Semi-automatic Hand Annotation Making Human-human Interaction Analysis Fast and Accurate

Stijn De Beugher[1], Geert Brône[2] and Toon Goedemé[1]
[1]*EAVISE, ESAT - KU Leuven, Leuven, Belgium*
[2]*MIDI Research Group - KU Leuven, Leuven, Belgium*

Keywords: Hand Detection, Tracking, Human-human Interaction, Human Pose.

Abstract: The detection of human hands is of great importance in a variety of domains including research on human-computer interaction, human-human interaction, sign language and physiotherapy. Within this field of research one is interested in relevant items in recordings, such as for example faces, human body or hands. However, nowadays this annotation is mainly done manual, which makes this task extremely time consuming. In this paper, we present a semi-automatic alternative for the manual labeling of recordings. Our system automatically searches for hands in images and asks for manual intervention if the confidence of a detection is too low. Most of the existing approaches rely on complex and computationally intensive models to achieve accurate hand detections, while our approach is based on segmentation techniques, smart tracking mechanisms and knowledge of human pose context. This makes our approach substantially faster as compared to existing approaches. In this paper we apply our semi-automatic hand detection to the annotation of mobile eye-tracker recordings on human-human interaction. Our system makes the analysis of such data tremendously faster (244×) while maintaining an average accuracy of 93.68% on the tested datasets.

1 INTRODUCTION

Many applications could benefit from image processing techniques in order to reduce manual input. In this paper we focus on a specific application, viz. the analysis of recordings in the field of human-human interaction. In this line of research, scholars are interested in the nonverbal behavior of humans during interaction and communication. An example of such a recording can be found in Figure 1. Research questions to be answered within this type of experiments are for example: Does the spectator notice the hand gesture of the right hand? If the presenter looks sideways, does that affects the viewing behavior of the spectator? In a presentation-training context, analysis of such a recording can be used to measure and assess inefficient use of non-verbal language, such as frantic hand gestures or immobile hands. In this paper we focus specifically on data captured by wearable or *egocentric* cameras, like for example GoPro cameras mobile eye-trackers. The analysis of the data captured by a mobile eye-tracker for example includes the annotation of the gaze cursor in terms of items that are instrumental for human-human interaction. If the gaze cursor overlaps with for example a human hand or a face, one has to annotate this event. Since such an analysis is extremely time-consuming, there is a growing interest in algorithms that reduce the manual workload. As mentioned above, we focus in this paper on the semi-automatic detection of specific body parts in images. Those detections could then be used as input for further analysis such as linking with gaze data or more complex analysis such as gesture detection.

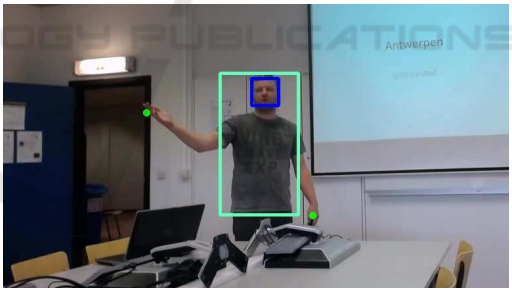

Figure 1: Typical frame captured by a mobile eye-tracker. Green dots are the hand detections, blue rectangle represents the face detection and green rectangle represents the upper body detection.

Recent developments in image analysis delivered highly accurate algorithms for both face and person

Beugher, S., Brône, G. and Goedemé, T.
Semi-automatic Hand Annotation Making Human-human Interaction Analysis Fast and Accurate.
DOI: 10.5220/0005718505520559
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 552-559
ISBN: 978-989-758-175-5

detection (Dollár et al., 2012), making this type of analysis relatively straightforward. Techniques for human hand detection on the other hand are far more complex. Most existing accurate hand detection algorithms use tools in order to facilitate the detections such as colored gloves or additional motion sensors. The use of such tools, however, may have an impact on the naturalness of recorded data, for both production and reception. Therefore we cannot allow items that may attract the visual attention of the participants in the experiment. The use of 3D cameras, which provide depth information and therefore facilitate the hand detection, is also not applicable since most of the egocentric cameras are 2D cameras.

In this paper, we present a semi-automatic hand detection algorithm based on an efficient combination of several techniques. We developed a system that automatically detects hands but asks for manual intervention when the confidence of a detection is below a certain threshold. Using such an approach reduces the manual analysis significantly while guaranteeing high accuracy. Since our approach relies on algorithms that are not computationally demanding, it is substantially faster as compared to the methods based on complex models. Driven by the wide applicability of our semi-automatic annotation tool, we made it publicly available[1] for other users.

The remainder of this paper is organized as follows: in section 2, we present a thorough comparison of existing hand detection algorithms. In section 3 the integration of the manual intervention is explained while in section 4 we discuss our hand detection algorithm in detail. In section 5 the results are discussed and in section 6 a final summary is given.

2 RELATED WORK

Traditionally, one can divide hand detection techniques into four categories. In this section we give an overview of existing techniques and explain the limitations of these approaches.

A well-known method for hand detection is the use of colored gloves, which is used as a marker that can be easily detected in images. In recent work (Wang and Popović, 2009) uses a multi-colored glove, enabling the detection of various hand orientations and poses. Since we focus on hand detection in natural and unconstrained scenes, we cannot afford the use of colored gloves, since they disturb the visual attention during a conversation.

A second approach of hand detection is making use of motion sensors (Stiefmeier et al., 2006). Typically multiple sensors, like ultrasonic transmitters and inertial sensor modules are placed on the user. Because of the same reason as mentioned above, it is not recommended to place additional sensors on the participants due to possible interference in the natural behavior.

The increasing popularity and public availability of 3D cameras paved the way for a third type of hand detection (Ren et al., 2013). These cameras provide useful depth information of a scene. Depth information facilitates hand detection and it even enables the detection of small items such as for example fingertips (Raheja et al., 2011). Although this is a promising approach, it is not applicable in our application since most of the egocentric cameras, like for example mobile eye-trackers, are not equipped with 3D sensors.

A last approach of hand detection is based on image processing in 2D images without the need of additional markers or sensors placed on the body. In (Kolsch and Turk, 2004) a hand tracking approach was described based on KLT features in combination with color cues. Such an approach yields good results as long as the hand is easily visible (large enough) in order to calculate an adequate number of features. This approach is not applicable in our type of experiments, where the hands represent only a small part of the image, as can be seen in Figure 1. In (Shan et al., 2007) a real-time hand tracking is presented using a mean-shift embedded particle filter. Their system is very fast (only 28ms per frame is needed) however the resolution of their test images is only 240×180 pixels. In their experiments they only detect and track a single hand, whereas in our application we need to track both hands with respect to the human pose. (Bo et al., 2007) presents a hand detection technique based on a combination of Haar-like features and skin segmentation. This approach is sufficiently accurate in controlled scenes, e.g. a clean white background on the other hand, the approach suffers from high false positive rates when applied to less constrained scenes. In the work of (Eichner et al., 2012) a technique for estimating the spatial layout of humans in still images is presented, using a combination of upper body detection and the detection of individual body parts. This method performs well on larger body parts (such as arms or heads), whereas smaller parts (e.g. hands) are much more challenging. The accuracy of this technique largely depends on the upper body detection, detection at a wrong scale will result in deviating limb detections. This approach works far from real-time: on average 25 seconds are needed for processing a single 1280×720 frame. A similar approach was proposed by (Yang and Ramanan, 2011). This

[1] http://www.eavise.be/insightout/Downloads

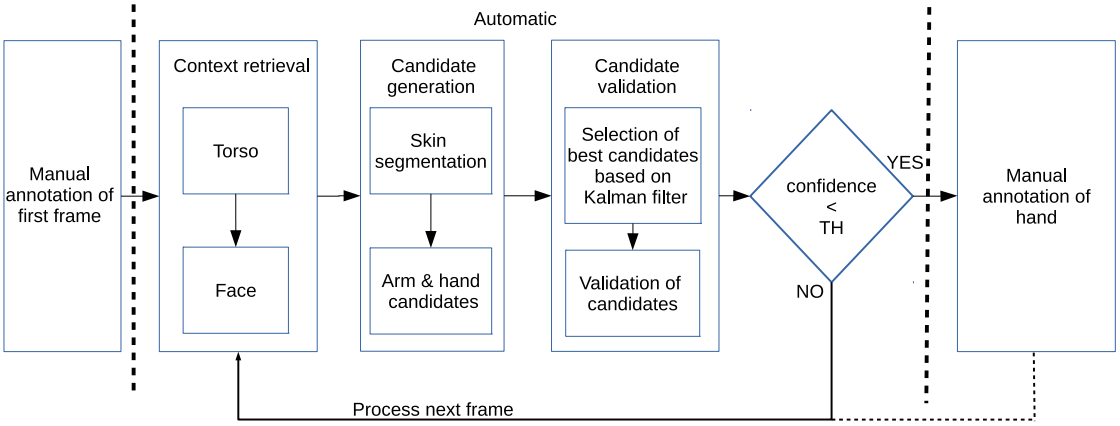

Figure 2: Workflow of our hand detection approach.

is a method for human pose estimation in static images based on a representation of part models, taking into account the relative locations of parts with respect to their parents. (e.g. elbow w.r.t. to shoulder), which results in accurate detections. However, the authors admit their approach has difficulties with some body poses e.g. raised or fully stretched arms. A highly accurate approach was proposed by (Mittal et al., 2011), combining a deformable-part-model (DPM) of a human hand with skin segmentation to generate hand candidates. Those candidates are then suppressed using a super-pixel based non-maximum suppression yielding accurate detections. This technique has a large computational cost due to the complexity of the DPM, yielding an average processing time of a frame of 1280×720 of about 290 seconds. In our recent work (De Beugher et al., 2015), we improved this approach greatly using a faster DPM calculation, a reduction of the search space based on a human upper body detector and avoiding the need for a super-pixel segmentation. Unfortunately, despite our efforts, the technique remains too slow for practical use (37 seconds for an image of 1280×720). The work of (Spruyt et al., 2013) is also a recent hand detection approach focused on real-time HCI. However, compared to the images we tackle, the difficulty of the datasets they used is limited, in that they do not involve typical challenges of real-life data, like e.g. changing camera angles and distances, (partial) occlusions of and by hands, etc. These are situations in which their approach fails. Furthermore, they do not provide the possibility to manually steer the detections in case of false detections.

Our approach differs significantly from all previously mentioned ones. We propose a hand detection methodology, which is both fast and accurate, and allows for manual intervention. We extensively optimized and combined previously described techniques, and integrated them with probabilistic information.

3 SEMI-AUTOMATIC APPROACH

As mentioned before, we present a semi-automatic hand annotation tool to process a sequence of consecutive frames. It is important to mention that we tackle an annotation application that is currently done completely manual. The goal is to reduce the amount of manual analysis as much as possible while maintaining top accuracy. Therefore we developed an algorithm to detect body parts that are instrumental for this type of analysis: hands, face and torso. The detection of a human face and torso can be done automatized with available accurate algorithms like (Felzenszwalb et al., 2010; Viola and Jones, 2001). The bounding boxes in Figure 1 show example detection results. The detection of human hands on the other hand is far more complex and it is even impossible to reach very high accuracy when using the most complex approaches. We developed a system that automatically detects hands in images based on simple color cues. However when the confidence of a hand detection drops below a preset threshold, our automatic analysis is paused and manual intervention is asked from the user, to manually annotate the corresponding hand. After this intervention, the automatic analysis is continued. We fine-tuned the parameters of our system to ensure the lowest amount of manual interventions as possible, while guaranteeing high accuracy. On top of those manual interventions, we ask the user to manually annotate the first frame of the recording, ensuring a good starting point for the detections. The integration of these manual interventions are indicated in Figure 2.

4 HAND DETECTION

As illustrated in Figure 2, our system is a combination of several processing blocks. A first step is the detection of a human upper body, which is used to identify the presence of a person and to reduce the search area. Next we apply a skin segmentation that is used to generate hand candidates. To further enhance the detections a tracker is used for temporal smoothness. Finally we validate those candidates using a) a comparison between a predicted position and the candidate and b) a validation of the relative position between joints like for example wrist, elbow and shoulder. Each step of this workflow will be discussed below.

4.1 Context Retrieval

The first stage of our approach is based on the work of (De Beugher et al., 2015) and is used to get context information. We use an accurate human upper body detector based on a DPM as proposed in (De Beugher et al., 2014) to detect the presence of a person in the images. The advantage of this model is that we can cope with images in which a person is not visible from head to foot, as in most of the images captured by a mobile eye-tracker. This torso detection is also used to reduce the search area for the hands: the width of the upper body is extended by factor 3.5 while the height is extended by factor 1.8. Those enlarge factors are determined empirically and ensure that an average human hand lies within the extended region. This step allows us to restric the search for hands within this region and to discard the rest of the image. In Figure 3a the original torso detection is displayed using the purple rectangle, while the blue rectangle illustrates the extended bounding box. Next to the torso detection, we also apply a face detection to find the face and viewing direction using both a frontal and profile Haar-based face model (Viola and Jones, 2001). Both face and upper body detections are stored since they are instrumental for human-human interaction.

4.2 Candidate Generation

We segment the image patch, which is the extended bounding box, in skin and non-skin using rules in three color spaces as introduced by (Rahim et al., 2006), shown in Figure 3b. After two dilation and two erosion steps, we fit a contour over each sufficiently large group of skin pixels (Figure 3c) on which a bounding ellipse is fitted (Figure 3d). Each endpoint of the major axes of an ellipse is treated as a hand candidate, as illustrated by the green dots in Figure 3e. The example shown in Figure 3d) contains four ellipses: two of them contain the correct hands, one coincides with the face and is therefore automatically discarded and one ellipse is false, found on an approximately skin-colored chair.

4.3 Candidate Validation

The final stage of our approach is developed to automatically select the best candidate for both left and right hand and to validate them. Temporal continuity of the image sequence is exploited using a Kalman filter. The selection of the best candidate for both left and right hand is done by choosing the hand candidate with the smallest distance to the Kalman filter prediction of the respective hand. This Kalman tracker uses either the detection or the manual intervention in the previous frame to predict the position using a constant velocity motion model. As mentioned above, each hand candidate belongs to a line (major axis of the ellipse). When we selected the best candidate for a hand, we use the other endpoint of the corresponding line for validation. The remaining endpoint can be seen as a joint, which corresponds to an elbow in case the person wears short sleeves, or corresponds to a wrist in case the person wears long sleeves.

We utilize probability maps to summarize possible positions of elbows and wrists w.r.t. the left and right shoulder. In order to filter false detections, we weight candidate joints to these probability maps and hereby we remove impossible joint positions. The location of the shoulders is estimated using both face and upper body detection: y-position of the shoulder corresponds to the bottom of the face-bounding box, while the x-position of each shoulder is obtained by the width of the torso detection. The probability maps are created using the original labeling of the publicly available Buffy dataset (Ferrari et al., 2009), more specifically we used the labelings of wrist, elbow and shoulder. The motivation to use this particular dataset comes from the large variety of human poses that are recorded in this dataset, as can be seen in the sample frames in Figure 4. For each image in this dataset we calculate the relative position of elbow and wrist w.r.t. the shoulder, this results in four sets each containing 1496 data points. In Figure 5 we show the data points for both left elbow and wrist. The red dot illustrates the position of the left shoulder. Data points in the upper image are the relative positions of the left wrist w.r.t the left shoulder. The data points in the bottom image are the relative positions of the left elbow w.r.t the left shoulder. Two mirrored sets of points are used for the right shoulder. Next we apply a Gaussian

a) b) c) d) e)

Figure 3: Generation of hand candidates: a) original image, b) skin segmentation, c) contour detection, d) fit ellipse, e) final hand candidates.

Figure 4: Example frames of the Buffy dataset (Ferrari et al., 2009) indicating the large variety of human poses within this set. From this labeled dataset, our probability maps (P_{Elbow}) and (P_{Wrist}) are derived.

smoothing resulting in a dense map. After a normalization of the dense map, this results in a probability map. In total we developed four probability maps: elbow w.r.t. shoulder (P_{Elbow}) and wrist w.r.t. shoulder (P_{Wrist}), each for both left and right side.

The relative position of the left joint is validated against each of the two probability maps of this side. The best probability result ($max(P_{Elbow}, P_{Wrist})$) is than used in the next steps. The same rules are applied on the right joint.

Both probability P_{Elbow}, P_{Wrist} and distance D to the Kalman prediction are taken into account for the confidence condition C:

$$C = \{(D > D_{max}) \wedge (pred > pred_{max})\} \vee \{(max(P_{Elbow}, P_{Wrist}) < P_{TH})\} \quad (1)$$

D_{max} stands for the maximum allowable distance between prediction and hand candidate. This maximum distance is dependent on the size of the person and is therefore calculated as follows: face width $\times$ 0.75. *pred* stands for the amount of consecutive predictions that are used (thus no valid detection was available), $pred_{max}$ stands for the maximum amount of predictions that is allowed. Finally, P_{TH} stands for the lowest probability value that is allowable. If condition C (equation 1) is not met, our system automatically pauses and asks for manual intervention, as described above. Otherwise, the next image is processed automatically.

By varying the above mentioned parameters, one can increase or decrease the amount of manual interventions as shown in Figure 6. This Figure reveals that when the strictness of the confidence is lowered, our system asks less manual interventions. It is obvious that this comes at a cost of lower accuracy.

We implemented an additional feature in our approach to reduce the amount of manual interventions. As mentioned before, the probability maps are developed using the data labels from the Buffy dataset. Although this dataset contains a large variety of human poses, it may occur that a particular pose of a wrist or an elbow corresponds to a low probability since this particular pose occurs only sporadic in the Buffy dataset. When the automatic processing is paused due to a too low probability result, the user can indicate that the particular joint position is nevertheless correct. In the latter case the probability map is updated making this joint position valid in future processing. It is important to notice that the results shown in Figure 6 are obtained without this feature.

A video of our system is available online [2].

5 RESULTS

In this section we present the results of our semi-automatic hand detection algorithm. In order to validate our approach we used several publicly available datasets. Three of them are introduced in (De Beugher et al., 2015) and are further referred as D1, D2 and D3. A final dataset was introduced in (Spruyt et al., 2013) and is further referred as

[2] *http://youtu.be/DsxdBc4gGjg*

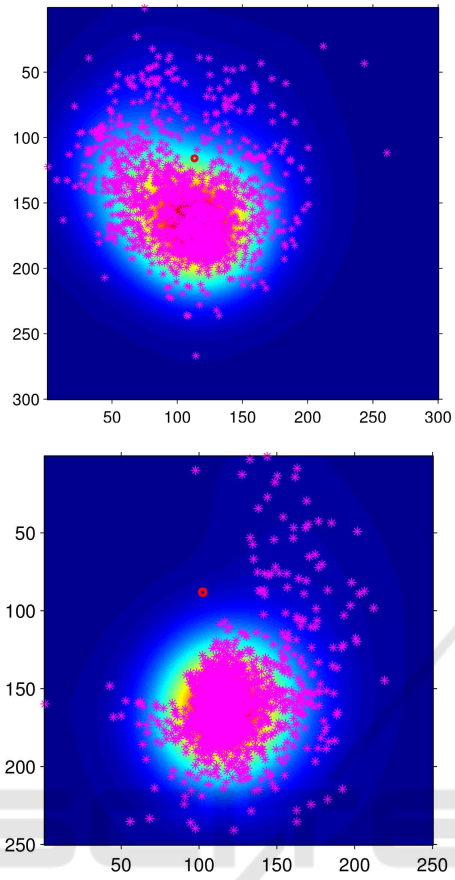

Figure 5: Top image shows data points and probability map of the left wrist w.r.t left shoulder. Bottom image shows the data points and probability map of left elbow w.r.t. left shoulder.

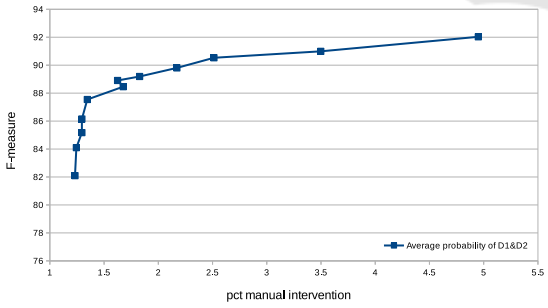

Figure 6: Influence of varying the parameters of the confidence calculation. Decreasing the amount of manual intervention comes at a cost of a lower accuracy.

D4. In total, the fingertips of 6000 hand instances are manually labeled in those sequences and are the groundtruth for our accuracy measurements. In our final implementation we have chosen to set both confidence calculation parameters $pred_{max}$ and P_{TH} to 5.

The leftmost columns of table 1, present the accuracy and the amount of manual intervention of our semi-automatic approach without the probabilistic validation. In the second column, the amount of manual interventions our entire system needed to reach the corresponding accuracy (F-measure) is shown. On average, our system asks for manual intervention in only 1.92% of the frames, and reaches an accuracy of 93.68%. A detection is considered valid if the distance between the detection and the annotation is below half-face width, which is a commonly used measure in other hand detection papers. As explained above, when the confidence of the detection drops below a certain threshold, manual intervention is asked. It is important to notice that the same parameters for the confidence calculation are used for all the datasets. In the third column of table 1 we show the performance of our earlier work (De Beugher et al., 2015) on datasets D1, D2 and D3. To allow for a fair comparison, we show the amount of manual interventions that is required in their system to achieve a similar accuracy as our work. As seen, we clearly outperform the competitor in accuracy while the manual work is significantly less. In Figure 7 we compare our approach, using the above mentioned settings, to our earlier approach (De Beugher et al., 2015). It is clear that our approach requires a substantially lower amount of manual interventions. In the two rightmost columns we show the accuracy of two full automatic approaches (Mittal et al., 2011; Yang and Ramanan, 2011) on the above mentioned datasets. It is clear that the accuracy of these automatic approaches is significantly lower than our semi-automatic approach.

In Figure 8 we show some example frames on the four datasets. The green circles represent the hand detections, yellow circles represent the joints (either wrist or elbow), red circles represent the shoulders (this is an estimation based on both upper body and face detection), and the pink circle represents the center of the face. We also draw the connections between the previously described points in order to symbolize the upper part of the human skeleton. In the rightmost image, we show the advantage of using two types of probability maps. Even when an arm is invisible in an image, our system is able to detect a correct hand and joint. In this case, the joint corresponds to a wrist. On top of the improvement in accuracy, we also present a significant improvement in computational speed. Our semi-automatic hand detection algorithm is about 244$\times$ faster as compared to our previous approach (De Beugher et al., 2015). This needed approximately 36 sec to process an image of 1280$\times$720, where our present approach only requires 150 ms to process the same frame. This improvement in computational speed is mainly achieved by abandoning the DPM model based hand detection.

Table 1: Comparison of our semi-automatic hand detection approach and (De Beugher et al., 2015), *man.* indicates the amount of hands of which manual interventions was required.

	Ours without prob.		**Ours with prob.**		**De Beugher(2015)**		**Mittal(2011)**		**Yang(2011)**	
	man.	F-measure	man.	F-measure	man.	F-measure	man.	F-measure	man.	F-measure
D1	1.63%	90.85%	**2.62%**	**95.76%**	4.2%	95.28%	0%	85.0%	0%	24.2%
D2	2.55%	83.57%	**1.84%**	**92.75%**	19%	92.13%	0%	46.5%	0%	46.5%
D3	0.65%	81.08%	**0.75%**	**88.31%**	8.6%	87.62%	n.a.	n.a.	n.a.	n.a.
D4	n.a.	n.a.	**2.47%**	**97.89%**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
avg.	1.61%	85.17%	**1.92%**	**93.68%**	6.72%	91.4%	0%	68.15%	0%	35.35%

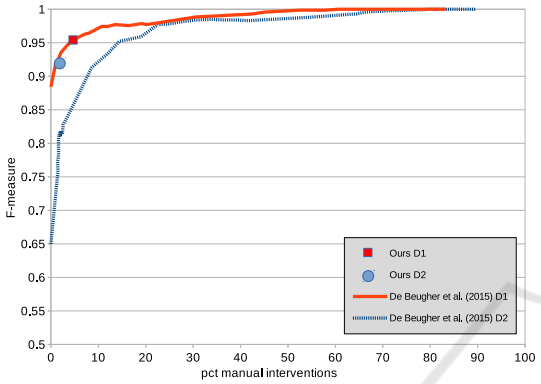

Figure 7: Results of (De Beugher et al., 2015) superposed with our present results on the same datasets.

In table 2 an overview of the speed results is given. Next to the speed of our earlier approach, we show also the computational time of two other hand detection techniques on an image of 1280×720. All timing results were acquired on the same hardware (Intel Xeon E5645).

Table 2: Execution times per frame averaged over all frames.

	avg. time/frame
Ours	**150ms**
(De Beugher et al., 2015)	36.67s
(Yang and Ramanan, 2011)	113s
(Mittal et al., 2011)	293.33s

Next to the improvement in accuracy and computational speed, our approach differs significantly from our earlier work (De Beugher et al., 2015) in several ways: (a) we no longer need the highly computational hand models, (b) skin detection is combined with contour detection, (c) apart from the hands, we also track the wrist and/or elbow, (d) we validate candidates against probabilitiy maps.

6 CONCLUSION AND FUTURE WORK

In this paper we proposed a novel semi-automatic hand detection algorithm for the annotation of egocentric recordings in the context of research on human-human interaction. Our approach is based on integrating manual supervision with an automatic hand detection algorithm. Our system automatically detects hands in images, but when the confidence drops below a certain threshold, our system asks for manual intervention. This yields maximally accurate annotations at the cost of a minimal amount of manual input. Our hand detection algorithm works as follows: first we apply a highly accurate upper body detection to reduce the search area. Next we use a skin segmentation to generate hand candidates. Finally a set of trackers is used to follow the hands over time, combined with knowledge of human poses (e.g. relative position between shoulder and wrist). This is combined to decide whether manual intervention is required.

We validated our approach using four publicly available datasets and compared against a number of recent hand detection algorithms. This validation reveals that our approach is more accurate than the competitor while being more than 244× faster, which makes it more applicable in real life applications. Moreover, our system requires an even lower number of manual interventions in order to achieve the same accuracy.

Our future work concentrates on further exploring the capabilities and boundaries of our approach. We plan to test our approach on more real-life eye-tracking recordings and to use our semi-automatic approach as annotation tool. On top of that we plan to use this algorithm as input for more complex analysis such as gesture detection.

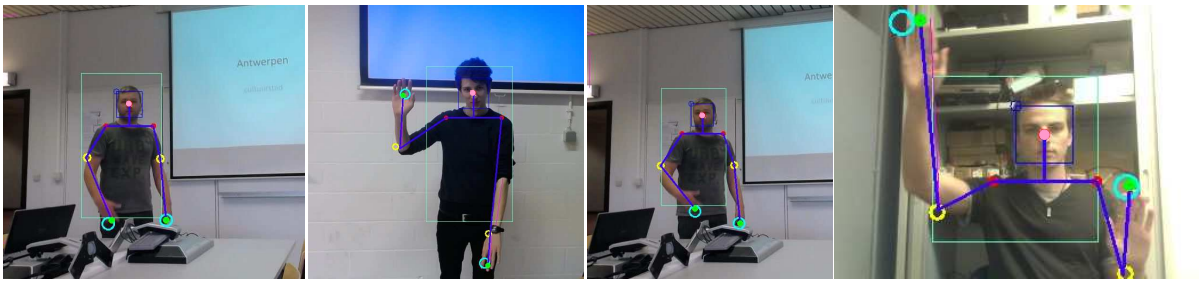

Figure 8: Examples of our detections on the four datasets. Green circles are the hand detections, yellow circles are the corresponding joints.

ACKNOWLEDGEMENTS

This work is financially funded by OPAK via the Into The Wild research project.

REFERENCES

Bo, N., Dailey, M. N., and Uyyanonvara, B. (2007). Robust hand tracking in low-resolution video sequences. In *Proc. of IASTED*, pages 228–233, Anaheim, CA, USA.

De Beugher, S., Brône, G., and Goedemé, T. (2014). Automatic analysis of in-the-wild mobile eye-tracking experiments using object, face and person detection. In *Proc. of VISAPP*, pages 625–633.

De Beugher, S., Brône, G., and Goedemé, T. (2015). Semi-automatic hand detection - a case study on real life mobile eye-tracker data. In *Proc. of VISAPP*, pages 121–129.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on PAMI*, 34(4):743–761.

Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2012). 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99:190–214.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on PAMI*, 32(9):1627–1645.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2009). Pose search: Retrieving people using their pose. In *Proc. of CVPR*, pages 1–8.

Kolsch, M. and Turk, M. (2004). Fast 2d hand tracking with flocks of features and multi-cue integration. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 10 - Volume 10*, CVPRW '04, pages 158 – 158, Washington, DC, USA. IEEE Computer Society.

Mittal, A., Zisserman, A., and Torr, P. (2011). Hand detection using multiple proposals. In *Proc. of BMVC*, pages 75.1–75.11. BMVA Press.

Raheja, J., Chaudhary, A., and Singal, K. (2011). Tracking of fingertips and centers of palm using kinect. In *Proc. of CIMSiM*, pages 248–252.

Rahim, N. A. A., Kit, C. W., and See, J. (2006). RGB-H-CbCr skin colour model for human face detection. In *Proc. of M2USIC*, Petaling Jaya, Malaysia.

Ren, Z., Yuan, J., Meng, J., and Zhang, Z. (2013). Robust part-based hand gesture recognition using kinect sensor. *IEEE Transactions on Multimedia*, 15(5):1110–1120.

Shan, C., Tan, T., and Wei, Y. (2007). Real-time hand tracking using a mean shift embedded particle filter. *Pattern Recognition*, 40(7):1958 – 1970.

Spruyt, V., Ledda, A., and Philips, W. (2013). Real-time, long-term hand tracking with unsupervised initialization. In *Proc. of ICIP*, pages 3730–3734.

Stiefmeier, T., Ogris, G., Junker, H., Lukowicz, P., and Troster, G. (2006). Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario. In *Proc. of ISWC*, pages 97–104.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. pages 511–518. Proc. of CVPR.

Wang, R. Y. and Popović, J. (2009). Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 28(3).

Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Proc of CVPR*, pages 1385–1392. IEEE.