Towards automatic annotation of communicative gesturing

Kristiina Jokinen
University of Tartu
Estonia
kristiina.jokinen@ut.ee

Graham Wilcock
University of Helsinki
Finland
graham.wilcock@helsinki.fi

Abstract

We report on-going work on automatic annotation of head and hand gestures in videos of conversational interaction. The Anvil annotation tool was extended by two plugins for automatic face and hand tracking. The results of automatic annotation are compared with the human annotations on the same data.

1 Introduction

Hand and head movements are important in human communication as they not only accompany speech to emphasize the message, but also coordinate and control the interaction. However, video analysis of human behaviour is a slow and resource-consuming procedure even by trained annotators using tools such as Anvil (Kipp 2001). There is an urgent need for more advanced tools to speed up the process by performing higher-level annotation functions automatically.

We use two Anvil plugins, a face tracker (Jongejan 2012) and a hand tracker (Saatmann 2014), that automatically create annotations for head and hand movements. Objects are recognized based on visual features such as colour and texture, and Haar-liked digital image features, using OpenCV framework. Motion trajectories are estimated by calculating the mean velocity and acceleration during the time span of a set of frames (we experimented with 7 frames as more than 10 makes the algorithm insensitive for quick, short movements). Movement annotations with respect to velocity and acceleration are marked on the appropriate Anvil track, to indicate the movement and its start and stop. The interface has controls for minimum saturation threshold and for how many frames to skip (Figure 1).

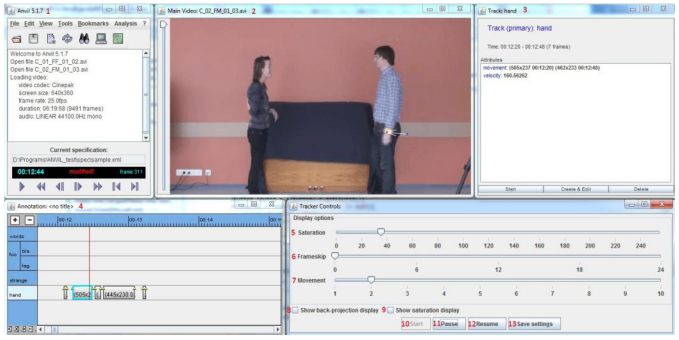

Figure 1 Anvil interface of the new hand tracker plugin.

2 Comparison of human and automatic annotations

Compared with human annotation the trackers are good at detecting some movements but prone to mis-detecting other movements. Problems occurred e.g. when the hue of the hands was similar to the background colour, or if the direction of the movement is reversed quickly, so that the time span is not long enough to detect velocity up to the thresholds (short head movements). Acceleration annotation did not recognize movements if they start and stop slowly. Changing the detection threshold can improve results, but is a trade-off as it prevents small movements being detected. However, the plugins will be of great help in multimodal analysis. Using the plugins reduces the time spent on annotating these movements, which in turn results annotations in increased productivity.

Proceedings of the 25th International Conference on Computational Linguistics, pages 124–125,
Dublin, Ireland, August 23-29 2014.

Here we present a more detailed analysis of the human and automatic annotations with reference to face tracking. The annotated hand and head movements are listed in Table 1. From the collected data we used four sample videos, each about six minutes long, altogether 45 303 frames. Table 2 shows the number of elements automatically recognized using velocity and acceleration, with precision scores, i.e. manually annotated gestures correctly recognized by the automatic annotation.

Head movements		
Nod down	Backward	Waggle
Nod up	Forward	Shake
Turn sideways	Tilt	Other

Hand movements
Both
Single
Complex
Other

Table 1. Annotation features for head and hand movements.

Gesture	Manual annotation	Velocity	Acceleration
NodDown	149	110 (74%)	108 (72%)
NodUp	42	15 (36%)	27 (64%)
TurnSide	40	29 (73%)	27 (68%)
HeadBackward	27	18 (67%)	14 (52%)
HeadForward	21	17 (81%)	18 (86%)
Tilt	57	35 (61%)	29 (51%)
Waggle	12	11 (92%)	8 (67%)
HeadOther	3	2 (67%)	1 (33%)
Total	**351**	**237 (73%)**	**232 (66%)**

Table 2. Manual and automatic head movement annotations for 4 videos. Precision: Velocity 73%, Acceleration 66%

Figure 2 shows two examples of the annotation results on the Anvil annotation board, one where the face tracker recognized head movements appropriately, and one where the face tracker "invented" movements which the human annotator does not recognize as communicative gestures.

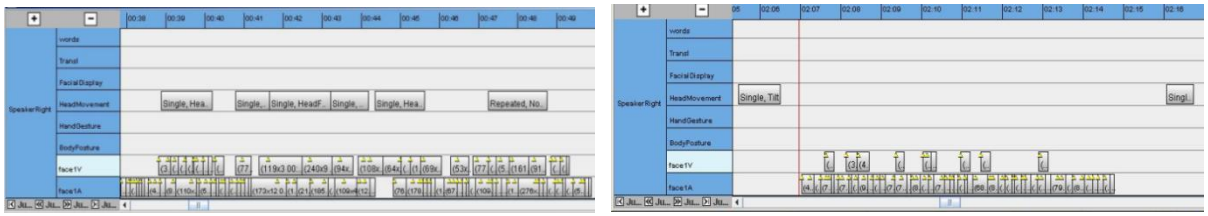

Figure 2. Face tracker detecting manual annotation categories (left) and inventing face movements (right).

3 Future work

Following the work outlined in Jokinen and Scherer (2012), we will compare the top-down linguistic-pragmatic analysis of movements with the bottom-up signal-level observations. We will also use a machine-learning approach to analyse if there are any systematics with the problematic cases. We may also explore if a recognized movement can be automatically interpreted with respect to communicative intentions. In human-robot interaction, the automatic gesture recognition model can be used to study the robot's understanding of the situation and of human control gestures, cf. Han et al. (2012).

References

Han, J., Campbell, N., Jokinen, K. and Wilcock, G. (2012). Investigating the use of non-verbal cues in human-robot interaction with a Nao robot, in *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, 679-683.

Jokinen, K. and Scherer S. (2012). Embodied Communicative Activity in Cooperative Conversational Interactions - studies in Visual Interaction Management. *Acta Polytechnica Hungarica.* 9(1), pp. 19-40.

Jongejan, B. (2012) Automatic annotation of face velocity and acceleration in Anvil. *Proceedings of the Language Resources and Evaluation Conference (LREC-2012). Istanbul, Turkey*.

Kipp, M. (2001). Anvil – A generic annotation tool for multimodal dialogue. *Proceedings of the Seventh European Conference on Speech Communication and Technology, pp. 1367-1370.*

Saatmann, P. (2014). Experiments With Hand-tracking Algorithm in Video Conversations. *Proceedings of the 5th Nordic Symposium on Multimodal Communication.*