Detecting hands in egocentric videos: towards action recognition

Alejandro Cartas *[1], Mariella Dimiccoli[2], and Petia Radeva[3]

[1,2,3]University of Barcelona and Computer Vision Centre, Barcelona, Spain

With the advances on wearable technologies in recent years, there has been a growing interest in analyzing human daily activities from data collected by wearable cameras [4, 9, 6]. Daily activities are crucial to characterize human behavior, and enabling their automatic recognition would pave the road to novel applications in the field of Preventive Medicine.

The hands are involved in a wide variety of daily tasks, such as typing on a self-phone keyboard, drinking coffee or riding a bike (see Fig. 1). Along with the objects being manipulated in a scene, the hands are often the main focus in the egocentric field of view. Consequently, their detection is a fundamental step towards action recognition. However, detecting hands in egocentric images is not a trivial task for three main reasons. First, the hands are intrinsically non-rigid and their shape appearance change continuously while manipulating objects. Second, the illumination conditions rapidly change in egocentric images as a consequence of the camera user movements across different locations. These changes also affect the appearance of the hands and their recognition, as stated by Li and Kitani [5]. Third, the complexity of the method also depends on the camera used and its position on the body (head, shoulders, or chest). For instance, if the camera is worn on the chest, the focus of attention is lost and the location of hands in the field of view becomes more unpredictable. Available methods for detecting hands in egocentric images [2, 5, 8] are mostly based on hand-crafted features such as color histogram, texture and HOG in different color spaces.

In this work, we propose a hand detector that exploits skin modeling for fast hand proposal generation and Convolutional Neural Networks for hand recognition. We tested our method on UNIGE-HANDS dataset [1] and we show that the proposed approach achieves competitive hand detection results.

1. Hand detection

Our hand detector consists in a three-task architecture outlined in Fig. 2. We first detect regions containing skin pixels. Later, we generate a set of hand proposals using these regions. Finally, we classify the hand proposals using a Convolutional Neural Network (CNN).

*Corresponding Author: alejandro.cartas@gmail.com.

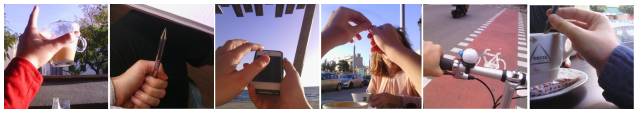


Figure 1. Hand-actions captured by a wearable camera.

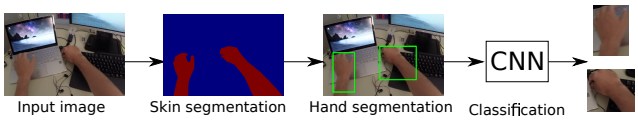
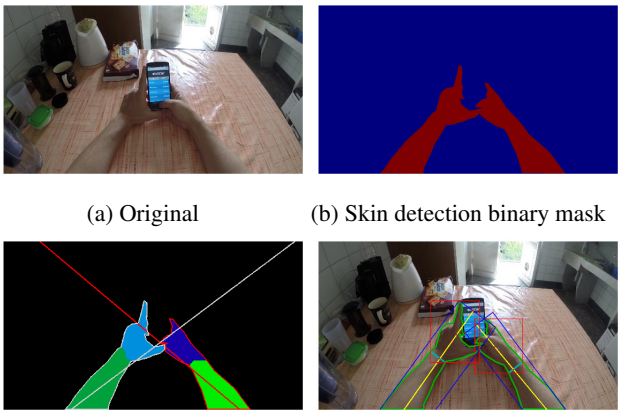


Figure 2. Outline of the proposed method for hand detection.



(a) Original (b) Skin detection binary mask

(c) Arm segmentation (d) Hand contour cut

Figure 3. Example of a hand proposal generation over a skin region containing two pixel connected arms.

1.1. Skin detection

For this task, we use the pixel-level skin detection (PERPIX) method introduced in [5]. We made a performance comparison on the UNIGEN dataset with other two methods originally designed for it [1], as seen on Table 1. Besides being available for use, the PERPIX method offers competitive results using less than the 10 percent of the training data used in [1].

1.2. Hand proposal generation

To generate hand proposals we first determine if the estimated skin-region contains two arms. This is achieved by fitting a straight line using the points from the boundary of the skin region. If the mean squared error of the fit is over a threshold, then the skin-region is considered as a two-arms region. A two-arms region is segmented into two by first fitting the points from its medial-axis into two lines that

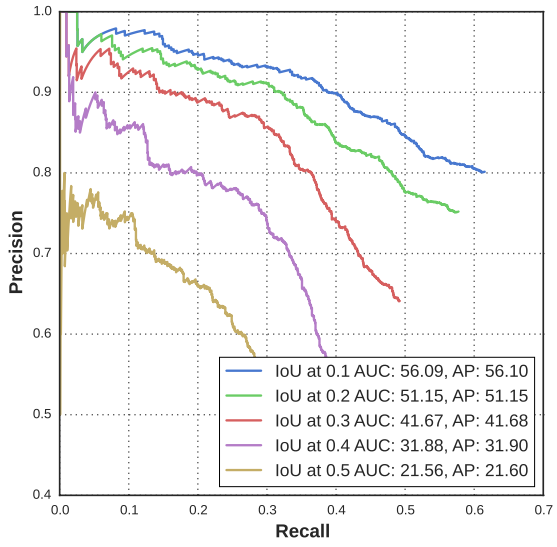


Figure 4. Detection results on the UNIGEN test set for different values of the intersection over union (IoU) ratios.

should represent each arm. To make a soft segmentation, we obtain small regions using the watershed algorithm. Finally, each small region is assigned to the closest line of the two (see Fig. 3c). Finally, we separate the hands from each one-arm regions by fitting a rectangular box from its contour. The side of the box closer to the center of the frame is considered to be the nearest to the hand. Consequently, we obtain the hand proposals by cutting the contour at different fixed distances from that side, see Fig. 3d.

1.3. Hand recognition

To classify a hand proposal we created a binary classifier by fine-tuning the CaffeNet network [3] pre-trained on ImageNet. Our training set was obtained by combining several datasets with bounding boxes of hands to obtain positive examples, and ImageNet [7] to obtain negative examples. The total number of images and bounding boxes is 761,946 and 872,414, respectively.

2. Experimental results

We evaluated our method over a subset of 2,000 images from the UNIGE-HANDS dataset [1] captured by a GoPro camera. Since this dataset comes with a label per image indicating if it contain skin or not, we manually annotated hands to allow a numerical evaluation. The number of images containing hands were 1,000 and in total they were over 1,739 hands.Fig. 4 shows precision-recall curves for four distinct values of intersection over union. We obtain an average precision (AP) of 0.216 when using the PASCAL VOC criteria.

	True Positive			True Negative		
	HOG-SVM	DBN	PERPIX	HOG-SVM	DBN	PERPIX
Office	0.893	**0.965**	0.962	0.929	0.952	**0.987**
Street	0.756	0.834	**0.897**	0.867	**0.898**	0.528
Bench	0.765	0.882	**0.925**	0.965	**0.979**	0.925
Kitchen	0.627	0.606	**0.709**	0.777	**0.848**	0.823
Coffee bar	0.817	**0.874**	0.641	0.653	0.660	**0.867**
Total	0.764	**0.820**	0.809	0.837	**0.864**	0.843

Table 1. Skin-segmentation performance comparison. The HOG-SVM and DBN results correspond to [1] and the PERPIX results were obtained using the Per-pixel regression method [5].

References

[1] A. Betancourt, M. Lopez, C. S. Regazzoni, and M. Rauterberg. A Sequential Classifier for Hand Detection in the Framework of Egocentric Vision. In *Conference on Computer Vision and Pattern Recognition*, volume 1, Columbus, Ohio, 2014. IEEE Computer Society.

[2] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*, pages 3281–3288. IEEE, 2011.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] 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 *Pattern Recognition (ICPR), 20th International Conference on*, pages 4113–4116. IEEE, 2010.

[5] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577. IEEE, 2013.

[6] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding Everyday Hands in Action from RGB-D Images. In *IEEE International Conference on Computer Vision*, Santiago, Chile, Dec. 2015.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[8] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara. Hand segmentation for gesture recognition in egovision. In *Proceedings of the 3rd ACM International Workshop on Interactive Multimedia on Mobile Portable Devices*, IMMPD, pages 31–36, New York, NY, USA, 2013. ACM.

[9] J. Zariffa and M. R. Popovic. Hand contour detection in wearable camera video using an adaptive histogram region of interest. *Journal of NeuroEngineering and Rehabilitation*, 10(1):1–10, 2013.