Semi-automatic Hand Detection
A Case Study on Real Life Mobile Eye-tracker Data

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

Keywords: Eye-tracking, Hand detection, Hand tracking, Human-human interaction, Gaze, (Semi-)automatic analysis.

Abstract: In this paper we present a highly accurate algorithm for the detection of human hands in real-life 2D image sequences. Current state of the art algorithms show relatively poor detection accuracy results on unconstrained, challenging images. To overcome this, we introduce a detection scheme in which we combine several well known detection techniques combined with an advanced elimination mechanism to reduce false detections. Furthermore we present a novel (semi-)automatic framework achieving detection rates up to 100%, with only minimal manual input. This is a useful tool in supervised applications where an error-free detection result is required at the cost of a limited amount of manual effort. As an application, this paper focuses on the analysis of video data of human-human interaction, collected with the scene camera of mobile eye-tracking glasses. This type of data is typically annotated manually for relevant features (e.g. visual fixations on gestures), which is a time-consuming, tedious and error-prone task. The usage of our semi-automatic approach reduces the amount of manual analysis dramatically. We also present a new fully annotated benchmark dataset on this application which we made publicly available.

1 INTRODUCTION

Detection of human hands in real-life images is an extremely challenging task due to their varying shape, orientation and position. Our motivation for developing a highly accurate hand detector comes from the wide applicability in a variety of disciplines including computer science, linguistics, sociology and psychology. Practical applications for such a technique include human-computer and human-robot interaction, gesture detection, automatic sign language translation, active gaming, etc. Recently, several highly accurate hand detection algorithms were developed for 3D images (Van den Bergh and Van Gool, 2011). Hand detection in 2D images, however, is far from a trivial task due the lack of depth context. Several attempts were made including skin-based detections (Wu et al., 2000), model-based detections (Bo et al., 2007; Karlinsky et al., 2010; Mittal et al., 2011) or pose estimation techniques (Yang and Ramanan, 2011). Unfortunately when applied to real-life images, their performance drops significantly.

On top of the challenging task we try to tackle, we aim to develop a generic method to achieve a high detection rate. It is well known that fully automatic approaches typically do not guarantee high accuracy in practical cases. However many applications could benefit from such a generic approach, e.g. the removal of privacy sensitive content such as faces in mobile mapping images, generation of ground-truth data, cartography by using object detection in aerial images, etc. To overcome this we expanded our framework with an intelligent mechanism which automatically demands for manual input when the confidence of a detection is below a threshold value. Using such an approach increases the detection rate significantly at the cost of a limited amount of manual interventions. For a certain target accuracy, our system computes the minimum amount of manual interactions.

In contrast to other techniques, we focus in this work on the detection of hands in video material. Using sequences of images gives us the opportunity to use the spatio-temporal relationship between consecutive frames to increase the detection rate. We use a 3-stage framework to generate the best possible result. First, we reduce the search space, using a human-torso detector. Second, we make a hypothesis using a sliding window approach of a hand model combined with a skin-based hand detection. Third, we use an advanced elimination approach to remove false detec-

De Beugher S., Brône G. and Goedemé T..
Semi-automatic Hand Detection - A Case Study on Real Life Mobile Eye-tracker Data.
DOI: 10.5220/0005306601210129
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 121-129
ISBN: 978-989-758-090-1

tions in combination with a tracker resulting in reliable detections.

To validate our framework, we present a (semi-) automatic analysis of mobile eye-tracker data in the context of human-human interaction studies. The analysis of these data generally requires substantial manual annotation work (Gebre et al., 2012; Jokinen, 2010; Al Moubayed et al., 2013; Brône and Oben, 2014). The eye-tracking community would greatly benefit from the implementation of techniques that reduce the manual annotation load, like e.g. the detection of gesture strokes (Gebre et al., 2012) and body language categorization (Williams et al., 2008). The presented framework aims to contribute to these developments and proposes a technique to (semi-) automatically detect hands in video data recorded by a mobile eye-tracker. By mapping eye gaze data on interlocutors' body parts that are instrumental to face-to-face communication (like hands and faces), a first step in the analytical process is realized, as it allows for basic calculations of visual distribution. These data can then serve as the basis for further analytical work (e.g. the analysis of visual fixations on certain gesture types).

Next to a fast and accurate hand detection framework, an important contribution of this paper is a generic (semi-)automatic detection approach. Furthermore, during our study, we noticed that it is hard to find fully annotated video material of human hands in real life recordings. Therefore we made our annotated dataset of eye-tracker recordings publicly available. This set contains two sets of data of which approximately 1000 frames were annotated[1].

This paper is organised as follows: In section 2, we discuss related work on hand detection. Section 3 clarifies our hand detection framework in detail. In section 4 we discuss our novel (semi-)automatic approach in which a minimal manual intervention step enhances the detection rate. Finally in section 5 we present the results on a pre-existing dataset and on our publicly available eye-tracker recordings that were performed to validate the approach.

2 RELATED WORK

In recent years several attempts have been made to develop an accurate hand detector for 2D images, mostly by decreasing the complexity of the problem. Examples are the use of artificial markers e.g. coloured gloves (Wang and Popović, 2009) or using a static camera enabling the use of background segmentation (Pfister et al., 2012). In this paper however, we focus on real-life applications where unmarked body parts need to be detected automatically, and therefore we only review the most popular methods that are applicable to natural settings.

A well known object detection technique is based on *Haar-like features* (Viola and Jones, 2001). This technique combines a set of weak classifiers to build a final strong classifier and uses a sliding window approach to search for specific patterns in the image. In (Bo et al., 2007) this technique is used as a basis for a hand detection algorithm, in combination with a skin detector to eliminate non-hand detections. Unfortunately the performance of this technique on unconstrained images is insufficient. Newer detectors outperform greatly Haar-based techniques.

A second approach is based on the *Deformable Parts Model* (Felzenszwalb et al., 2010), which is an extension of *Histograms of Oriented Gradients* (HOG) (Dalal and Triggs, 2005). This approach allows for the definition of a model of an object which is invariant to various postures or viewing angles. In (Mittal et al., 2011) this technique is used to create two models of a hand, both with and without its surrounding region, e.g. the wrist (see figure 1). In addition, they use a skin detector based on the average skin colour of the face. This skin detector is used to improve the detection rate by searching for arms in the image. Finally a super-pixel based non-maxima suppression (NMS) is used in which overlapping bounding boxes are suppressed. A drawback of this method is the high computational cost: processing a frame of 360 x 640 pixels takes up to 4 minutes, of which the greater part is spent on superpixel calculations.

Another hand detection approach was presented by (Spruyt et al., 2013). In this paper an invariant hough forest detector was used, resulting in a robust detection of the hand locations. Nevertheless, in our application the detection of the hand orientation is also of great importance on top of the location itself. Therefore we can not use such a basic approach.

In (Eichner et al., 2012) the human pictorial structure is used. This approach searches for limbs in a human torso using the spatial relation between them. This method performs well on larger body parts (such as arms or heads), whereas smaller parts (e.g. a hand) are much more challenging. There are two major drawbacks of this method: a) the requirement that all body parts are visible in the image and b) they have a limited set of body poses that are detectable.

A pose estimation algorithm is proposed by (Yang and Ramanan, 2011). This method is highly accurate since it has several parts for each limb and uses con-

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

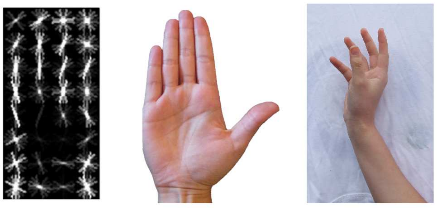


Figure 1: Illustration of the hand models. The left image is the HOG representation of the hand model. The middle image illustrates the hand model, while the right image is an illustration of the context model (hand and its surrounding region including the background and wrist.)

textual co-occurrence relations between them. This method is designed for static images and its accuracy decreases drastically when motion blur is present, caused by moving body parts. The authors also admit their model has difficulties with some body poses (e.g. raised arms).

Based on a comparison of the previously described techniques, we opted for the work of (Mittal et al., 2011) as a starting point for our algorithm. This approach achieves decent accuracy and its source code is publicly available so we can easily compare our method against it. In the next section we discuss the modifications we made in order to improve the detection results drastically, and how we extended to video.

3 HAND DETECTION FRAMEWORK

An overview of our hand detection algorithm is given in figure 2. The general idea is that we first detect a human torso in the image, giving a robust reference for the detection of smaller body parts. Next we detect the face resulting in an indication of the hand sizes. After that, we detect hands using a model introduced by (Mittal et al., 2011) in combination with a skin-based detection. Then we apply an advanced elimination scheme in order to remove false detections. Finally we use a Kalman filter to track left and right hand using the spatial relationship of consecutive frames.

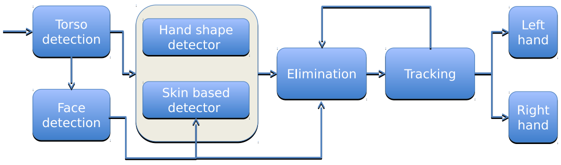


Figure 2: Graphical representation of the proposed hand detection framework. The three stages: torso and face detection, hand detection and a combination of elimination and tracking.

3.1 Torso Detection

The first stage in our approach is the detection of a human torso, for which we use our own torso detector as we proposed in (De Beugher et al., 2014). This torso detector is a part-based model (Felzenszwalb et al., 2010), trained using only the upper 60% of the labeled bounding boxes of human bodies of the standard PASCAL VOC dataset[2]. Using this model, rather than the more widely used full person detector, has the advantage that we can cope with images in which a person is not completely visible (from head to foot) such as, for example, in most of the images captured by a mobile eye-tracker (see figure 5) in a natural setting.

3.2 Face Detection

The next stage is a face detection step (Viola and Jones, 2001), which is used as a way to further improve the accuracy of the hand detections. In the work of (Mittal et al., 2011), the face detection is only used for skin segmentation. If a face is detected, they apply a skin colour based proposal method to improve their detection results. In our approach on the other hand, we also make use of the proportions of the face by rejecting hand detections which have an abnormal size compared to the size of the face. This is based on the general rule that a human face has, about the same size as an outstretched human hand.

3.3 Hand Detection

When the torso and face location are known, we run our actual hand detection algorithm. Instead of searching for hands in the entire image, we define a search area by expanding the torso detection bounding box in both vertical and horizontal orientation. As mentioned before, we started from the work of (Mittal et al., 2011). This means we use the same part-based deformable model of a hand, as illustrated in the left part of figure 1. In their approach, an additional context model is used. However, the experiments we ran for this study showed that the addition of this model introduces a significant amount of false detections, so that we opted not to use it.

The hand model was developed to detect upstanding hands, but in real-life recordings any hand-orientation is possible. Therefore we rotate the enlarged region around the detected torso in steps of 10 degrees per rotation, as illustrated in figure 3, yielding an accurate detection of hands in any orientation.

[2] The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Dataset http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html

Table 1: Accuracy of the hand model versus rotation angle of the images.

Step size	Precision	Recall	Time/frame
10 deg.	79,20 %	78,86 %	42 s
20 deg.	75,78 %	75,47 %	21 s
30 deg.	71,24 %	71,13 %	14 s
45 deg.	62,82 %	62,55 %	9,3 s
90 deg.	48,72 %	48,50 %	5 s

Using a larger step size decreases the computational cost, but also affects the accuracy of the detector as shown in table 1. This table shows the performance of the hand model on a set of 100 annotated frames of 1280 x 720 pixels. To further decrease the computational cost related to this type of model evaluation, we used the acceleration approach of (Dubout and Fleuret, 2012).

The hand model performs well as long as a hand is clearly visible in the image. However, when a hand is not visible or strongly deformed — for example due to motion blur caused by fast movements of the arms — these models show low detection rates. To overcome this problem, we developed an additional hand detection technique as shown in figure 4. This technique segments the image in skin and noskin based on three different colour spaces as introduced by (N. A. Abdul Rahim, 2006). In this work, skin colour is defined in both Red Green Blue (RGB), Hue Saturation Value (HSV) and Luma Chroma blue Chroma red (YCbCr) colour space resulting in a robust detection mechanism for skin, even under different lighting conditions. Using this approach is an improvement compared to the work of (Mittal et al., 2011), because we no longer depend on the accuracy of the face detector for skin segmentation. We apply this segmentation to the stretched torso detection as shown in figure 4(b). Next, we skeletonize this result using a sequence of several erosion and dilation steps in order to get an accurate estimation of the skeleton, as illustrated in figure 4(c). In a following step, we apply the information obtained from the face detector. We use the correlation between the human body parts to classify the skeletonized image. If a skeletonized part has a length which is similar to the height of the face, we classify it as a hand (as illustrated by the top row in figure 4. Parts that are larger than a face are automatically treated as an arm (as illustrated by the bottom row in figure 4). For each part that is classified as an arm, we estimate a hand at both endpoints of the arm, as illustrated in figure 4(d). Estimated detections at the wrong endpoints are rejected using the elimination and tracking described in the next sections.

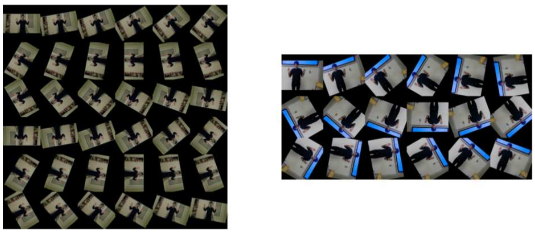


Figure 3: Illustration of the rotation of our images in order to detect hands in any orientation. Left: step size is 10° per rotation. Right: step size is 20° per rotation.

3.4 Elimination

After the above-mentioned steps, a large amount of hand detections is obtained, as seen in figure 5(a). The task of this elimination stage is to reject non-hand detections and to cluster overlapping detections. The output of this elimination operation is a reduced number of hand candidates as shown in figure 5(b). In our elimination process we apply the following steps:

- Remove hand detections which have an insufficient number of skin pixels, using the same skin detection algorithm as described in the previous step.
- Remove hand detections which have a divergent size with respect to the size of the face.
- Cluster overlapping detections based on their overlap and distance between their centers.
- Reduce the contribution of clusters that coincide with the face. We noticed that a face is often detected by the hand model. Only eliminating these detections is not a viable option since persons can hold their hands in front of the face. Therefore we reduce the score of those overlapping clusters by a predefined factor to minimize the impact.
- Remove hand detections which are too far from the predicted location by the Kalman trackers.

In the elimination step, we reduced the number of hand detections. Finally we classify the remaining detections in a left and right detection using the Kalman tracker information as explained in the next section.

3.5 Tracking

Our tracking stage is one of the most important contributions in order to improve the detection results. This is realized by steering the detections based on previous detections using a Kalman filter (Kalman, 1960). This mathematical filter is used to predict the position of the hands, which is needed when a detection is missing due to e.g. occlusions. A second advantage of

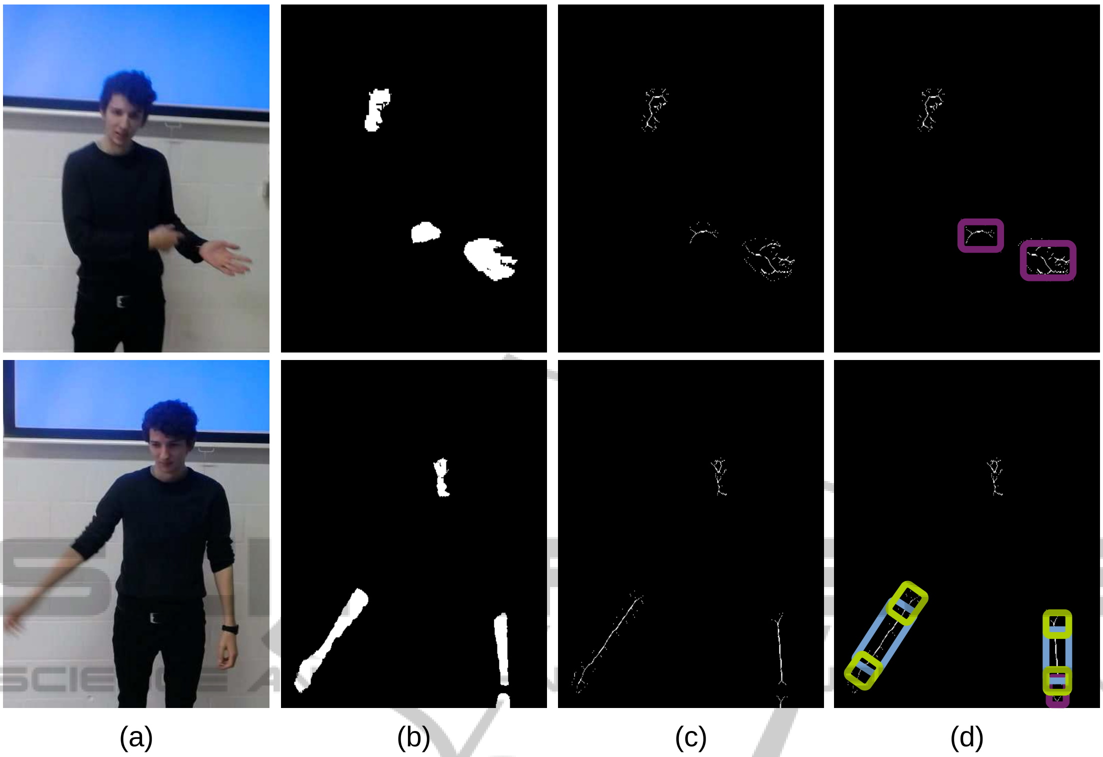


Figure 4: From left to right: original image(a); binary image based on skin segmentation(b); skeletonization(c); arm and hand estimation(d). Purple boxes illustrate the hand classifications, blue boxes the arm detections and green boxes the estimated hands at the endpoints of the arm.

using a Kalman filter is that the noise on the measured position of the detections is filtered out, resulting in more stable detections. For each torso detection we define two Kalman trackers: one for the left hand and one for the right hand in order to track each hand individually. We use a Kalman filter with the following state vector and update matrix, assuming a constant velocity motion model:

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix} \qquad A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where x and y are the position of the hand and v_x and v_y are the velocity of the hand. For each of the remaining clusters, as described in the previous section, we calculate the cost, based on the distance, to assign them to one of the Kalman trackers. By choosing the cluster with the lowest cost, we select the best candidate for each tracker.

To summarize this section we give an overview of our contributions as compared to the approach of (Mittal et al., 2011):

- Reduced computational footprint of our algorithm by avoiding both super-pixel calculation and the validation of the context model without loss in accuracy.

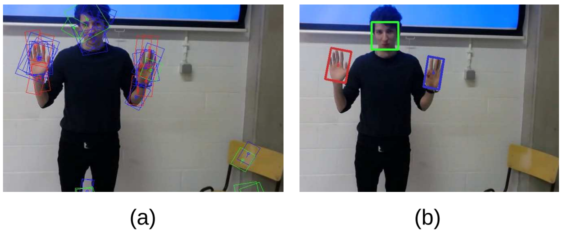


Figure 5: Left: large amount of detections before elimination; Right: Final detections after elimination step.

- Reduced search space by using a human-torso detector and only searching for hands in a region around the torso detection. This resulted in a reduced computational time and it reduced the number of false detections.
- Skin based detection is performed even when no face is detected, resulting in more detection candidates.
- Elimination of false-detections using the size of the face.
- Kalman tracker for both left and right hand that belongs to each torso detection.

4 SEMI-AUTOMATIC ANALYSIS

As mentioned before, we aim to develop a framework that achieves a detection rate up to 100%. Obviously it is unfeasible to develop an algorithm that achieves perfect accuracy on each dataset. Therefore we expanded our hand detection framework with a generic mechanism that allows for manual intervention resulting in a much higher accuracy. The key idea is that when the confidence drops under a specific (user-defined) threshold, our algorithm requests manual input. The user then has to manually annotate the missing detection. Relying only on the detection score results in a too large amount of manual interventions. To overcome this, we also take into account the distance between a detection and the predicted position (coming from the Kalman trackers). The formula of the confidence score is shown in equation 2:

$$M = \alpha log(D_{max} - D) + \beta S_i \qquad (2)$$

where:

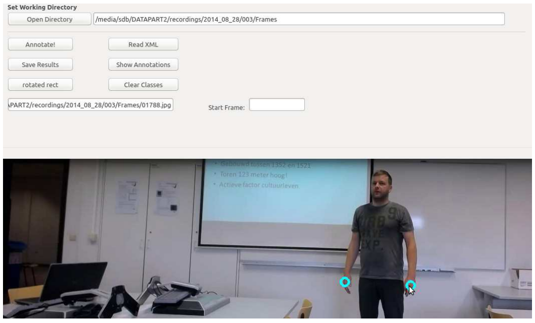
$$D = \begin{cases} D_{max} - 1, & \text{if } d(C_i, C_{i-1}) \geq D_{max} \\ d(C_i, C_{i-1}), & \text{otherwise} \end{cases}$$

D_{max} stands for the maximum allowed distance between the current detection and a detection in the previous frame, C_i and C_{i-1} define respectively the center of the current and the previous detection. α and β are used to change the weight of the distance and detection score. In our experiments, we empirically determined the optimal value of those parameters: α = 0.5 and β = 1.0.

The general concept of this approach is that a detection is likely to be valid if either the distance to the predicted location (based on previous detections) is low or if the detection score is high. If this value is below a user-defined threshold, manual input is requested. Thus by varying this threshold we can change the amount of manual interventions from zero (fully automatic detection) up to the number necessary to achieve full accuracy ((semi-)automatic detection). As illustrated in figure 6, the user is requested to manually annotate the missing detections when confidence score M is below a certain threshold. After this manual intervention the state vector of the corresponding kalman tracker is reset, thus resulting in a stable reference point for further detections.

5 EXPERIMENTAL RESULTS

As mentioned in the introduction, we validate our hand detection framework using a data set of recordings. First we introduce our datasets, next we discuss the accuracy of our framework compared to other techniques.



/media/sdb/DATAPART2/recordings/2014_08_28/003/Frames

PART2/recordings/2014_08_28/003/Frames/01788.jpg

Figure 6: Interface for manual intervention in which one can manually annotate the detection items.

5.1 Dataset

During our research we noticed that it is very hard to find video material containing hand annotations for each frame. In (Mittal et al., 2011) a dataset of annotated movie frames is presented. Unfortunately, the available frames are not consecutive, which makes them unsuitable for our approach, designed for a sequence of frames. We also examined some video recordings from the MPI archive[3], but those were annotated in terms of gestures (start and endpoint of the gesture) and contain no additional information of hand locations.

To overcome the lack of fully annotated video material, we set up a series of experiments. In each experiment a mobile eye-tracker was used to record the field of view of the test person. This eye-tracker records images at a resolution of 1280 x 720 pixels. In the first experiment two persons stood face-to-face at a distance of 3 meters from each other. The person who wore the eye-tracker was told to look attentively at the interlocutor while this person made movements with his hands. The second experiment was performed in a more natural setting. In this experiment, a powerpoint presentation was given with the spectator wearing a mobile eye-tracker used as recording device. An illustration of these is given in figure 7.

For each experiment we manually annotated left and right hand in more than 500 consecutive frames. This results in a reference dataset of more than 2000 annotated hand instances which can be used as reference dataset for benchmark tests. The annotation consist of a bounding rectangle oriented with respect to the wrist. Since it is hard to find publicly available hand-annotated video material, we made our dataset

[3] http://corpus1.mpi.nl

Figure 7: Left image is a frame from our first dataset, right image is a frame from the second dataset.

publicly available[4] for other researchers.

5.2 Results

To validate our framework, we have performed a series of experiments. First we tested our hand detection algorithm without tracking of the hands nor manual intervention. We did this experiment on both our own datasets and one publicly available dataset: the '5-signers' dataset. Examples of the detections on those datasets are shown in figure 8 and 9 respectively. This is a collection of non-consecutive frames from five news sequences (39 frames each) with different signers (Buehler et al., 2008). The validation is done using the F-measure:

$$F = \frac{2TP}{2TP+FP+FN} \tag{3}$$

In each frame of our datasets one person and two hands are visible. Since our framework was designed to detect two hands for each torso instance, the number of false positives (FP) and false negatives (FN) are equal, hence the F-measure is reduced to the precision. A hand detection is considered valid if it is within half hand width from the ground-truth location of the hand. We compare the results to the performance of two state of the art techniques. The publicly available hand detection algorithm of (Mittal et al., 2011) was used in which we use the two best detection scores as candidates for left and right hand. We also compare to the pose estimation proposal of (Yang and Ramanan, 2011) in which we classify the outermost bounding boxes of the arms as hands.

Our algorithm performs better than the other techniques in terms of accuracy. We outperfom the pose estimation technique, although a note on the bad performace of the approach of (Yang and Ramanan, 2011) should be made. The detection code we have used was developed to detect poses of persons from head to foot, whereas in the images of Dataset 1 the legs of the person are not completely visible as shown in the left part of figure 7. The results of this comparison is shown in table 2. Next, we compared our hand detection algorithm with tracking of the hands to the other techniques. We did those experiments on our own datasets, since we need sequences of frames. It is clear that the accuracy increases significantly when the tracking is applied, as shown in the right column of table 2.

Table 2: Accuracy of our hand detection algorithm compared to other techniques. Dataset 1 & 2 contains 1000 annotated hand-instances each, the '5-Signer' dataset contains 390 hand-instances.

	Mittal	Yang	**Ours**	**Ours incl. tracking**
Dataset 1	85%	24.2%	**83.4%**	**88.2%**
Dataset 2	48.9%	46.5%	**52.9%**	**65.3%**
5-Signers	77.6%	n.a.	**81.1%**	**n.a.**

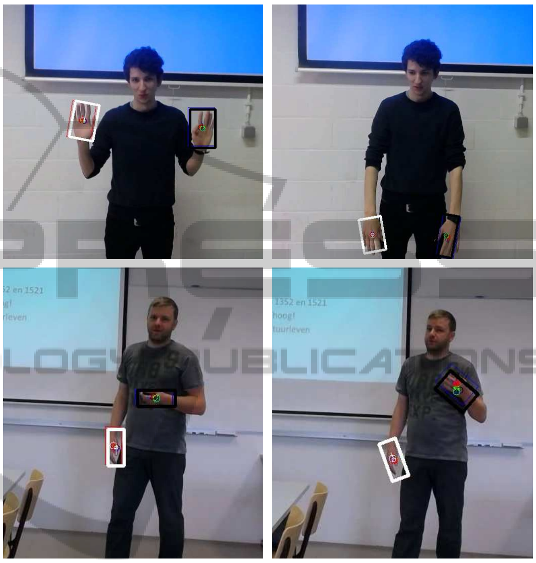


Figure 8: Examples of hand detections on our own recorded datasets. Top row are images from our first dataset, bottom row are images from our second dataset.

We also compared the execution speed of our algorithm, as shown in table 3. It is clear that the execution time of our algorithm is drastically lower compared to the other techniques on the same hardware (Intel Xeon E5645). Our approach is much faster compared to the work of (Mittal et al., 2011) since amongst others we no longer depend on the superpixel calculation. We also outperform the computational cost of (Yang and Ramanan, 2011) by a factor of 3.

Furthermore we present the extensive results of our (semi-)automatic approach on both our own datasets as shown in figure 10. In this graph we plot the accuracy in function of the number of manual in-

[4] http://www.eavise.be/insightout/Datasets/

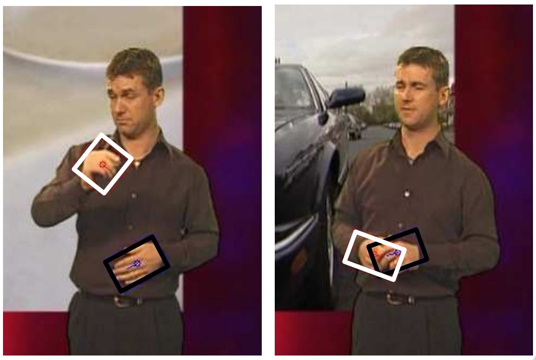


Figure 9: Examples of hand detections on the 5-Signer dataset.

Table 3: Execution times per frame averaged over all frames.

	Mittal	Yang	**Ours**
Avg time/frame	293.33 s	113 s	**36.67 s**

terventions expressed in a percent of the numbers of frames in the set. As mentioned before, by thresholding the result of equation 2, we can change the amount of necesary manual interventions. It is obvious that a higher amount of manual interventions results in a higher accuracy. We should also note the improvement in accuracy between no manual intervention and the lowest amount of manual interventions. For Dataset 1, the accuracy increases form 90% to 93% at the cost of only 7 manual interventions, Dataset 2 on the other hand has an accuracy improvement of 12% at the cost of only 14 manual interventions. We observe that each manual intervention restarts the tracker such that the hands in the following frames are again detected automatically.

6 CONCLUSION

We present a novel approach for the detection of human hands in real-life 2D-image sequences. We used the work of (Mittal et al., 2011) as a baseline and extended this approach in order to improve the accuracy and to lower the computation cost. First we use a torso detector to reduce the search area, next we use a face detector whose information is used to reject wrong hand detections. Furthermore we use an advanced hand and arm detection mechanism, based on skin detection, to detect hands in images where motion blur occurs. Finally, an evaluation scheme is used to reject wrong hand detections and a advanced tracking mechanism for left and right hand is introduced to steer the detections based on previous frames. We report good accuracy as compared to state-of-the-art techniques while the computation cost is drastically less.

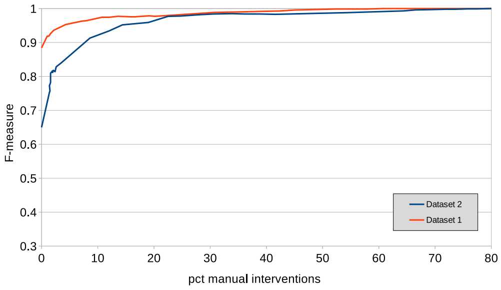


Figure 10: Result of our (semi-)automatic approach in which accuracy is improved by manual interventions.

In order to further improve the accuracy, we expanded our hand detection framework with a generic mechanism that finds the optimal places to ask for manual intervention resulting in a much higher accuracy with minimal manual effort. By calculating a score based on the detection score and distance to the predicted detection, we measure the reliability of detection. By thresholding this value, we can change the amount of manual interventions.

The validation of our approach was done using a series of datasets. We used two own recorded datasets, which we made publicly available, and one pre-existing dataset. We report good accuracy on all datasets outperforming the other techniques in both accuracy and execution time.

Our future work concentrates on further reducing the computational cost of the hand detection algorithm. Furthermore we will work on the integration of the eye gaze data. Using such an approach enables the automatic analysis of mobile eye-tracker data in terms of visual fixations on hands.

ACKNOWLEDGEMENTS

This work is partially funded by KU Leuven via the projects Cametron and InSight Out. We also thank Raphael Den Dooven for his contributions.

REFERENCES

Al Moubayed, S., Edlund, J., and Gustafson, J. (2013). Analysis of gaze and speech patterns in three-party quiz game interaction. In *Interspeech 2013*.

Bo, N., Dailey, M. N., and Uyyanonvara, B. (2007). Robust hand tracking in low-resolution video sequences. In *Proc of the third conference on IASTED International Conference: Advances in Computer Science and Technology*, pages 228–233, Anaheim, CA, USA.

Brône, G. and Oben, B. (2014). Insight interaction. A multimodal and multifocal dialogue corpus. *In Language Resources and Evaluation*.

Buehler, P., Everingham, M., Huttenlocher, D., and Zisserman, A. (2008). Long term arm and hand tracking for continuous sign language tv broadcasts. In *Proceedings of the British Machine Vision Conference*, pages 110.1–110.10. BMVA Press.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

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 the 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

Dubout, C. and Fleuret, F. (2012). Exact acceleration of linear object detectors. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 301–311.

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 Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Gebre, B. G., Wittenburg, P., and Lenkiewicz, P. (2012). Towards automatic gesture stroke detection. In *the Eighth International Conference on Language Resources and Evaluation*, pages 231–235.

Jokinen, K. (2010). Non-verbal signals for turn-taking and feedback. In *Proc. of the Seventh International Conference on Language Resources and Evaluation*.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. In *Transaction of the ASME Journal of Basic Engineering*, volume 82, pages 35–45.

Karlinsky, L., Dinerstein, M., Harari, D., and Ullman, S. (2010). The chains model for detecting parts by their context. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 25–32.

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

N. A. Abdul Rahim, C. W. Kit, J. S. (2006). RGB-H-CbCr skin colour model for human face detection. In *MMU International Symposium on Information and Communications Technologies (M2USIC)*, Petaling Jaya, Malaysia.

Pfister, T., Charles, J., Everingham, M., and Zisserman, A. (2012). Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *British Machine Vision Conference*.

Spruyt, V., Ledda, A., and Philips, W. (2013). Real-time, long-term hand tracking with unsupervised initialization. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3730–3734. IEEE.

Van den Bergh, M. and Van Gool, L. (2011). Combining rgb and tof cameras for real-time 3d hand gesture interaction. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV)*, WACV '11, pages 66–72, Washington, DC, USA. IEEE Computer Society.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. pages 511–518. IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Wang, R. Y. and Popović, J. (2009). Real-time hand-tracking with a color glove. In *ACM SIGGRAPH 2009 Papers*, pages 63:1–63:8.

Williams, G., Bregler, C., Hackney, P., Rosenthal, S., Mcdowall, I., and Smolskiy, K. (2008). Body signature recognition.

Wu, Y., Liu, Q., and Huang, T. S. (2000). An adaptive self-organizing color segmentation algorithm with application to robust real-time human hand localization. In *in Proc. of Asian Conference on Computer Vision*, pages 1106–1111.

Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE.