tao, Timothy Volonakis, Bo Tan, Yanguo Jing, Kevin Chetty, and Melvyn Smith. Home activity monitoring using low resolution infrared sensor. *arXiv preprint arXiv:1811.05416*, 2018.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[20] Anna A Trofimova, Andrea Masciadri, Fabio Veronese, and Fabio Salice. Indoor human detection based on thermal array sensor data and adaptive background estimation. *Journal of Computer and Communications*, 5(04):16, 2017.

[21] Piotr Wojtczuk, Alistair Armitage, T David Binnie, and Tim Chamberlain. Pir sensor array for hand motion recognition. In *Proc. 2nd Int. Conf. on Sensor Device Technologies and Applications*, pages 99–102, 2011.

[22] Przemyslaw Woznowski, Xenofon Fafoutis, Terence Song, Sion Hannuna, Massimo Camplani, Lili Tao, Adeline Paiement, Evangelos Mellios, Mo Haghighi, Ni Zhu, et al. A multi-modal sensor infrastructure for healthcare in a residential environment. In *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pages 271–277. IEEE, 2015.

[23] Ni Zhu, Tom Diethe, Massimo Camplani, Lili Tao, Alison Burrows, Niall Twomey, Dritan Kaleshi, Majid Mirmehdi, Peter Flach, and Ian Craddock. Bridging e-health and the internet of things: The sphere project. *IEEE Intelligent Systems*, 30(4):39–46, 2015.