Semi-automatic annotation of eye-tracking recordings in terms of human torso, face and hands

Stijn De Beugher, Geert Brône, and Toon Goedemé

1 Introduction/Related Work

Research on interactive communication increasingly focuses on the role of eye gaze as an important signal in interaction management, reference and grounding (see [1] for an overview). Interlocutors may use eye gaze as a means to take, hold or give the floor in conversation (turn management), to refer to objects or persons in the conversational space (gaze cueing) or to give and elicit feedback (grounding). The use of unobtrusive eye-tracking technology (like eye-tracking glasses or table-top systems) has proven to be an invaluable resource for obtaining detailed information on the distribution of visual attention of multiple participants simultaneously ([3], [4], [2]). One of the key challenges in the use of mobile eye-tracking technology, however, resides in the processing and annotation of the obtained data stream. An example of an automatic annotation algorithm can be found in our previous work, see [5] for more details. In this paper we further extend our semi-automatic hand annotation algorithm [7], to apply it in the processing of mobile eye-tracking recordings. We integrate the detections of human face, body and hands with gaze data and produce ELAN compatible annotation files.

2 Our Approach

As mentioned above we present the integration of a semi-automatic annotation tool and the gaze data of mobile eye-tracker recordings in order to reduce the manual annotation cost. The annotation classes we tackle are: human torsos, faces and hands. The detection of the human torso and faces is built on our previous work [6]. In [7] we reduced the computational cost of those algorithms by utilizing the Kalman-tracker predictions to reduce the search area of both face and torso. We also implemented an extension to the face detection to left, right and frontal view. The detection of human hands on the other hand is a novel implementation of an accurate

Stijn De Beugher
EAVISE - KU Leuven, Belgium e-mail: stijn.debeugher@kuleuven.be

Geert Brône
MIDI Research Group - KU Leuven, Belgium e-mail: geert.brone@arts.kuleuven.be

Toon Goedemé
EAVISE - KU Leuven, Belgium e-mail: tgoedeme@esat.kuleuven.be

Published at SAGA Workshop 2015, proceedings archived under DOI: 10.2390/biecoll-saga2015_0.
Individual article archived under DOI: 10.2390/biecoll-saga2015_1

segmentation combined with advanced tracking mechanisms as well as a validation of human poses. The foundation of our hand detection algorithm is a highly accurate skin segmentation [8] in combination with Kalman-trackers for both hands and arms. Finally we validate hand candidates against a probability map of possible hand/arm positions with respect to the human pose. We embedded those algorithms in a semi-automatic tool, which calculates the confidence of the hand detections. This confidence is a combination of the distance with respect to the detections in the previous frames and the result of the validation against the probability maps. If the confidence drops below a certain threshold, our automatic analysis is halted and the user is asked for manual correction. This threshold is determined empirically using several eye-tracking experiments, but can be adapted by the user making the system more or less strict. After this intervention, our system automatically continues processing the remaining frames. Using such an approach results in a highly accurate system at a minimal cost of manual interventions. For technical details see [7].

As explained above, our semi-automatic system processes each frame captured by a mobile eye-tracker and searches for faces, human torso and hands. Next we map the gaze coordinates of each frame on top of our detections. Our system automatically calculates whether the gaze coordinates overlap with one of our detection classes. The integration of our semi-automatic detections and the gaze data is shown in figure 1. The red dot illustrated the gaze cursor while the rectangles illustrate the detection class that overlaps with the gaze cursor. The green rectangle corresponds to a face detection, while the yellow and blue rectangles corresponds to respectively right and left hand. We applied our algorithm on two different sequences. Each sequence contains images of one person in front of the participant wearing the eye-tracker. The current implementation of our algorithm supports only a single person in the video, however the software is written to support multiple persons in the future. In a next step, our system clusters consecutive frames in which the gaze cursor overlaps with the same detection class. When the length of the cluster is larger than a user defined threshold (standard value of a visual fixation is 150 ms), the cluster is stored as a valid annotation. We assign the class label as annotation value. Finally we export these data to a file that is compatible with annotation tools such as ELAN, making our tool integratable with existing annotations.

3 Results

Although the main contribution of this paper is the integration of semi-automatic hand detection with gaze data, we first present the accuracy results of the hand annotations. We validate our approach on a dataset containing 4000 hand labels and achieve an average accuracy of around 90% at a cost of only 1.7% of manual annotations. The average processing time per frame (1280×720) is around 150ms and includes face, torso and hand detection. Our approach is substantially faster compared to our previous research [6] in which more than 30 seconds were needed to perform the same detections.

Published at SAGA Workshop 2015, proceedings archived under DOI: 10.2390/biecoll-saga2015_0.
Individual article archived under DOI: 10.2390/biecoll-saga2015_1

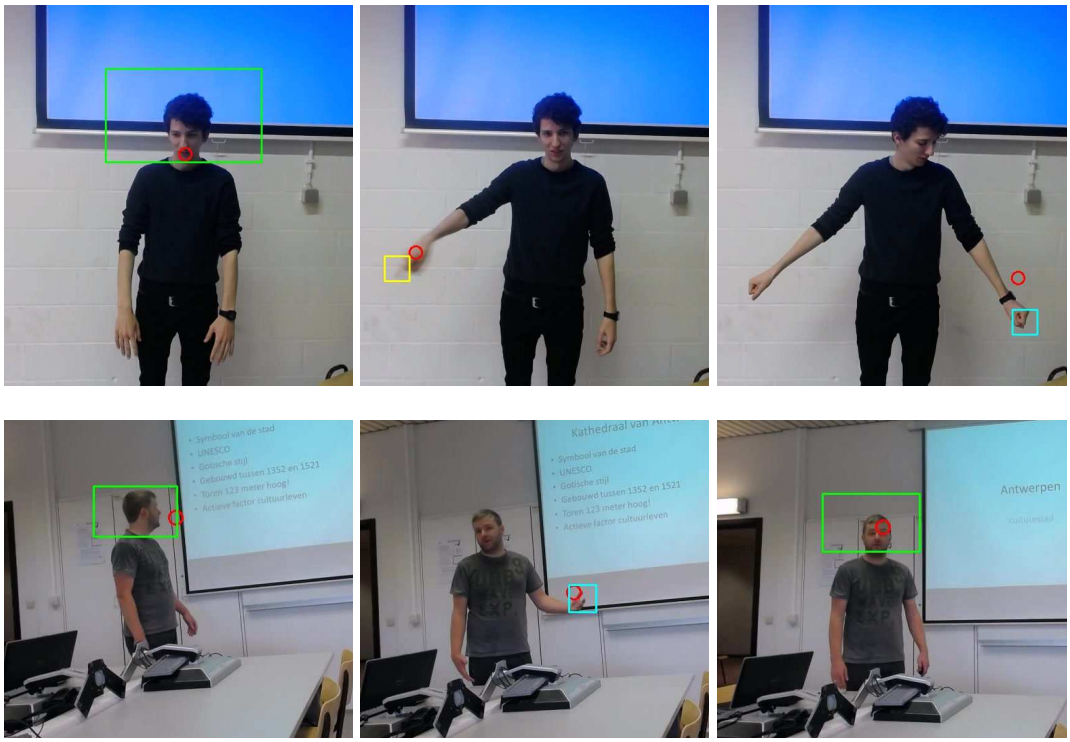

Fig. 1 Results on two sequences of our dataset. Red dot indicates the gaze cursor, coloured rectangles illustrate the detection class that overlaps with the gaze cursor: e.g. face (green), left hand (blue) and right hand (yellow).

Next to the accuracy of the hand detections, we also present the accuracy of our total system: face-, person- and hand detections and the integration of the gaze data. We used our tool to analyze a video sequence of approximately 1m30s and mapped the gaze data on top of the detections. Next we exported those results into an ELAN compatible format. We removed all the labels for validation and asked a participant to manually assign a label to each annotation. To validate our system we applied a statistical analysis, of which the results can be found in table 1. This table reveals the high accuracy of our system making it applicable in real-life situations. A final note should be made to the processing time: the semi-automatic analysis of the video took 5 minutes, while the manual labeling took 20 minutes. The software tool developed in this work will be made publicly available.

Table 1 Statistical analysis of semi-automatic analysis vs manual analysis.

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha
automatic vs manual analysis	95.6%	0.904	0.904	0.904

Published at SAGA Workshop 2015, proceedings archived under DOI: 10.2390/biecoll-saga2015_0.
Individual article archived under DOI: 10.2390/biecoll-saga2015_1

4 Conclusion

In this paper we proposed the integration of a novel semi-automatic hand annotation tool [7] and mobile eye-tracking recordings. Our semi automatic annotation tools detects human faces, bodies and hands in images with a high accuracy (around 90%) at the cost of only 1.7% of manual interventions at a speed of 6.5 frames per second. Next we integrate those detections with the gaze data to automatically decide to which part of the body a person is looking at. A comparison of our semi-automatic approach and a fully manual annotation reveals that our system is highly accurate (we scored 0.904 on a Cohen's Kappa test) and that our system is at least 4 times faster than manual analysis.

References

1. Rossano, F (2012). Gaze in Conversation. The Handbook of Conversation Analysis 308-329.
2. Holler, J Kendrick K. H. (2015). Unaddressed participants gaze in multi-person interaction: optimizing recipiency. Frontiers in Psychology 6-98.
3. Jokinen, K (2010). Non-verbal signals for turn-taking & feedback. In Proc. of 7th International Conference on Language Resources & Evaluation.
4. Oertel C, Salvi G (2013). A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue. Accepted for publication In Proc. of the 15th ACM International Conference on Multimodal Interaction ICMI. Syndney, Australia.
5. De Beugher S, Brône G, Goedemé T (2013). Object recognition and person detection for mobile eye-tracking research. A case study with real-life customer journeys. In Proc. of the First International Workshop on Solutions for Automatic Gaze Data Analysis 2013 (SAGA 2013)
6. De Beugher S, Brône G, Goedemé T (2015). Semi-automatic hand detection - a case study on real life mobile eye-tracker data. In Proc.of VISAPP, pages 121-129.
7. De Beugher S, Brône G, Goedemé T (2015) Fast and accurate semi-automatic hand annotation for human-human interaction analysis. To be published.
8. N. A. Abdul Rahim, C. W. Kit,J. See (2006). RGB-H-CbCr skin colour model for human face detection. In Proc. of M2USIC.

Published at SAGA Workshop 2015, proceedings archived under DOI: 10.2390/biecoll-saga2015_0.
Individual article archived under DOI: 10.2390/biecoll-saga2015_1