In-House Deep Environmental Sentience for Smart Homecare Solutions toward Ageing Society

Philip Easom[1], Ahmed Bouridane[1], Feiyu Qiang[2], Li Zhang[1], Carolyn Downs[2] and Richard Jiang[2]
[1]*Department of Computer and Information Sciences, Northumbria University, Newcastle-upon-Tyne, NE1 8SB, UK*
[2]*LIRA Center, Lancaster University, Lancaster, LE1 4YW, UK*

***Abstract*—With an increasing amount of elderly people needing home care around the clock, care workers are not able to keep up with the demand of providing maximum support to those who require it [43]. As medical costs of home care increase the quality is care suffering as a result of staff shortages [43], a solution is desperately needed to make the valuable care time of these workers more efficient. This paper proposes a system that is able to make use of the deep learning resources currently available to produce a base system that could provide a solution to many of the problems that care homes and staff face today. Transfer learning was conducted on a deep convolutional neural network to recognize common household objects was proposed. This system showed promising results with an accuracy, sensitivity and specificity of 90.6%, 0.90977 and 0.99668 respectively. Real-time applications were also considered, with the system achieving a maximum speed of 19.6 FPS on an MSI GTX 1060 GPU with 4GB of VRAM allocated.**

Keywords—Smart homecare, AIoT, environmental sentience

I. Introduction

With the ageing global population following an increasing trend, the amount of people aged over 60 is expected to double by 2050 [1]. In 2017, approximately 1 in 8 people were aged 60 or over [1]. This percentage is expected to increase to 1 in 5 by 2050 [1].

Due to the increase in the ageing population and the demographic disparity between people of working age compared to rising numbers of over 60s, care home workers are increasingly unable to provide support to all residents within care homes or in domiciliary settings as demands increasingly outstrip staffing levels and ability [43].

Within the US alone, the predicted costs of medical care are projected to rise from 3% to 5.5% by 2050 [2] as a result of the aging population, so a promising solution to assist in delivering effective care to the elderly is needed in support care workers, reduce costs and increase efficiency globally.

In a staffing levels survey by UNISON [3], 49% of respondents reported not having adequate time with each patient, 45% felt there were not enough staff to deliver care and 65% reported care needs were being left unmet because of understaffing. Developing a solution to allow care workers to provide support to each person as efficiently as possible is urgently needed [43].

Approximately 28-35% of people aged 65 and over fall at least once each year [4], with the probability of falling increasing as they age. Elderly people within care homes are more likely to fall over than older people living in the community. One study [5] found that 15% of people who had fallen were on the floor for over an hour.

Serious injury is also found to be heavily associated with the length of time a person spends on the floor unable to get up [4]. With the disproportion between care workers and elderly people widening each year, assistance and potential medical attention might not reach the person quickly enough. As hospitalization costs from fall related injuries are expected to increase to $240 billion by 2040 [6], being able to efficiently care for elderly people could help reduce injuries and the cost associated with them. Smart home surveillance [42] becomes urgently needed by the ageing society.

Deep learning is a modern subfield of machine learning that focuses primarily on deep artificial neural networks (ANN's) and their counterparts [23]-[28]. Advances into convolutional neural networks (CNN's) have been substantial since one of the original models LeNet5 [7], first released in 1998.

With recent improvements in the field of deep learning to deep convolutional neural network capabilities and processing speed, the use of these networks has huge potential to be used within many healthcare applications. CNN's have been shown to provide model solutions for human behaviour recognition [8], fall detection [9, 10] and the maintenance of independent living for seniors [11] which are all important in providing efficient caring [43].

A base system model that is able to be developed and tailored towards an all-round solution is proposed here. The aim of the system is to be able to detect common salient objects[29][30] found within a home environment to a high degree of accuracy. The proposed system should also be able to process detected objects in real-time, to give maximum applicability.

II. Related Work

Related work for the system proposed above will include discussion into different deep convolutional neural network models. The accuracy of the models will be evaluated, as well as their real-time suitability.

Image classification models can produce very promising accuracy results that could be applied to this system. Models such as GoogLeNet [12] and ResNet [13] both won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014 and 2015 respectively. These models performed extremely well, with ResNet achieving an error rate of 3.6% [13] in the challenge. Whilst these models provide excellent accuracy results, they lack the real-time suitability that a system such as this would need.

Convolutional neural network models have also been developed to provide better real-time capability that other. Faster R-CNN [14], which is an improved version of R-CNN [15] and Fast R-CNN [16], increased the speed of this system to 7 frames per second (FPS) whilst keeping a respectable mean average-precision (mAP) of 73.2. They achieved this by the addition of a region proposal layer, that took the final feature map of the system and produced region proposals for objects from that.

Mask R-CNN [17] is an extension of the R-CNN model series [14, 15, 16] that attempts to produce detections on a pixel level. The result of this are masks that cover the detected

objects. This has potential to be used in many applications such as dense human pose estimation [18] which can be adapted to provide a substantial monitoring of fall detection platform. However, the speed of this model is limited as the model was only able to achieve ~5 FPS when processing images.

Regression object detection models can be better suited to real-time tasks than the other models discussed but can suffer drawbacks in accuracy as a result. You Only Look Once (YOLO) [19] is a regression model that gives bounding box predictions and confidence scores associated with them for the entire image. This model is able to achieve 78.6 mAP alongside a processing speed of 40 FPS.

III. System Methods

The system structure makes use of the deep convolutional neural network 'YOLOv2', as described in [20] by Redmon et al. YOLOv2 is an improved version of the original model 'YOLO' [19]. The network consists of 24 convolutional layers and 4 max pooling layers. The first 20 layers are pre-trained on the ImageNet 1000 class dataset and the final four are initialized with randomized weights.

The convolutional layers within the network employs various kernel sizes ranging from 1x1 to 3x3. The activation function used within the network is a leaky ReLU. This is used after every convolutional layer.

Batch normalization is also used within the network during training to assist in the convergence of the network whilst reducing the need of other regularization methods to be used [21]. The algorithm for batch normalization is displayed below:

$$BN_j = \frac{\gamma(z_j - \mu)}{\sqrt{\sigma}} + \beta$$

Where μ is the batch mean, γ represents a scale factor, σ is the batch variance, β is the offset and z_j is the output from the convolutional layer. These parameters are learned and updated throughout training.

The training data for the system is 7 of the 20 classes from the Pascal VOC Dataset, as well as a custom object dataset including an extra object class. These classes are all objects commonly found within a home environment so provide specificity to the task.

Backpropagation was used to update the weight parameters of the network. Stochastic gradient descent was also used alongside this with a learning rate of 1×10^{-5}. The steps used to train the system are as follows:

1) Load the initialized network
2) Load the training data and annotations
3) Process training image through each layer, performing batch normalization after each convolutional layer
4) Perform backpropagation and update weights of each layer.
5) Repeat steps 3 and 4 on each training image and continue process until convergence is reached

IV. Prepare Your Paper Before Styling

The proposed system was trained for 40 hours on an MSI GTX 1060 GPU with 4GB memory allocated to the training. 4 training checkpoints were saved to test the systems convergence rate and performance. These were at 29,000, 69,000, 100,000 and 121,200 training iterations. This was approximately 37 full epochs of the training data.

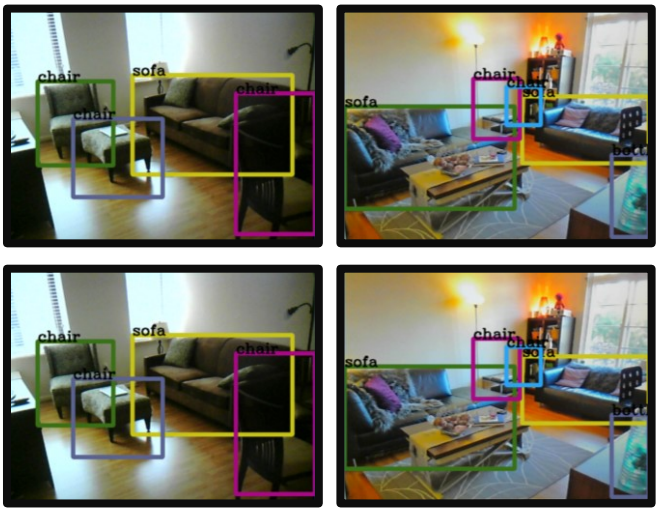


Fig. 1 Example detections taken from the test images

Table 1 Dataset classes and instances

Class	Training Instances	Testing Instances
Person	17,401	211
Chair	3056	180
Bottle	1561	120
Dining Table	800	140
TV Monitor	893	127
Sofa	841	130
Potted Plant	1202	121
Lamp	486	73

Table 2 System accuracy

Iterations	Accuracy	Sensitivity	Specificity	Precision
29,000	56.3%	0.52908	0.96166	0.67182
69,000	90.6%	0.90977	0.99668	0.97459
100,000	85.7%	0.86929	0.99493	0.94489
121,200	88.5%	0.88413	0.99688	0.97616

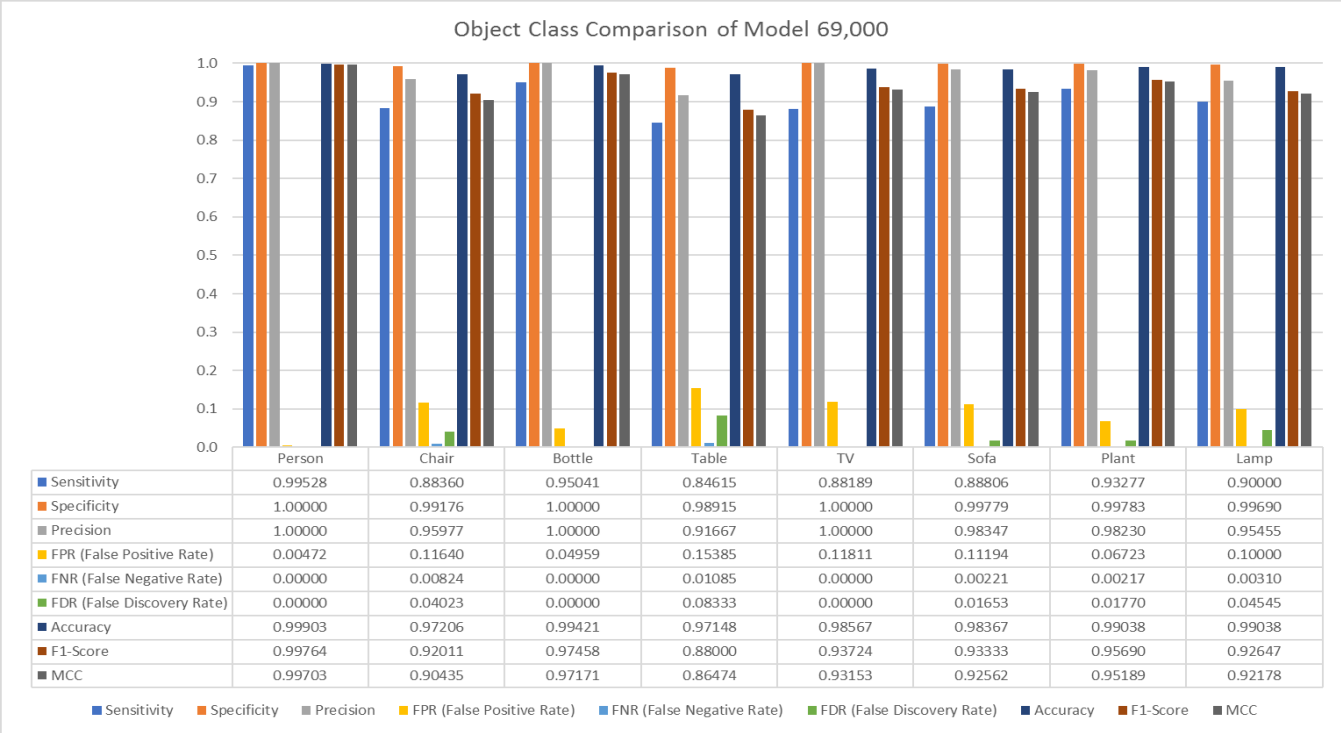
Table 3 System processing speed

	Objects Detected (FPS)	No Objects Detected (FPS)
CPU	2.0	2.1
GPU	10.7	19.6

The performance of the system was evaluated by assessing its' ability to correctly detect objects within test images used. The test images used were taken from the Pascal VOC test dataset, alongside other images sourced online. Accuracy was determined as an average of correctly to incorrectly detected objects. Sensitivity, Specificity and Precision values can be calculated as follows:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$



	Person	Chair	Bottle	Table	TV	Sofa	Plant	Lamp
Sensitivity	0.99528	0.88360	0.95041	0.84615	0.88189	0.88806	0.93277	0.90000
Specificity	1.00000	0.99176	1.00000	0.98915	1.00000	0.99779	0.99783	0.99690
Precision	1.00000	0.95977	1.00000	0.91667	1.00000	0.98347	0.98230	0.95455
FPR (False Positive Rate)	0.00472	0.11640	0.04959	0.15385	0.11811	0.11194	0.06723	0.10000
FNR (False Negative Rate)	0.00000	0.00824	0.00000	0.01085	0.00000	0.00221	0.00217	0.00310
FDR (False Discovery Rate)	0.00000	0.04023	0.00000	0.08333	0.00000	0.01653	0.01770	0.04545
Accuracy	0.99903	0.97206	0.99421	0.97148	0.98567	0.98367	0.99038	0.99038
F1-Score	0.99764	0.92011	0.97458	0.88000	0.93724	0.93333	0.95690	0.92647
MCC	0.99703	0.90435	0.97171	0.86474	0.93153	0.92562	0.95189	0.92178

Fig.2. Detection of common objects in a smart home.

$$Precision = \frac{TP}{(TP + FP)}$$

where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP is the number of false positives.

The processing speed of the system was also tested for its' suitability to have real-time applications. Results were produced for the speed of the system on an MSI GTX 1060 GPU with 4GB memory allocated, as well as an Intel I5 8600k CPU. These results were given in frames per second (FPS) and an average value from 10 seconds of the system running was produced.

V. System Evaluation and Discussion

Table 2 shows the results of the systems' accuracy, sensitivity, specificity and precision at the training iterations shown. Among all 4 models, model 69,000 performs the best overall. After model 69,000, the performance of the system starts to drop before rising again slightly at model 121,200. From these results, it can be determined that the model converges at approximately 69,000 training iterations as model 29,000 detects objects poorly. The decrease in accuracy after this point could be a result of the network overfitting to the training data.

Table 3 shows the results from measuring the systems processing speed. The CPU values do not differ much with the presence of detections or not, whereas the GPU values do decrease substantially with the addition of detected objects.

From the chart in Fig,2, you can see that the person class performs the best overall, which is not surprising as the system correctly detected all of the images within this class. An interesting point within this model is that the lamp class performed much better than the chair and table. This was slightly unexpected as the lamp had been trained on a dataset with inferior quality to all other classes. However, the dining table is a fairly hard object to detect as it is usually surrounded by chairs, so is not so obvious to a machine as other objects that stand out more. As you can see, the sensitivity result for the chair and sofa class is quite low. This shows that the system is not so confident in its prediction, which is expected since the difference between a chair and sofa is relatively small. The dining table class also produced the highest FNR result, which could agree with the comment that they are usually surrounded by chairs, so the system is not so confident in its' prediction and struggles to see the object between the chairs.

The results of the proposed system show potential to have applications within home care. The accuracy of detections is reliable enough to have applications within home care, as shown in figure 1 Further development could produce a system that would be able to process the locations of these detections relative to one another. For instance, if there was a chair and a person within a certain boundary or shape, then the system would produce a status that the person was sitting down. There are many other common scenarios where the system could produce notifications like these, such as drinking, watching tv or walking.

The system could also be adapted to trigger alerts on significant actions such as falling over. Convolutional neural networks have been shown to adapt to this problem well [9, 10].

By implementing additional features within the system, care homes would be able to allocate staff time more efficiently. Higher priority residents would be able to receive the quality care needed instead of the staff monitoring other residents at risk of falls but otherwise of lower dependency, meaning their time could be used more effectively. Being able to manage care workers time more efficiently would save money as there would be a lower demand for more staff. With

the average UK care worker salary being between £12,500 and £25,000 per annum [22], the implementation of this system could save substantial amounts of money, whilst improving care quality.

VI. Conclusions

A system has been created that once fully developed, has the potential to provide many benefits to the care of elderly people globally. The system makes use of the deep convolutional neural network YOLOv2 [20] and shows potential for multiple caring applications, reducing well-documented pressure and stressors for care-workers [43]. The promising accuracy of the system means it can be adapted to provide useful information about the persons' current status, allowing care workers' time to be used and managed more efficiently. This could lead to the quality of care increasing to the people who need it most, whilst reducing costs from not requiring to employ larger amounts of staff.

In our future work, we will consider 3D perception of home environment[31], identify any motions in the surveillance home[32]~[33], introduce sounds/speech analysis with visual perception[34], identify habitants via biometrics[35]~[38], and detect the emotion changes[39]. Furthermore, we will consider its hardware implementation [39]~[42] aiming at a low-power IoT edge deployment, where various functions will be integrated as a whole for smart homecare.

[1] World Population Ageing. United Nations, 2017. Available at http://www.un.org/en/development/desa/population/ publications/pdf/ageing/WPA2017Infochart.pdf

[2] D. W. Elmendorf. The 2014 long-term budget outlook. DTIC Document, 2014.

[3] UNISON's staffing levels survey. UNISON, 2015. Available at https://www.unison.org.uk/content/uploads/2015/04/ TowebRed-Alert-Unsafe-Staffing-Levels-Rising1.pdf

[4] WHO global report on falls prevention in older age. World Health Organization, 2007. Available at http://www.who. int/ageing/publications/Falls_prevention7March.pdf

[5] Fleming, J. and Brayne, C., 2008. Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. Bmj, 337, p.a2227.

[6] Cummings, S.R., Rubin, S.M. and Black, D., 1990. The future of hip fractures in the United States. Numbers, costs, and potential effects of postmenopausal estrogen. Clinical orthopaedics and related research, (252), pp.163-166.

[7] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

[8] Rege, A., Mehra, S., Vann, A. and Luo, Z., Vision-Based Approach to Senior Healthcare: Depth-Based Activity Recognition with Convolutional Neural Networks.

[9] X. Li, T. Pang, W. Liu and T. Wang, "Fall detection for elderly person care using convolutional neural networks," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, 2017, pp. 1-6.

[10] Núñez-Marcos, A., Azkune, G. and Arganda-Carreras, I., 2017. Vision-Based Fall Detection with Convolutional Neural Networks. Wireless Communications and Mobile Computing, 2017.

[11] Luo, Z., Rege, A., Puisol, G., Milstein, A., Fei-Fei, L., Lance-Downing, N. 2017. Computer Vision-based Approach to Maintain Independent Living for Seniors.

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015, June. Going deeper with convolutions. Cvpr.

[13] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[14] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[15] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[16] Girshick, R., 2015. Fast r-cnn. arXiv preprint arXiv:1504.08083.

[17] He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017, October. Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2980-2988). IEEE.

[18] Güler, R.A., Neverova, N. and Kokkinos, I., 2018. Densepose: Dense human pose estimation in the wild. arXiv preprint arXiv:1802.00434.

[19] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[20] Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. arXiv preprint.

[21] Loffe, S. and Szegedy, C., 2015, June. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456).

[22] Care Worker Job Profile. National Careers Service, 2018. Available: https://nationalcareersservice.direct.gov.uk/job-profiles/care-worker

[23] C. Chiang, C. Barnes, P. Angelov and R. Jiang, "Deep Learning based Automated Forest Health Diagnosis from Aerial Images," in IEEE Access, doi: 10.1109/ACCESS.2020.3012417.

[24] R. Jiang and D. Crookes, "Shallow Unorganized Neural Networks Using Smart Neuron Model for Visual Perception", 2019, IEEE Access.

[25] G Storey, R Jiang, A Bouridane, C T Li, "3DPalsyNet: A Facial Palsy Grading and Motion Recognition Framework using Fully 3D Convolutional Neural Networks", 2019, IEEE Access.

[26] Jiang, Z., Chazot, P. L., Celebi, M. E., Crookes, D. & Jiang, R., "Social Behavioral Phenotyping of Drosophila with a 2D-3D Hybrid CNN Framework", 2019, IEEE Access.

[27] G Storey, A Bouridane, R Jiang, "Integrated Deep Model for Face Detection and Landmark Localisation from 'in the wild' Images", 2018, IEEE Access.

[28] G Storey, R Jiang, A Bouridane , "Role for 2D image generated 3D face models in the rehabilitation of facial palsy", 2017, IET Healthcare Technology Letters.

[29] R Jiang, D Crookes, "Deep Salience: Visual Salience Modeling via Deep Belief Propagation", AAAI, 2773-2779, 2014.

[30] R. Jiang and D. Crookes, "Visual saliency estimation through manifold learning", Proc. Natl. Conf. Artif. Intell., pp. 2773-2779, 2012.

[31] R Jiang, M Parry, P Legg, D Chung, I Griffiths, "Automated 3D animation from snooker videos with information theoretic optimization", 2013, IEEE Trans. Comput. Intell. AI Games.

[32] R Jiang, D Crookes, N Luo, MW Davidson, "Live-cell tracking using SIFT features in DIC microscopic videos", 2010, IEEE Transactions on Biomedical.

[33] S Al-Maadeed, R Almotaeryi, R Jiang, A Bouridane, "Robust human silhouette extraction with Laplacian fitting", 2014, Pattern Recognition Letters.

[34] R. Jiang, A. H. Sadka, D Crookes, "Multimodal Biometric Human Recognition for Perceptual Human–Computer Interaction", IEEE Trans. Systems, Man, & Cybernetics, Part C, Vol.40, Issue 5, 2010.

[35] R Jiang, D Crookes, N Luo, "Face recognition in global harmonic subspace", 2010, IEEE Transactions on Information Forensics and Security.

[36] R Jiang, S Al-Maadeed, A Bouridane, D Crookes, ME Celebi, "Face recognition in the scrambled domain via salience-aware ensembles of many kernels", 2016, IEEE Transactions on Information Forensics and Security.

[37] R Jiang, A Bouridane, D Crookes, M Celebi, HL Wei, "Privacy-protected facial biometric verification via fuzzy forest learning", IEEE Transactions on Fuzzy Systems, 2016.

[38] S Al-Maadeed, M Bourif, A Bouridane, R Jiang, "Low-quality facial biometric verification via dictionary-based random pooling", 2016, Pattern Recognition.

[39] R Jiang, ATS Ho, I Cheheb, N Al-Maadeed, S Al-Maadeed, A Bouridane , “Emotion recognition from scrambled facial images via many graph embedding”, 2017, Pattern Recognition.

[40] D Crookes, S Trainor, R Jiang, “An Environment for Rapid Derivatives Design and Experimentation”, 2016, IEEE Journal of Selected Topics in Signal Processing.

[41] R Jiang, D Crookes, “Using signed digit arithmetic for low-power multiplication”, , Electronics Letters, 2007, Vol.43, No.11, p.613.

[42] I Zafar, U Zakir, I Romanenko, R Jiang, E Edirisinghe, “Human silhouette extraction on FPGAs for infrared night vision military surveillance”, Circuits, Communications and System (PACCS), 2010 Second Pacific-Asia Conference on, 2010.

[43] Pavlidis, G., Downs, C., Kalinowski, T.B., Swiatek-Barylska, I., Lazuras, L., Ypsilanti, A. and Tsatali, M., 2020. A survey on the training needs of caregivers in five European countries. *Journal of nursing management*, *28*(2), pp.385-398.