Automated Detection of Hands and Objects

in Egocentric Videos, for Ambient Assisted Living Applications

Thi Hoa Cuc Nguyen, Jean-Christophe Nebel
and Gordon Hunter
School of Computer Science and Mathematics
Kingston University
Kingston upon Thames, U.K.
cucnth87@gmail.com , J.Nebel@kingston.ac.uk ,
G.Hunter@kingston.ac.uk

Francisco Florez-Revuelta
Department of Computer Technology
University of Alicante, Alicante, Spain
francisco.florez@ua.es

***Abstract*—The need for technology assisted (or ambient assisted) living is increasing all the time as the population ages and the number of people with dementia and other conditions impairing memory and cognitive ability increases. In such applications, amongst others, it is necessary to identify and assess potentially hazardous situations. These include scenarios involving a person's hands and their interactions with various objects. In this paper, we describe our novel approach to identify human hands and objects in videos of people performing a variety of everyday tasks. We compare the performance of our method using different strategies with that of other state of the art approaches. We conclude that, when the proposed approach takes advantage of a pre-trained model, hand detection is performed accurately (94%), providing reliable information for assisted living applications.**

Keywords—egocentric vision; deep convolutional network; ambient assisted living; object detection; hand detection

I. INTRODUCTION

The population of the World is ageing, and in some countries such as Japan, the number of people aged 65 or over already exceeds the number aged under 25 [1]. This trend is expected to continue, and also be reflected in most developed and many developing countries. As the population ages, more and more elderly people are likely to be living alone as they become more infirm and, in some cases, develop dementia. Such senior citizens, along with younger people with disabilities, including mental handicaps, will require assistance when carrying out everyday tasks, and be monitored so that they do not put themselves into hazardous situations. In many common activities – so-called "Activities of Daily Living" (ADL) – people have to manipulate everyday objects using their hands. Having a responsible person supervising what an infirm person is doing all the time is prohibitively expensive and infringe the infirm person's privacy, so there is a need for reliable automated systems to monitor an infirm person's activities, identify when the person is putting him/herself in danger, and either warn them or their carers of the hazard or offer the person advice on carrying out the activity successfully. Such a system would rely on locating and identifying hands and everyday objects in order to infer the activity which is taking place before the relevant hazards could be identified and alerts issued as appropriate. In this paper, we discuss previous approaches used for hand and object detection in egocentric videos (i.e. videos made from the viewpoint of the person of interest) and propose our own method, based on a Faster Region-based Convolutional Neural Network (Faster R-CNN) [2], to perform the same tasks, comparing its performance with those of others, benchmarked with respect to standard Ambient Assisted Living (AAL) video datasets. The rest of this paper is structured as follows. In the following section, we give a brief review of related work in this area. In section III, we introduce the datasets and methodology we have used. The results of our experiments are presented in section IV. Finally, we put forward our conclusions and propose how this work can be extended in the future.

II. RELATED WORK

A. Challenges in Object and Hand Recognition

Automated recognition of objects in videos is a challenging task, since the appearance of objects can vary considerably with viewpoint, illumination and distances, and parts of an object may be occluded from view. For the recognition of human hands, the situation is even more complicated, since the configuration of each hand may change – the relative positions of the fingers may vary, each finger may be held straight or bent to various degrees, and the hands may touch or hold other objects.

B. Existing Techniques for Object Detection and Recognition

Many techniques have been proposed and implemented for the detection and tracking of objects in videos. A large number of these have been comprehensively reviewed in [3]. In this section, we focus on those approaches directly relevant to this study.

Convolutional Neural Networks (CNNs) are a class of multi-layered feed-forward Artificial Neural Networks (ANNs) that have been applied to analysing images using at least three layers of nodes and are designed to minimise the pre-processing required [11, 12] compared with traditional image classification algorithms. Moreover, a major advantage of the

CNN approach is its ability to perform feature design and selection without requiring any human effort. An example of a CNN is VGG16 [5], which has 16 layers of neurons as shown in figure 1. *ReLU* is a Rectified Linear Unit, with an activation function defined as $f(x)=max(0,x)$, the *softmax* function (a normalised exponential function applied to convert the components of a vector into values between 0 and 1, such that the normalised components sum to 1) is used in the final layer to highlight the largest values and suppress values which are significantly below the maximum value. This network has achieved good results for image recognition tasks in Imagenet ILSVRC 2015 Competitions [6].

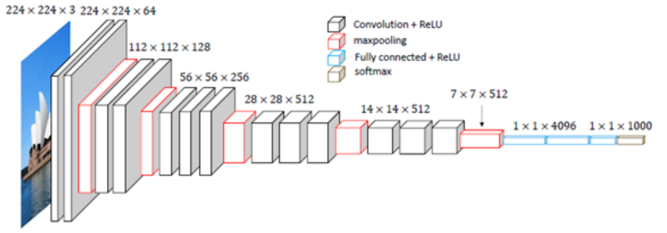

Figure 1: The architecture of VGG16 model. ReLU (rectified linear unit) is an activation function defined by $f(x)$=max(0,x), the *softmax* function is used in the final layer to highlight the largest values and suppress values which are significantly below the maximum value (from https://junjiwon1031.github.io/2017/09/08/Single-Shot-Multibox-Detector.html)

A *Region-based Convolutional Neural Network* (*R-CNN*) is a CNN designed for detecting multiple objects in a single image [13]. R-CNNs first extract multiple regions of interest in an image using a region proposal technique such as selective search [14], then apply a CNN to each region to give each an object category. Selective search is a technique for generating possible object locations for use in object detection by combining exhaustive search and segmentation. It extracts around 2000 bounding boxes which are each then classified into object categories by a many layered CNN such as the VGG16 shown in Figure 1. This method has achieved excellent object detection accuracy when trained with a very large dataset, but the training process is very slow. However, improved versions such as the Fast-RCNN [4] and Faster-RCNN [2] networks are faster and give better recognition performance.

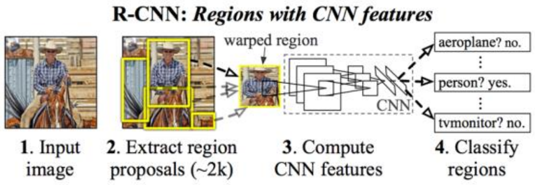

Figure 2: Task pipeline for the R-CNN approach [13].

Instead of using CNN networks for each proposed region separately in the R-CNN model, the Fast-RCNN model first processes the whole image with several convolutional (conv) and max pooling layers to produce a conv feature map. Pooling layers are applied after convolutional layers to reduce the dimensionality of the convolutional layers. Max pooling layers use a *max filter* that keep the maximum value from each subregion, e.g. 2x2 filter, as shown in Figure 3. For each region proposal, a region of interest pooling layer extracts a fixed-length feature vector from the feature map. Each feature vector is fed into a sequence of fully connected layers that finally branch into two output layers: one that produces probability estimates for each of K object classes, and another layer that outputs four numbers encoding bounding box positions for one of these K classes. Due to these innovations, Fast-RCNN significantly improves the training and testing speed compared to RCNN model. Fast-RCNN model trains a VGG16 network nine times faster than the corresponding RCNN network, and is 213 times faster in test time when ignoring the time spent on extraction of region proposals (i.e. selective search).

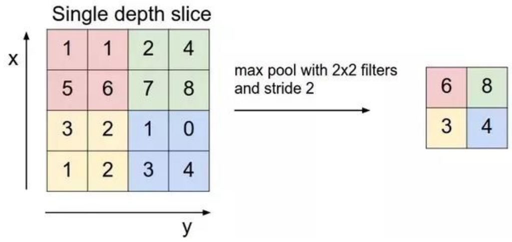

Figure 3: Max pooling each 2x2 subregion of a 4x4 image : “stride 2” means the (2x2) filter is moved on 2 pixels each time. Max pooling means that the largest value in each 2x2 subregion is selected to represent that whole subregion.

Faster-RCNN [2] networks use a many layered fully convolutional neural network to propose regions (Region Proposal Networks - RPNs) instead of selective search, motivated by the observation that convolutional feature maps can also be used for generating region proposals [2]. RPNs learn to propose regions from training data, and thus benefit from deeper and more expressive features. After region proposals are obtained, a Fast-RCNN detector is used for classifying those proposed regions. A Faster-RCNN with RPNs has achieved better results both in speed and accuracy than either Fast-RCNN or RCNN models, and has become a practical and state of the art model for object detection due to its resource-efficient end-to-end training and testing process.

Finally, we note that the Deformable Part-Based Object Detector (DPBOD), described in [20, 21], represents an object as a mixture of multiple deformable parts, then uses a Histogram of Oriented Gradients (HOG) descriptor computed on a dense grid of uniformly-spaced cells. This has been used widely in object detection and recognition applications, including to ADL scenarios [8, 22], with considerable success.

III. IMPLEMENTATION AND EXPERIMENTAL SET-UPS

A. *Implementation Details*

We implemented a Faster-RCNN model including a VGG16 network in Python (with Cython C extensions) on a NVIDIA Titan X GPU with 12GB of GDDR5X memory and a memory speed of 10Gbps. The implementation made use of the Caffe (a deep learning framework developed by Berkeley AI Research), Pycaffe (a set of Python extensions for Caffe) and CUDA (a parallel computing and application programming interface for the GPU created by Nvidia) libraries.

B. *Datasets*

We used three large datasets in this study. For object recognition (Experiments *O1* and *O2*), we employed the Imagenet 200DET dataset [9], a set of (labelled) training and (unlabeled) evaluation videos produced for the “Object Recognition from Video” challenge of the IMAGENET Large

Scale Visual Recognition Challenge [6, 10], plus the ADL dataset [8] for experiment *O2*. For hand detection (Experiments *H1* and *H2*) we used the EgoHands dataset [7] of 4800 images, plus the IMAGENET dataset [9] for pre-training of a VGG16 model.

C. *Experimental Set-Ups*

We performed four sets of experiments – two sets each for object detection and for hand detection – as follows:

a) Experiment O1: We adapted the last layer of a Faster-RCNN model to detect objects from the 200 categories of the Imagenet 200DET dataset [9], a large dataset widely used to benchmark object detectors. The VGG16 image classification model that was already trained on 1.2 billion images of Imagenet's Image Classification Dataset was used to calculate the weights. We trained the model on 70% of the dataset (350,000 images) with 2,100,000 iterations, which corresponded to 6 epochs, with a batch-size of one. The detector model was tested on the remaining 30% of the dataset. We chose the minimum batch-size to keep the highest quality of the network, as it has been observed in practice that larger batch-sizes may degrade the model's ability to generalize. As the ADL dataset [8] only has ground-truth annotations for only 16 types of objects among the dataset's 200 categories, it could not be used for testing here.

b) Experiment O2: We adapted the last layer of the Faster-RCNN network to detect objects from the 42 categories of the ADL dataset. Weighting was performed using a VGG16 image classification model that was already trained on 1.2 billion images of Imagenet's Image Classification Dataset [9]. We trained the detector on 6 videos and tested on 14 videos from the ADL dataset [8]. Due to availability of a powerful GPU, we decided to run the experiments for 100 epochs to achieve good accuracy, as recommended by other researchers. As there are a total of around 12,000 frames in 6 videos, we ran the experiments for 1,200,000 iterations to complete 100 epochs of training, since we set a batch-size of one.

c) Experiments H1 and H2: in both experiments, the last layer of the Faster-RCNN network was altered in order to detect two object classes, namely the the left-hand and right-hand classes. We used the EgoHands dataset [7] to train the hand detector and we ran each experiment with 480,000 iterations, corresponding to 100 epochs as EgoHands contains 4,800 images, and a batch-size of one image per batch. In each case, the system was trained on 75% of the EgoHands dataset and tested on the remaining 25% of the same dataset. While in experiment *H1* the Faster-RCNN with VGG16 network was trained without specifying any weights, in experiment *H2* it was trained with initial weights for a VGG16 image classification model which had already been pre-trained on 1.2 billion images of Imagenet's Image Classification Dataset [9].

IV. Results and Discussion

We evaluated the performance of our object detection models using the mean Average Precision (mAP) metric from the PASCAL VOC 2007 detection benchmarks [15, 16].

TABLE I: Mean Average Precision (mAP) values obtained (as percentages) for each object category in Experiment *O1*.

Object Category	**mAP**	**Object Category**	**mAP**
Bed	33.9	TV	42.7
Wine Bottle	28.7	TV remote control	35.0
Water Bottle	20.5	Toaster	33.2
Refrigerator	23.1	Mug/Cup	33.1
Laptop Computer	51.4	Stove/Cooker	14.3
Microwave Oven	40.5	Vacuum Cleaner	11.1
Pan	10.8	Washing Machine	36.8
Pitcher/Jug	28.5	Keyboard	40.1
Soap	16.2	Computer mouse	22.8

This evaluation measure is algorithm-independent, unlike for example the Detection Error Trade-off used for evaluating pedestrian detectors [17, 18] which is only suitable for sliding window methods. Furthermore, a comparison of the mAP measure and the "area under curve" (AUC) measure [19] on PASCAL VOC2006 showed that the mAP measure highlighted differences between methods to a greater extent [16]. Therefore, mAP has been used for evaluating detectors in this section. Table 1 shows object detection results for experiment *O1*, where we have 70% of the Imagenet 200DET dataset for training and 30% of the same dataset for testing.

For experiment *O2*, we compared the performance of our FasterRCNN model with a VGG16 network with that of the Deformable Part-Based Object Detector (DPBOD), also applied to the ADL dataset and trained on 1200 training instances (1200 bounding boxes) per object category to detect 24 types of objects, described in [8]. However, it should be noted that precise details on training and testing procedures were not specified in [8].

TABLE II: Mean Average Precision (mAP) values obtained (as percentages) for each object category in experiment *O2*. Note that the DPBOD model of [8] was only applied to 10 object categories in the leftmost column.

Object Category	**Faster-RCNN mAP**	**DPBOD mAP [8]**	**Object Category**	**Faster-RCNN mAP**
Tap/Faucett	**54.2**	40.4	Pan	13.5
Soap Liquid	25.8	**32.5**	Book	12.8
Refrigerator	**24.5**	19.9	Tooth-paste	9.3
Microwave	18.7	**43.1**	Dish	7.8
Oven/Stove	**52.1**	38.7	Detergent	6.5
Bottle	10.2	**21.0**	Trash Can	3.8
Kettle	16.5	**21.6**	TV remote control	3.6
Mug/Cup	18.2	**23.5**	Knife/Fork/Spoon	3.4
Washer/Dryer	38.4	**47.6**	Food/Snack	3.4
TV	68.5	**69.0**	Pitcher/Jug	1.8
Laptop	44.5	-	Bed	1.4
Mobile (Cell) 'Phone	20.3	-	Cloth	1.1
Door	19.7	-	Dental Floss	0.2

As Table II shows, the performance of our Faster-RCNN method compares favorably with that of the DPBOD [8] for several categories, the former outperforming the latter in 3 cases, and achieving a very similar performance in one further case. However, for many categories DPBOD considerably outperformed our Faster-RCNN. It should be noted that our model was only trained on 6 videos and tested on 14 videos, and no details of training and testing procedures are available for the DPBOD model. As experiment *O1* suggests, our Faster-RCNN detector should perform better if more training data were used.

In experiment *H1*, where no initially pre-trained model was used for the weights, our hand detector failed to detect any hands at all, yielding a mAP of zero! However, in experiment *H2*, where weights which resulted from training a VGG16 image classification model on 1.2 billion images from Imagenet's Image Classification Dataset were used, an overall mAP value of 94.0% was obtained. This compares favorably with the corresponding result from [7] – the paper in which the EgoHands dataset was first introduced – where the mAP only reached 80.7%.

V Conclusions and Future Work

In this paper we have presented, implemented and tested a novel approach, based on a Region-Based Convolutional Neural Network, for detection of objects in general, and human hands in particular, in egocentric videos, with a view to applying these in an Ambient Assisted Living context. Although the performance of our model for the detection of objects is disappointing, for the specialised task of detecting human hands, our model is quite successful, outperforming that of [7], which is considered the current state of the art. As a consequence, such model provides reliable information on which higher level reasoning can be applied.

We plan to extend this work by applying our hand detector to the task of recognizing activities of daily living (ADL) being carried out in egocentric videos, with a view to building a complete Ambient Assisted Living system, using egocentric videos, i.e. videos obtained from the point of view of the person requiring ambient assistance.

Acknowledgment

T.H.C.N. is grateful to the Faculty of Science, Engineering and Computing at Kingston University for financial support whilst working on this project.

References

[1] Index Mundi (2018) "Japan – Demographic Profile" https://www.indexmundi.com/japan/demographics_profile.html

[2] S. Ren, K. He, R. Girshick & J. Sun (2015). "Faster R-CNN: Towards real-time object detection with region proposal networks", in *Advances in Neural Information Processing Systems* (*NIPS 2015*) pp 91-99

[3] T.H.C. Nguyen, J.-C. Nebel & F. Florez Revuelta (2016) "Recognition of Activities of Daily Living with Egocentric Vision: A Review", Sensors, 16, 72; doi:10.3390/s16010072

[4] R. Girshick, (2015) "Fast R-CNN", *Proceedings of IEEE International Conference on Computer Vision* (*ICCV*), pp 1440-1448, doi : 10.1109/ICCV.2015.169

[5] K. Simonyan & A. Zisserman (2014) "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556.

[6] IMAGENET Large Scale Visual Recognition Challenge (2015) - http://image-net.org/challenges/LSVRC/2015/

[7] S. Bambach, S. Lee, D. Crandall & C. Yu (2015) "Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions", *Proceedings of IEEE International Conference on Computer Vision (ICCV), pp 1949-1957*

[8] H. Pirsiavash & D. Ramanan (2012) "Detecting activities of daily living in first-person camera views", *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Providence, Rhode Island, USA, 16–21 June 2012; pp. 2847–2854.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, et al. (2015) "Imagenet large scale visual recognition challenge", Int. J. Comput. Vision, 115, 211–252

[10] IMAGENET Large Scale Visual Recognition Challenge (2017) - http://image-net.org/challenges/LSVRC/2017/

[11] P. Sermanet, P., D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, (2013) "Overfeat: Integrated recognition, localization and detection using convolutional networks", arXiv preprint 2013, arXiv:1312.6229

[12] M. Oquab, L. Bottou, I. Laptev, J. Sivic (2015) "Is object localization for free ? Weakly-supervised learning with convolutional neural networks", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA, June; pp. 685–694.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. (2016) "Region-based convolutional networks for accurate object detection and segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 38, No. 1, pp 142 - 158

[14] J. Uijlings, K. van de Sande, T. Gevers, A. Smeulders (2013) "Selective search for object recognition", International Journal of Computer Vision, Vol. 104, no. 3, pp. 154-171

[15] M. Everingham, L. Van-Gool, C.K.I. Williams, J. Winn, & A. Zisserman (2007) "PASCAL Visual Object Classes Challenge 2007 (VOC2007)", http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[16] M. Everingham, L. Van-Gool, C.K.I. Williams, J. Winn, & A. Zisserman (2010) "PASCAL Visual Object Classes Challenge", International Journal of Computer Vision, Vol. 88, pp. 303–338

[17] W. Yao & Z. Deng (2012) "A Robust Pedestrian Detection Approach Based on Shapelet Feature and Haar Detector Ensembles", Tsinghua Science and Technology, Vol. 17, No. 1, pp 40-50, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6151906

[18] N. Dalal & B. Triggs (2005) "Histograms of oriented gradients for human detection", *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR),* pp. 886–893

[19] S. Paisitkriangkrai, C. Shen & A. van den Hengel (2013) "Efficient pedestrian detection by directly optimizing the partial area under the ROC curve", *Proceedings of International Conference in Computer Vision (ICCV) 2013*, Sydney, Australia, https://arxiv.org/pdf/1310.0900.pdf

[20] P.F. Felzenszwalb, R.B. Girshick, D. McAllester & D. Ramanan (2010) "Object detection with discriminatively trained part-based models", IEEE Transactions on Pattern Anal. Mach. Intell. 2010, 32, 1627–1645

[21] P. Felzenszwalb, D. McAllester, & D. Ramanan (2008) "A discriminatively trained, multiscale, deformable part model", *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Anchorage, Alaska, USA, June 2008; pp. 1–8.

[22] K. Matsuo, K. Yamada, S. Ueno & S. Naito (2014) "An attention-based activity recognition for egocentric video", *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Columbus, Ohio, USA, June 2014; pp. 565–570.