All Hands on Desk
Hand Segmentation and Activity Localization in EgoHands Dataset

Final Project Report

Chuhua Wang Sven Bambach David J. Crandall
School of Informatics, Computing, and Engineering
Indiana University
{cw234,sbambach,djcran}@indiana.edu

1. Introduction

Wearable cameras, such as Google Glass, GoPro, and Narrative Clip [1], have been developed to record daily-life egocentric videos. A large number of the contents in egocentric videos include interactions between users, which generates a vast amount of data. Unlike third-person videos (*e.g*. surveillance videos or movie scenes), in which cameras are usually stabilized, egocentric videos contain unstable camera motion, unusual viewpoints, poor scene composition and constantly changing illuminations. Identifying human actions in egocentric videos plays essential roles in multiple applications, such as robotic control, augmented reality(AR) and surveillance. However, most of the works focus on human body recognition and activity recognition, but not on hand configurations.

Because we use our hands regularly to interact with others, hand poses and trajectories contain a wealth of information. For example, spatial information and temporal characteristics can be delivered when pointing at an object using a finger [18]. In this sense, physical movement of the hands intent to convey information on the activities of users [15]. To understand interactions between users in egocentric videos, accurately extracting hands is the first and critical step to locate activities. Therefore, hand detection and segmentation are the fundamental tasks.

Many approaches have been devoted to accomplish hand detection and segmentation, including probabilistic models [11], mathematical models [22, 3], or deep learning models [1, 10]. Current methods used to analyze these data may be improved by increasing the accuracy. Therefore, we proposed this project to analyze data of hand configuration by using an updated method, which could help us better understand the scenes, situations, and interactions between users in egocentric videos.

This project is based on the Bambach's *et al*. work [1], in which they presented a deep learning model for hand detection, segmentation, and activity recognition. In this project, We used the state-of-the-art Mask R-CNN [8] for hand detection and pixel-wise segmentation between different hand types: my left hand, my right hand, other's left hand, and other's right hand in EgoHands dataset. In addition to hand recognition, the same model was also used for activity localization based on given segmentation or segmentations.

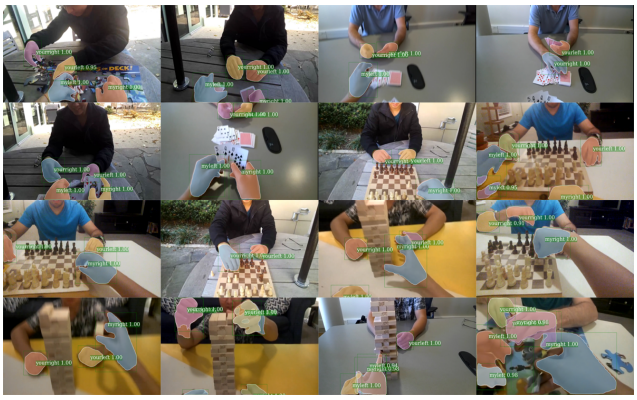

Figure 1. *Examples of hand detection and segmentation.*

This project is a joint project of my independent study, and the project is completed by myself, with the help of Dr. Sven Bambach, and under the supervision of Professor David Crandall.

2. Background and related work

2.1. Mask R-CNN

Mask R-CNN is the current state-of-the-art segmentation algorithm. It was proposed by Ka *et al*. [8], which extends Faster R-CNN [19] by adding a paralleled branch in addition to object detection and classification. In order to solve

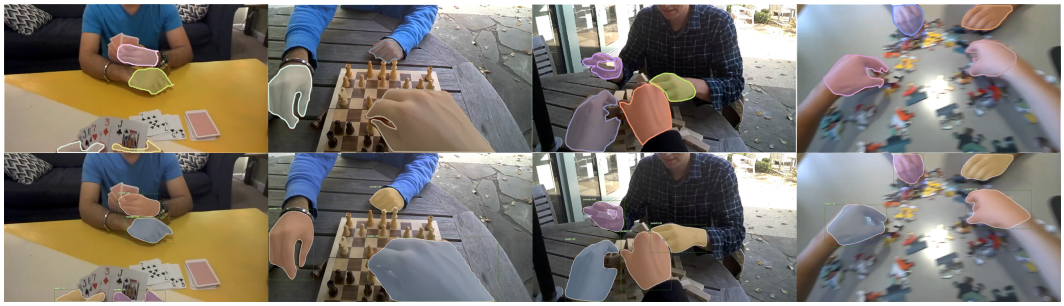

Figure 2. *Visualizations of EgoHands dataset. First row:* Ground truth hand segmentation masks of selected frames from the dataset, where colors indicate the different hand types. *Second row:* Predicted results of corresponding frames. Masks are shown in different color. The bounding boxes are also shown in green color. Category and confidences are shown on the top left of the bounding boxes.

the misalignment caused by RoIPool, they also updated the RoIPool method by using bilinear interpolation to estimate pixel values.

In this project, we used Mask R-CNN to locate and distinguish between hands in egocentric videos as well as to locate different activities based on hand segmentation(s). The Mask R-CNN implementation from Detecton [7] was adapted in this project.

2.2. Hand detection and segmentation

Previously, Ren and Gu [20] used motion to separate the moving hands and objects from the background and showed that the segmentation improved the accuracy of object recognition systems assuming the background is static. In 2013, Li and Kitani *et al*. [12] classified hand detection approaches into three categories: local appearance-based detection, global appearance-based detection and motion-based detection. Their pixel-level detection approach showed success in a wide range of illumination changes and hand deformations. However, in their research they assumed the videos contained the egocentric viewer only. Lee *et al*. [11], proposed a probabilistic model to detect and disambiguate hands type in egocentric videos. However, the experiment only took place in a laboratory setting. Bambach *et al*. [1], proposed a skin-based approach that first generates a set of bounding boxes that may contain hand regions used Convolutional Neural Networks (CNN) to locate and distinguish between hands as well as different activities in egocentric video. They also introduced a first-person dataset with dynamic interactions between people, along with fine-grained ground truth. Deng *et al*. [5], used Faster R-CNN [19] backbone and a multi-component SVM to generate region proposals and estimate hand rotation. Huang *et al*. [9] proposed region growth approach for hand detection. Recently, Khan and Borji [10] proposed a new hand segmentation model which gives improved results comparing to the previous methodologies, and they added new annotations to existing EgoHands dataset by labeling each hand pose with 16 new activities. Also, Garcia-Hernando *et al*. presented experimental evaluations for RGB-D and pose-based hand action recognition in the first-person setting. [6]

2.3. Activity localization

Pirsiavash and Ramanan [17] presented an algorithm to solve the problem of detecting activities of daily living (ADL), and their results outperform existing approaches for wearable ADL recognition. Serra *et al*. [22] introduced a random-forest based hand classification algorithm and a SVM based gesture recognition method. Baraldi *et al*. [2] presented a method for monocular hand gesture recognition by using dense trajectories approach in egocentric view. The recognition task can be accomplished in near real-time. Ma *et al*. [14] have developed a twin stream CNN network architecture to learn features that capture object attributes and hand-object configurations. However, all the works above didn't include interaction between the egocentric viewer and others.

Ryoo and Matthies [21] developed a kernel-based activity recognition method to recognize interactions targeted to the egocentric viewer. However, the recognition task requires a full human body movement captured by the camera. Bambach *et al*. [1] built CNNs to classify activities based on hand segmentation with background information masked out. Khan and Borji [10] used a RefineNet and achieved a higher accuracy compare to Bambach's *et al*. work.

3. Methods

3.1. Dataset overview

This project used data in the EgoHands dataset, which contains 48 first-person 30-fps videos with complex interactions between two people captured by head-mounted cameras [1]. Each video lasts 90 seconds. The videos are randomly partitioned into 36 training, 4 validation, and 8 test groups, with the actors, activities, and locations evenly

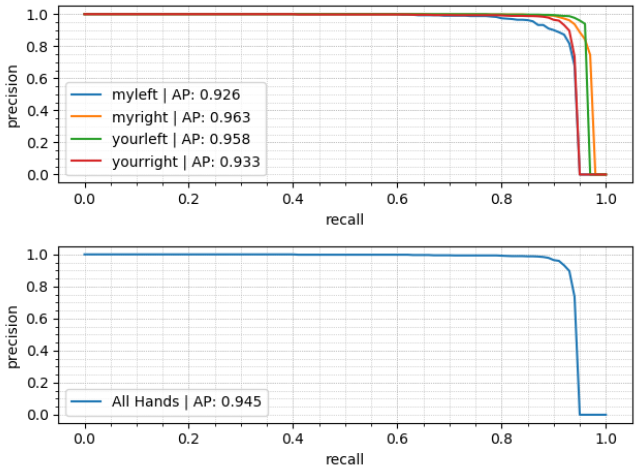

Figure 3. *Precision-Recall curves for detecting hands*. *Top:* Results for four hand types, where colors indicate the different hand types. *Bottom:* Results for all hand types. IoU threshold is set to 0.5

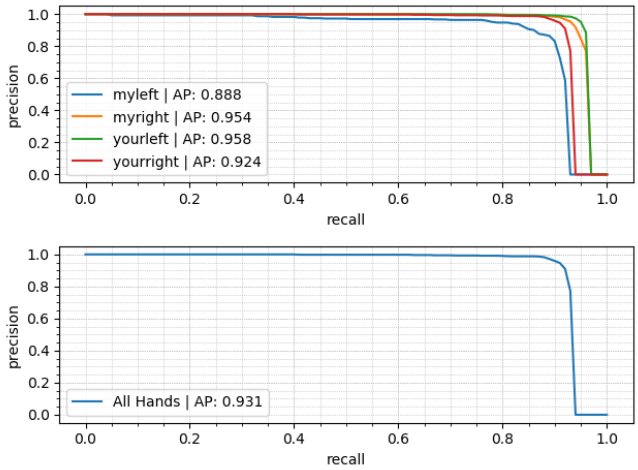

Figure 4. *Precision-Recall curves for segmenting hands*. *Top:* Results for four hand types, where colors indicate the different hand types. *Bottom:* Results for all hand types. IoU threshold is set to 0.5

distributed. One hundred frames were selected from each video (about one frame per second) and annotated, which results in a total of over 15,000 hand instances. Thus, the final dataset used in this project included 40 groups of frames as the training/validation data and 8 groups of frames as the testing data (each group has 100 frames).

3.2. Hand detection and segmentation

We formulated the hand detection and segmentation as an object detection task. The 4 hand types are simply 4 object classes. In order to accomplish the task, the data was split into 3 subsets: training, validation and testing. The training data for hands detection and segmentation includes 3600 frames and 11,358 instances. Validation and testing set contain 400 frames and 800 frames.

Adopting Mask R-CNN for hand detection and segmentation. We adapted Mask R-CNN from the Detectron [7] library on Caffe2 [16]. It accepts custom dataset in COCO style json format [13] for a better performance. Because the EgoHands dataset was created in Matlab format, the first step was to convert the data into a proper format. Each instance in the json file contains only one segmentation for one hand type (*e.g*. My left, My right, Your left and Your right). Thus, each hand is independent of each other. The size of each input image is 720 pixels tall by 1280 pixels wide.

After converting, we trained MSRA's original ResNet-101 pre-trained model and fine-tuned on EgoHands dataset. The pre-trained model was used to generate feature map from the original image. Then the feature map was fed into Regional Proposal Network (RPN). The RPN passes sliding windows over the feature map, and outputs 9 anchors at each window, which classified whether an anchor contains a hand type. The anchor ratio was set to $(0.5, 1, 2)$. The model then used proposed anchors for classification. Simultaneously, the feature map was also sent to a Fully Convolutional Network(FCN) for mask generation.

The model was trained on 1 TITAN X GPU. Each mini-batch involves 2 images per GPU and 256 RoIs per image. We used a weight decay of 0.0001 and a momentum of 0.9. The learning rate was set to 0.005 for 60,000 iterations. We used 2000 RoIs per image for training and 1000 for testing. Option for horizontally-flipped images was turned off so that the model could learn the difference between different hand types. Training Mask R-CNN took about 5 hours on EgoHands dataset.

During training, 13 images were filtered due to no foreground. This is because hands are blocked by table or other objects, thus no annotations were created for these frames.

3.3. Activity localization

Activity recognition is one vital element to understand the interaction between users in an egocentric video. Hand poses cue a rich information of what we are doing, especially when we are interacting with others. Bambach *et al*. believed the absolute and relative spatial information of hands also provided evidence and helped to recognize the activities [1]. Khan and Borji extended Bambach's *et al*. task and performed experiments on coarse-level and fine-level activity recognition. [10].

Our goal here is to find out **1)** whether activities can be detected and localized based on hand segmentations alone, without using any information on the context; **2)** whether the model can locate the activity as well as recognize different hand types. The non-hand background was removed using ground truth segmentation,

and replaced with mean gray, which was calculated by the mean value of images in ImageNet [4] and COCO dataset [13]([102.9801, 115.9465, 122.7717])

The activity localization can also be formulated as an object detection problem. We proposed 3 different approaches to train the network (See Figure 5):

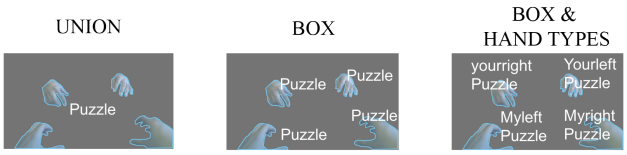

Figure 5. *Examples of input images and different approaches of activity localization* Non-hands backgrounds were replaced with mean gray. The three approaches: Union, Box, and Box with hand types, were used to train the network

Union: All hands were labeled as 1 of the 4 activities: Cards, Chess, Jenga and Puzzle. In this case, the model learned the spatial information between each hand, and we aimed to tell whether the spatial information helped to increase the precision of our task.

Box: Each hand was labeled as 1 of the 4 activities. This approach is similar to hand detection and segmentation. The only difference is that all hands in one individual have the same class.

Box with hand types: Each hand was labeled as 1 of the 4 hand types and activities. This approach combined 4 activities and 4 hand types together, with 16 classes in total.

In order to compare the performance, original images were also used to train the model, the process is shown in Figure 6. We used the same training configuration as for detection and segmentation, some results are shown in Figure 7.

4. Results

4.1. Evaluation

For both two tasks, we report mean Average Precision(mAP) and precision-recall curve based on Intersection-Over-Union (IoU). The IoU is computed by the intersection between predicted bounding box area with the ground truth. Then we compared the results with selected baseline methods.

4.2. Hand detection and segmentation

We chose Bambach's *et al*. work as our baseline [1] for detection, in which they used CNN with window proposal technique to classify each proposal window. The bottom pane of Figure 3 shows the precision-recall curve for detection when IoU is greater than 0.5. Our mAP achieves a better performance than the baseline (0.945 vs 0.807). In addition, the upper pane shows the precision-recall curve

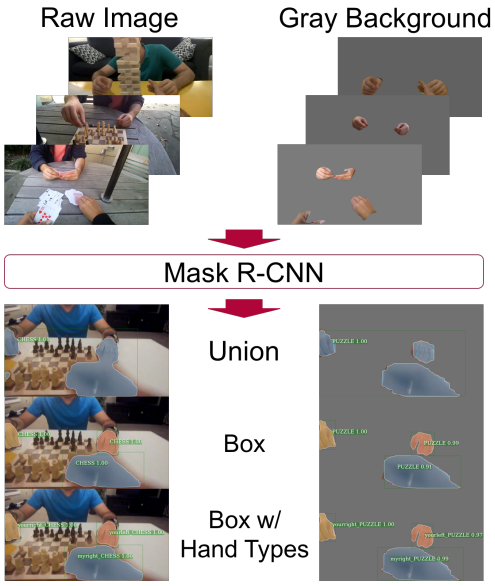

Figure 6. *Three different approaches and two different datasets for training*

for all 4 hand types. All of them significantly outperform the baseline. The comparison is shown in Table 2.

For segmentation, we chose a recent study by Khan and Borji [10] as our baseline. They formulated the segmentation task as a dense prediction problem where they used RefineNet for a pixel-wise segmentation. The bottom pane of Figure 4 shows our approach achieved the mAP at 0.931, which outperforms Khan's *et al*. approach by 5 percent.

Table 1 shows the performance comparison on detection and segmentation with two baselines. Figure 1 and 2 show some results of detection and segmentation.

	Method	**mAP**
Detection	Bambach *et al*.	0.807
	Ours	**0.945**
Segmentation	Khan *et al*.	0.879
	Ours	**0.931**

Table 1. Performance comparison on detection and segmentation.

4.3. Activity localization

We trained 6 models with 2 sets of data (original images and gray background images) and 3 different approaches. The benchmark is shown in Table 3 We didn't choose the baseline and reproduce their average precision for activity localization due to insufficient time. The example for different models can be found in Figure 7.

Method	My Left	My Right	Your Left	Your Right
Bambach *et al*.	0.64	0.727	0.813	0.781
Ours	**0.926**	**0.963**	**0.958**	**0.933**

Table 2. Detection AP on four different hand types

	Union	Box	Box with Hand Types
Raw Image	0.966	0.94	0.923
Gray Image	0.852	0.516	0.507

Table 3. mAP of Activity Localization of all 6 models

5. Discussion

5.1. Hand detection and segmentation

Our results outperform all previous work in detection and segmentation. However, by examining the top pane of Figure 3 and Figure 4, we found the precision for 'my left hand' detection and segmentation are relatively low compared to other 3 hands. Bambach *et al*. provided some insights in their work: The hands of egocentric viewer(especially non-dominant hand) are frequently truncated by the frame boundaries, and this may result in decreased precision(*e.g*. row 2, column 4 of Figure 1). We also hand picked some failure examples of double detection (*e.g*. row 3, column 2, 3 and 4 of Figure 1), which also results in a lower precision.

5.2. Activity localization

The 'Union' approach performs much better than the other 2 approaches (See Table 3). When multiple segmentation were fed to the network as 1 instance, the model learned the spatial information between hands.

The model was able to recognize hand types as well as activities using the 'Box with hand types' approach. However, without spatial and background information, the mAP only achieved 0.507.

Although the model was able to complete detection and segmentation tasks without background(*e.g*. column 2 in Figure 7), we found the model learned context information from background to improve the precision by comparing models trained by raw images and gray background images.

The lower precision in models using gray background images was mostly due to misclassification on different activities. The 'mask' branch of Mask R-CNN is separated from detection and classification. Thus, it generates mask no matter the category is. When background was given, the context information helped to classify objects. We need to investigate what features were learned from the background and to figure out how can we improve the precision with limited context information.

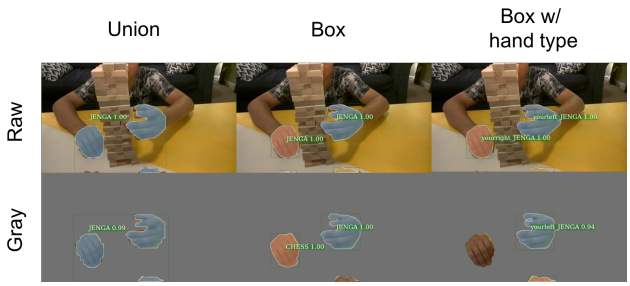

Figure 7. *Some activity localization results*, *First row:* Results using original images as training data. *Second row:* Results using gray background images as training data. *First column:* Results using Union approach. *Second column:* Results using Box approach. *Third column:* Results using Box with Hand types approach.

6. Conclusion

In this project, we based on previous work, and used current state-of-the-art for hand detection and segmentation method in EgoHands dataset. Our results showed that Mask R-CNN improved the performance noticeably and outperformed all works so far on this dataset. To better understand the activities in egocentric videos, we used the same algorithm and trained them with different approaches for activity localization. We found the spatial information significantly improved the performance on the task. Also, the model learned the background information when multiple segmentations were fed in as one instance. However, exact features that were learned by the model remain unclear.

In future work, we plan to **1)** Train new models with limited context information given and attempt to interpret what information model learned from the background that improved the precision. **2)** With the new dataset proposed by Khan and Borji [10], we will add new pose annotations and consummate the dataset.

7. Acknowledgments

Thanks to Sven Bambach for meeting with me, and thanks him and David Crandall for their expertise. Further, thanks to David Crandall for providing server for training my model.

8. Source code

All source codes are available in Github (https://github.com/whale9067/All-Hands-on-Desk).

References

[1] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 3, 4

[2] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS*. IEEE, 2014. 2

[3] A. Betancourt, P. Morerio, E. Barakova, L. Marcenaro, M. Rauterberg, and C. Regazzoni. Left/right hand segmentation in egocentric videos. *Computer Vision and Image Understanding*, 154:73–81, 2017. 1

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 4

[5] X. Deng, Y. Yuan, Y. Zhang, P. Tan, L. Chang, S. Yang, and H. Wang. Joint hand detection and rotation estimation by using CNN. *CoRR*, abs/1612.02742, 2016. 2

[6] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *arXiv preprint arXiv:1704.02463*, 2017. 2

[7] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. `https://github.com/facebookresearch/detectron`, 2018. 2, 3

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 1

[9] S. Huang, W. Wang, and K. Lu. Egocentric hand detection via region growth. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 639–644. IEEE, 2016. 2

[10] A. U. Khan and A. Borji. Analysis of hand segmentation in the wild. *arXiv preprint arXiv:1803.03317*, 2018. 1, 2, 3, 4, 5

[11] S. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 543–550, 2014. 1, 2

[12] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *Computer vision and pattern recognition (cvpr), 2013 ieee conference on*, pages 3570–3577. Ieee, 2013. 2

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 4

[14] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. *arXiv preprint arXiv:1605.03688*, 2016. 2

[15] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, 2007. 1

[16] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 3

[17] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012. 2

[18] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015. 1

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015. 1, 2

[20] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3137–3144. IEEE, 2010. 2

[21] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2730–2737. IEEE, 2013. 2

[22] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara. Hand segmentation for gesture recognition in ego-vision. In *Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices*, pages 31–36. ACM, 2013. 1, 2