Action Detection from Egocentric Videos in Daily Living Scenarios

G. Kapidis[1, 2], E. van Dam[1], R. Poppe[2], L. P. J. J. Noldus[1], R. C. Veltkamp[2]

1 Noldus Information Technology, Wageningen, The Netherlands; 2 Department of Informatics and Computer Science, University of Utrecht, Utrecht, The Netherlands
{georgios.kapidis, elsbeth.vandam, lucas.noldus}@noldus.nl, {r.w.poppe, r.c.veltkamp}@uu.nl

Introduction

Human Action Recognition (HAR) as a computer vision subfield has seen impressive progress due to the growth of deep learning based methods. The field of egocentric vision has received significant attention from the activity recognition community as the amount of relevant work indicates ([1]–[8]). Recent detection methods rely on neural network structures to describe the content of video frames as well as the sequential relations in terms of objects [2], [7], activities [1], [3], [4], [6], locations [5], [7], interactions [8] or combinations of different scene aspects. In the domain of activities of daily living and specifically in that of older adult healthcare, egocentric vision is becoming acknowledged as a means of activity analysis and understanding [9]. A nice review for egocentric activity recognition can be found in [10].

In this work, we focus on HAR from an egocentric video perspective for the analysis of indoor scenes and activities. We are using Long Short-term Memory (LSTM)[11] a deep learning method designed for sequential data that allows us to analyze videos in the temporal domain, efficiently making use of the inherent structure that exists in a sequence of video frames. First experiments for distinguishing between standing/sitting/no-action classes in the Activities of Daily Living (ADL) dataset [12] reveal promising results towards classification of the performed actions of the camera wearer.

Methodology

We are following the example of [5] to produce a dataset of objects that can be classified into activity groups for each video frame. We do this by applying a state-of-the-art object detector on the videos comprising the ADL dataset, saving the detections and using them to produce train and test splits for the LSTM training scheme.
More specifically, we use the YOLO (You Only Look Once) object detector [13] trained to detect the ADL20 set of objects from [5] which consists of typical objects found in an indoor scene e.g. fridge, oven, television, door. To train the LSTM we process the object detections per frame, into a binary vector with length the number of available objects (20 in this case) with ones in the indices of detected objects for the current frame and zeros otherwise. Since the LSTM is able to model the temporal dynamics when having a sequence as input, we take advantage by using sets of five frames. Furthermore, we change the frame rate to a single frame per second (1 fps) with the effect of enlarging the period covered from one training example from 33 milliseconds to 1000, without complicating the input by having a very long sequence. For our training scheme, a single LSTM layer is sufficient to model the three proposed action groups. The action groups are the result of mapping the action classes proposed in [12]. The activity mapping is shown in Table 3.

Table 3: ADL action classes mapped to standing and sitting action groups with the addition of the background class for the frames that have been annotated as having no occuring activity. In the 3rd and 4th row the frames for each class in the train and test sets, respectively.

Group	No-action	Standing	Sitting
Activity	Background	Combing hair, make up, brushing teeth, dental floss, washing hands/face, drying hands/face, enter/leave room, adjusting thermostat, laundry, washing dishes, moving dishes, making tea, making coffee, drinking water/bottle, drinking water/tap, making hot food, making cold food/snack, mopping in kitchen, vacuuming, taking pills, making bed, cleaning house, using mouth wash, grabbing water from tap	Eating in kitchen, watching tv, using computer, using cell, reading book, writing, putting on shoes/socks, drinking coffee/tea

R.A Grant et al. (Eds.): Measuring Behavior 2018, ISBN 978-1-910029-39-8
Manchester, UK, 5th-8th June 2018

Train frames	1208	6701	3734
Test frames	2890	10338	7791

Experiments, Results and Discussion

After experimenting with the training hyperparameters of the LSTM, we discovered the optimal values for the number of training iterations, learning rate, sequence size, mini-batch size and dropout as shown in Table 4. The network consists of one LSTM layer, followed by a fully connected layer without a non-linear activation so it can be used for inference. The loss function is weighted softmax cross entropy, with Adam [14] for optimization. The weighting of each class is an additional hyperparameter to account for the imbalance of the action groups in the train set.

Table 4: LSTM hyperparameters and architecture

Training steps (epochs)	1500 (~62)
Learning rate	10^{-4} to 10^{-6} in 100 steps with polynomial decay
Sequence size	5
Mini-batch size	96
Dropout	0.85
LSTM layers	1
Hidden units	20

Our single best execution in terms of overall accuracy has 73.08% of the frames in the test set classified correctly. It is important to note the imbalance of the test set as shown at the 4th row of Table 3. Per class results are shown in Table 5. The low recall in the no-action case does not affect the overall accuracy significantly, because the number of test samples is low compared to the other two classes.

Table 5: Per class recall and precision results, max is 1.

	No-action	Standing	Sitting
Recall	0.01	0.92	0.75
Precision	0.11	0.71	0.77

The preliminary results show that it is possible to classify the frames when a person is ‘standing’ versus when ‘sitting’ successfully based on the underlying objects and activities; however, the ‘no-action’ frames are easily mistaken as one of the other groups, due to the aforementioned class imbalance. Another reason is the objects that signify each action group and the background. The background does not depend on specific objects to be detected, rather on random occurrences of objects, which are characteristic for other actions. A ‘fridge’ and an ‘oven’ found in a ‘no-action’ frame, would suggest a food related action at test time, which belongs in the ‘standing’ group because the number of training samples with this label is greater and as such, distinctive of the ‘standing’ action group.

In future work we intend to study the objects that individually describe each activity and action group in order to identify which object classes (and lack thereof) represent specific actions.

Acknowledgement

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676157.

References

[1] M. Ma, H. Fan, and K. M. Kitani, “Going deeper into first-person activity recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1894–1903.

[2] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, “First Person Action-Object Detection with EgoNet,” in *Proceedings of Robotics: Science and Systems*, 2017.
[3] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, “Jointly Learning Energy Expenditures and Activities Using Egocentric Multimodal Signals,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6817–6826.
[4] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, “Compact CNN for indexing egocentric videos,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.
[5] G. Kapidis, R. Poppe, E. van Dam, L. Noldus, and R. Veltkamp, “Where Am I? Comparing CNN and LSTM for Location Classification in Egocentric Videos,” in *SmarterAAL’18 Workshop, IEEE International Conference on Pervasive Computing and Communication (PerCom)*, Athens, Greece, 2018.
[6] G. Vaca-Castano, S. Das, and J. P. Sousa, “Improving egocentric vision of daily activities,” in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 2562–2566.
[7] G. Vaca-Castano, S. Das, J. P. Sousa, N. D. Lobo, and M. Shah, “Improved scene identification and object detection on egocentric vision of daily activities,” *Comput. Vis. Image Underst.*, vol. 156, pp. 92–103, Mar. 2017.
[8] R. Yonetani, K. M. Kitani, and Y. Sato, “Recognizing micro-actions and reactions from paired egocentric videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2629–2638.
[9] S. Karaman, J. Benois-Pineau, R. Mégret, V. Dovgalecs, J.-F. Dartigues, and Y. Gaëstel, “Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases,” in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 4113–4116.
[10] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, “Recognition of Activities of Daily Living with Egocentric Vision: A Review,” *Sensors*, vol. 16, no. 1, p. 72, Jan. 2016.
[11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[12] H. Pirsiavash and D. Ramanan, “Detecting activities of daily living in first-person camera views,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2847–2854.
[13] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.
[14] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *ArXiv14126980 Cs*, Dec. 2014.