Recognition of activities of daily living from egocentric videos using hands detected by a deep convolutional network

Thi-Hoa-Cuc Nguyen[1], Jean-Christophe Nebel[1], and Francisco Florez-Revuelta[2]

[1] School of Computer Science & Mathematics,
Kingston University, London, Kingston-Upon-Thames, KT1 2EE, UK
Email: {k1458932, j.nebel}@kingston.ac.uk
[2] Department of Computer Technology, University of Alicante,
P.O. Box 99, E-03080 Alicante, Spain
Email: francisco.florez@ua.es

Abstract. Ambient assisted living systems aim at supporting older and impaired people using computer technology so that they can remain autonomous while maintaining healthy living. Egocentric cameras have emerged as a powerful source of data to monitor individuals performing activities of daily living since they tend to focus on the area where the current activity takes place while showing manipulated objects and hand positions. While research has focused on activity recognition based on object recognition, this study has investigated the automatic acquisition of additional features modelling interactions between hands and objects using a deep convolutional network. Experiments conducted on a realistic dataset have demonstrated that, not only do those features improve activity recognition, but they can be accurately extracted.

Keywords: egocentric vision; deep convolutional network; ambient assisted living; activity recognition

1 Introduction

Ambient assisted living (AAL) systems aim at supporting older and impaired people using computer technology so that they can remain autonomous while maintaining healthy living. Many AAL systems are deployed in homes where they monitor individuals performing activities of daily living (ADLs), such as cooking or washing. Although those system employ different sensors, usage of cameras is becoming more popular [1] especially that automatic solutions ensuring individuals privacy are now available [2]. While traditionally video-based AAL systems rely on fixed cameras installed in the environment, the increasing availability of wearable cameras addresses many of their shortcomings, such as their limited field of view and the occlusions created by either the environment or an individuals body. Conversely, a camera attached to a person can record activities from an egocentric point of view, showing usually manipulated objects

and hand positions. As a consequence, ADL recognition from data captured by egocentric cameras has become an active field of research [3], where current systems generally rely on, first, detecting and identifying the objects present in each frame and, second, using temporal information to infer the performed activities.

With the remarkable progress of deep learning and Convolutional Neural Networks (CNNs) in particular [4], significant improvement has been achieved in image recognition tasks [5]. However, even the Imagenet dataset, and its 15 million annotated images, is still far from being sufficient to train a CNN able to handle a realistic AAL video. Therefore, one can expect that recognition of ADLs relying on object identity as main information source will only progress substantially with the availability of more exhaustive object datasets. However, experiments conducted with videos, the objects of which were manually annotated, have revealed that additional information such as indication about active objects is required to perform accurate recognition of ADLs [6].

This work reports research regarding the value of automatic extraction of additional features related to hand-hand and hand-object interactions in the context of ADL recognition. First, a CNN-based method is proposed to detect the left and right hands visible on a video captured by an egocentric camera. Second, using a bag-of-visual words model, a system for recognition of activities of daily living is developed taking advantage of both object identity and interaction features. Finally, experiments are conducted on an AAL dataset captured by a wearable camera to evaluate the impact of exploiting those new features.

2 Methodology

2.1 Hand detection using Faster-RCNN with VGG-16

As Faster-RCNN is a freely available and efficient object detection framework [5], and human hands can be treated as objects, this model has been selected to detect both left and right hands in videos. Faster-RCNN operates in two steps: first, a deep fully CNN, VGG-16 [4], proposes rectangular regions of interest associated to an objectness score; second, the Fast R-CNN detector [7] takes advantage of those regions to detect and classify objects.

To detect left and right hands, the last layer of the Faster-RCNN was adapted to output left-hand and right-hand object classes using, for training, a large dataset of hands in egocentric videos. The system returns bounding boxes tagged as either left or right hand associated with a confidence score. Bounding boxes with a score above a threshold are further processed to filter out noisy results which may be produced by blurred data caused by fast hand motion: a hand detection is accepted only if it has been continuously detected for some frames.

2.2 Activity recognition enhanced by hand information

The bag-of-visual words (BoVW) model has been a popular approach applied for image classification by treating an image as a document and features as words [8].

A BoVW is a vector representing occurrences of vocabularies which are generated from local image features. BoVW is mainly composed of five steps: (i) feature extraction, (ii) feature pre-processing, (iii) codebook generation, (iv) feature encoding, and (v) pooling and normalisation [9]. Motivated by the idea that an activity can be modelled by using relevant images (key features) in a video, the BoVW model is used to generate bags of key features and then combined with the Sequences of key features approach to recognise activities [10][2].

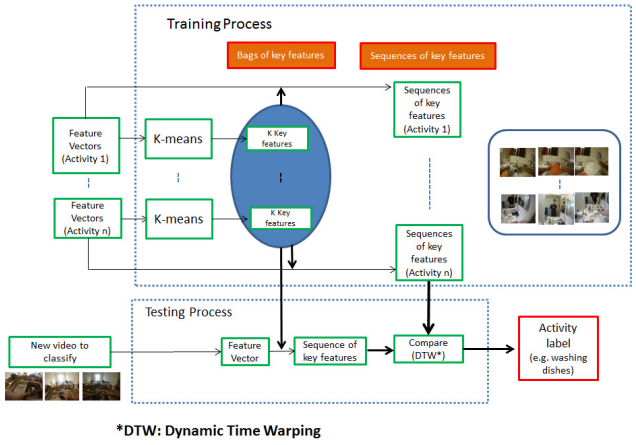

Fig. 1. Sequences of key interactions for activity recognition.

As Figure 1 shows, during the training process, bags of key features and sequences of key features for each activity are built. First, all feature vectors extracted from all videos of each activity class are fed to a k-means clusteringto obtain K representative vectors for each activity. Then, these vectors, for all the activities, are merged together into a bag of key features. Second, for each frame of a sequence, the nearest neighbour key feature in the bag is found. The successive nearest neighbour key features constitute the sequence of key features. This way, a set of sequences of key features is obtained for each activity class.

To classify an unseen video into a specific activity, three steps are followed (Figure 1). First, a feature vector is extracted for each frame of the new video. Then, vectors are mapped to the key features generated during the training process. Second, the corresponding sequence of nearest neighbour key features is built. Third, Dynamic Time Warping (DTW) is applied to compare the new sequence with every sequence of key features created during the training. The label of the nearest neighbour sequence supplies the label of the new video.

The presence of specific objects provides rich information about performed activities. However, it is proposed to provide further information by including features exploiting spatial relationships between both, objects and hands, and left and right hands. It is expected that such data will allow modelling inter-

actions occurring in each activity. Then, motivated by the idea that ADLs can be represented as a series of key interactions between hands and objects, the sequences of key features approach is applied to predict activities.

Egocentric videos of ADLs allow analysis of two main types of interactions. First, since hands are generally present in the centre of the images, they should provide useful information about the activity that is performed: left hand-right hand interactions are usually necessary to complete activities. Second, objects are usually manipulated by hands when activities are performed. Therefore, information interactions between hands and objects would allow not only the detection of active objects, but also discard objects which are irrelevant to the current activity. Since any object that is manipulated should eventually be in contact with a hand, the positions of both hands are used to identify active objects in a scene. Here, the active object associated to the left hand, resp. right hand, is defined as the closest object to its left, resp. right.

In this study, three features are extracted from each frame (see Figure 2): the distance in pixel between the left and right hands (*d1*), i.e. the distance between the centres of their bounding boxes, and the distances between each hand and its active object (*d2* and *d3*), i.e. the distances between the centre of each hands bounding box to the centre of its active objects bounding box.

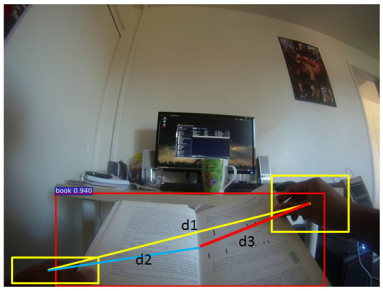

Fig. 2. Example of interaction features. *d1*: distance between left and right hands; *d2*: distance between left hand and book; *d3*: distance between right hand and book.

3 Experimental results

3.1 Experimental setup

Hand detection The Faster-RCNN model with VGG-16 network with two classes (left/right hand) was implemented using Caffe, the popular deep learning framework developed by Berkeley AI Research (BAIR) [11], its Pycaffe library supporting the Python version of Faster-RCNN [5]. A single GPU NVIDIA Titan X 12GB GDDR5X (memory speed of 10Gbps) and CUDA were used as high performance processing platform to train this very deep convolutional network.

Regarding training, the EgoHands dataset [94] which contains 48 videos of first-person interactions between two people recorded by a Google Glass was selected. Indeed, it provides 15,053 annotated (left/right) and segmented hands in a total of 4,800 images. 480,000 iterations were conducted as part of the training process: they correspond to 100 forward and backward passes of all the training examples (so called epochs) as the dataset contains 4,800 images and a batch-size of one image by batch was set. The weights of the network were tuned by Caffe during the back propagation phase. As weight pre-training is a procedure known for reducing computational time and producing improved performance, it was applied taking advantage of a VGG16 image classification model already trained using the Imagenets Image Classification Dataset. Finally, the outputs generated by the network were post processed considering only detections with a confidence above 0.8 if they appear continuously on at least 5 frames.

The evaluation of the hand detector was conducted using 75% of EgoHands as training set and the remaining 25% as testing set. Performance was assessed using the standard mean Average Precision (mAP) which is algorithm-independent and specially adapted at highlighting differences between methods [12].

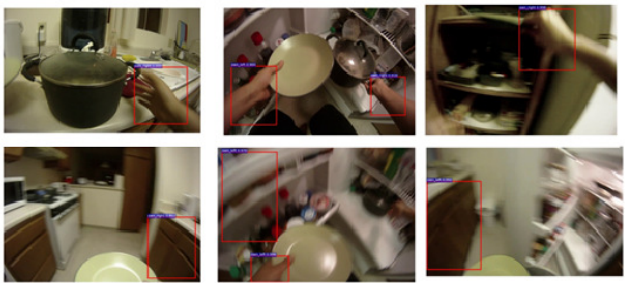

Fig. 3. Outputs of the hand detector: correct (top) and incorrect (bottom) detections.

Activity recognition The most suitable dataset available to evaluate recognition of ADLs with wearable cameras is the Activities of Daily Living Dataset (ADL) [6], which comprises 1 million frames recorded by a GoPro camera. Not only does it provide annotation for 18 activities of daily living, such as *washing dishes* and *combing hair*, but objects visible in each frame are annotated within 32 distinct object categories. Even though methods for object detection have improved, performance is still relatively low: only an mAP of 0.66 was recently achieved on the large Imagenet dataset [13]. In the case of the challenging ADL dataset, where blurred and dark images are not uncommon, there is not even any suitable set for training more than a third of the objects, e.g. dental-floss, detergent and clothes. As a consequence, automatic object detection on the ADL dataset is still extremely inaccurate ($mAP < 0.2$) and, as many have done before [6], actual ground-truth object annotations are used for activity recognition.

In order to demonstrate the added value provided by the inclusion of hand-hand and hand-object interaction information, four experiments were conducted where different feature vectors were fed to the sequence of key features classifier. First, as a baseline, each frame was represented by a feature vector containing 32 binary values encoding the presence of objects belonging to the 32 categories in the ADL dataset. Second, a feature vector was created containing only the spatial distance in pixels between the left and right hands (*d1*). Third, distances in pixels between each hand and its associated active objects (*d2*) and (*d3*) were added to the previous vector (*d1*). Finally, the presence of objects belonging to the 32 categories was added to distance information, i.e. *d1*, *d2* and *d3*.

For each of those experiments, a sequence of key interactions model was trained to recognise the 18 types of activities present in the 20 videos of the ADL dataset. Evaluation was performed from pre-segmented activities using the twofold cross-validation method calculating average accuracy.

3.2 Performance

Hand detection The trained Faster-RCNN with VGG-16 network performs extremely well as evidenced by a mean Average Precision of 0.940. This corresponds to an improvement of over 16% when compared with the mAP of 0.807 reported at the introduction of the EgoHands dataset. Although there are rare occasions when the system produces misdetections, especially in blurry images (see Figure 3, bottom row), such performance suggests that this classifier should be able to provide valuable information to an activity recognition system.

Activity recognition Average accuracy of activity recognition using various feature vectors is reported in Table 1. Performance for the experiment conducted with the second feature vector reveals that, even without any object information, the distance between the two hands is quite informative to identify an activity among the 18 possible types (prediction with 21.0% accuracy). This is an interesting result since object detection and annotation is particularly challenging on realistic ADL datasets. If additional information about interaction with active objects is provided, accuracy increases to 23.5%. Still, comparison with baseline performance shows that object annotations are even more discriminative than knowledge about interactions (26.3%). Finally, combination of the two types of features provides richer information for ADL recognition since it allows increasing annotation-based performance by 12.5% leading to an accuracy of 29.6%.

Figure 4 shows the confusion matrix obtained for the experiment integrating all features. It reveals that relatively good results are obtained for a few activities including *drying hands* (74.2%) and *laundry* (50%). However, they are mediocre for most activities, e.g. *combing hair* (12.5%), *making cold food or snack* (11.8%) and *brushing teeth* (18.1%). Moreover, the model could not detect activities involving small objects such as *dental floss*, leading to not a single recognition (0%). The main reason for these unsatisfactory performance is the lack of training data for many activities such as *make up* and *using cell-phone*. In addition, some

Table 1. Average accuracy of activity recognition using various feature vectors

Feature vector	**Average accuracy (%)**
Object presence (baseline)	26.3
Distance between left and right hands (*d1*)	21.0
Distances between left and right hands (*d1*), and each hand and its active object (*d2* & *d3*)	23.5
Object presence and distances (*d1*, *d2* & *d3*)	29.6

ADLs are very difficult to discriminate since they involve similar objects and interactions, e.g. *moving dishes* and *washing dishes*. Finally, the ADL dataset includes simultaneous activities such as *drinking* and *watching TV*, the concept of which is not addressed by the proposed framework.

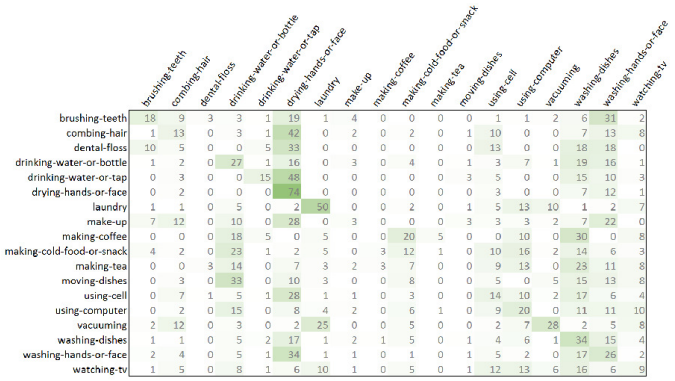

Fig. 4. Confusion matrix associated to classification of 18 activities based on both object presence and interactions.

Usage of a more sophisticated classifier based on a spatial pyramid to approximate temporal correspondence achieved on the ADL dataset an average accuracy of 55.8% on pre-segmented activities when using the ground truth for object annotation [6]. In addition, if the ground-truth for active objects is added, the systems performance increases to 77%. This is in line with the outcome of our study which suggests that, if the proposed approach for automatic hand-hand and hand-object interactions were to be employed, performance of classifiers relying only on object identification could be substantially improved.

4 Conclusion

Usage of egocentric cameras to recognise activities of daily living has the potential to become essential to ambient assisted living. However, research so far

has mainly relied on the recognition of objects visible in each video frame. Despite significant progress brought by deep learning, object recognition is still a challenge for realistic ADL datasets. This study has demonstrated that, not only does the addition of features modelling interactions between hands and objects improve activity recognition, but they can be automatically and accurately extracted using a deep network. Although the outcome of activity recognition remains unsatisfactory, exploitation of those features within state-of-the-art activity recognition systems should impact significantly on their performance.

References

1. Cardinaux, F., Bhowmik, D., Abhayaratne, C., Hawley, M.S.: Video based technology for ambient assisted living: A review of the literature. Journal of Ambient Intelligence and Smart Environments **3**(3) (2011) 253–269
2. Chaaraoui, A.A., Padilla-López, J.R., Ferrández-Pastor, F.J., Nieto-Hidalgo, M., Flórez-Revuelta, F.: A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. Sensors **14**(5) (2014) 8895–8925
3. Nguyen, T.H.C., Nebel, J.C., Florez-Revuelta, F., et al.: Recognition of activities of daily living with egocentric vision: a review. Sensors **16**(1) (2016) 72
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
5. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6) (June 2017) 1137–1149
6. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. (June 2012) 2847–2854
7. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision. (2015) 1440–1448
8. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. Volume 1., Prague (2004) 1–2
9. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding **150**(Supplement C) (2016) 109–125
10. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters **34**(15) (2013) 1799–1807 Smart Approaches for Human Action Recognition.
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 675–678
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision **88**(2) (Jun 2010) 303–338
13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3) (Dec 2015) 211–252