Automatic identification of head movements in video-recorded conversations: can words help?

Patrizia Paggio
University of Copenhagen
University of Malta
paggio@hum.ku.dk

Costanza Navarretta
University of Copenhagen
costanza@hum.ku.dk

Bart Jongejan
University of Copenhagen
bartj@hum.ku.dk

1 Introduction and background

Head movements are the most frequent gestures in face-to-face communication, and important for feedback giving (Allwood, 1988; Yngve, 1970; Duncan, 1972), and turn management (McClave, 2000).Their automatic recognition has been addressed by many multimodal communication researchers (Heylen et al., 2007; Paggio and Navarretta, 2011; Morency et al., 2007).

The method for automatic head movement annotation described in this paper is implemented as a plugin to the freely available multimodal annotation tool ANVIL (Kipp, 2004), using OpenCV (Bradski and Koehler, 2008), combined with a command line script that performs a number of file transformations and invokes the LibSVM software (Chang and Lin, 2011) to train and test a support vector classifier. Successively, the script produces a new annotation in ANVIL containing the learned head movements. The present method builds on (Jongejan, 2012) by adding jerk to the movement features and by applying machine learning. In this paper we also conduct a statistical analysis of the distribution of words in the annotated data to understand if word features could be used to improve the learning model.

Research aimed at the automatic recognition of head movements, especially nods and shakes, has addressed the issue in essentially two different ways. Thus a number of studies use data in which the face, or a part of it, has been tracked via various devices and typically train HMM models on such data (Kapoor and Picard, 2001; Tan and Rong, 2003; Wei et al., 2013). The accuracy reported i these studies is in the range 75-89%.

Other studies, on the contrary, try to identify head movements from raw video material using computer video techniques (Zhao et al., 2012; Morency et al., 2005). Different results are obtained depending on a number of factors such as video quality, lighting conditions, whether the movements are naturally occurring or rehearsed. The best results so far are probably those in (Morency et al., 2007), where an LDCRF model achieves an accuracy from 65% to 75% for a false positive rate of 20-30% and outperforms earlier SVM and HMM models.

Our work belongs to the latter strand of research in that we also work with raw video data.

2 Movement features

Three time-related derivatives with respect to the changing position of the face are used in this work as features for the identification of head movements: velocity, acceleration and jerk. Velocity is change of position per unit of time, acceleration is change of velocity per unit of time, and jerk is change of acceleration per unit of time. We expect that a sequence of frames for which jerk has a high value in the horizontal or vertical direction will correspond to the most effortful part of the head movement, often called *stroke* (Kendon, 2004).

3 Data, test setup, and results

The data come from the Danish NOMCO (Paggio et al., 2010), a video-recorded corpus of conversational interactions with many different annotation layers (Paggio and Navarretta, 2016), including type of head movement (nods, turns. etc).

For this work, two videos in which one of the participants is the same were selected at random, and only the head movements performed by this one participant are considered. One video is used for training, and the other for testing. In both videos, OpenCV is used to analyse each frame for the x and y coordinates of the participants's head, and based on these coordinates velocity, acceleration and jerk measures are calculated for each

Proceedings of the 6th Workshop on Vision and Language, pages 40–42,
Valencia, Spain, April 4, 2017. ©2017 Association for Computational Linguistics

Category	true	false
movement	29,980	11,960
non-movement	235,640	108,420
sum	265,620	120,380

Table 1: Distribution of true and false move and non-move sequences in milliseconds.

frame and added to the video annotation. In the video used for training, each frame is added a boolean feature indicating presence or absence of head movement in the manual annotation.

A first inspection of the classification results showed that in several cases the classifier detected sequences of movement interrupted by empty frames, where the manual annotation consisted of longer spans of uninterrupted movement. Therefore, empty spans (*margins*) of varying length were considered part of the movement annotation in the subsequent experiments, all performed with SVM. In all experiments, using all three movement features together yield the best results. When margin = 2 the ratio true positive/true negative is maximal. A maximum accuracy of 68%, however, is reached for a much higher value of the margin, 17 frames, or 0.68 seconds. For comparison, a baseline model always selecting non-movement would reach an accuracy of 64%. Counts for true and false movement and non-movement sequences detected by the classifier are shown in Table (1).

Even though we can do better than the baseline, the accuracy is still not adequate. Considering the fact that the annotators who created the gold standard had access to the audio channel when they identified the head movements, it is worth considering whether word features could be used to train more sophisticated and accurate models.

	true	false
words	0.58	0.46
no words (incl. filled pauses)	0.42	0.54
words (incl. filled pauses)	0.75	0.73
no words	0.25	0.27
filled pauses and fb words	0.07	0.05
other words and no words	0.93	0.95
stressed words	0.31	0.25
unstressed words and no words	0.69	0.75

Table 2: Proportions of different word and no word categories in true and false movement sequences

	true	false
words	0.36	0.57
no words (incl. filled pauses)	0.64	0.43
words (incl. filled pauses)	0.56	0.76
no words	0.44	0.24
filled pauses and fb words	0.04	0.04
other words and no words	0.96	0.96
stressed words	0.20	0.28
unstressed words and no words	0.80	0.72

Table 3: Proportions of different word and no word categories in true and false non-movement sequences

4 Head movements and words

The relation between head movements and words was investigated by looking at how different kinds of words are distributed over sequences of movement vs non-movements. We thus considered distributions where the word category includes only real words, also filled pauses, only filled pauses and feedback words, and finally only stressed words. In all cases, we are only looking at the speech stream of the person performing the movement. The last two distributions show the least interesting effects. Thus, feedback words have almost equal, and very low, probability to occur in movement and non-movement sequences. In the case of stressed words, we see that their probability of occurring with movement is slightly higher than with non movement (31% vs 20%). If we look at the distribution of all words vs no words including filled pauses, we see that words have a 58% probability of occurring with movement, as opposed to a only 36% probability of occurring with non-movement. Finally, if we take words including filled pauses against no words, the probability of word occurrence with movement is 75% vs 56% with non-movement. Thus, distinguishing between real words and no words including filled pauses has the potential to differentiate best between presence and absence of movement in that we see that in this case the mutual proportion between word and no words goes in opposite directions depending on the sequence type. The differences in the distribution in this case are significant on a chi-square test in both movement and non-movement sequences. All the probabilities are summed up in Tables (2) and (3) .

To conclude, we have presented an approach where an SVM classifier is trained to recognise movement sequences based on velocity, acceleration, and jerk. A preliminary investigation of the overlap between temporal sequences classified as either movement or non-movement and the speech stream of the person performing the gesture shows that using word features may help increase the accuracy of the model, which is now 68%.

References

Sames Al Moubayed, Malek Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, J-C Martin, Stanislav Ondas, Catherine Pelachaud, Jérôme Urbain, and Mehmet Yilmaz. 2009. Generating robot/agent backchannels during a storytelling experiment. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3749–3754. IEEE.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation*, 41(3–4):273–287.

Jens Allwood. 1988. The Structure of Dialog. In Martin M. Taylor, Franoise Neél, and Don G. Bouwhuis, editors, *Structure of Multimodal Dialog II*, pages 3–24. John Benjamins, Amsterdam.

G. Bradski and A. Koehler. 2008. *Learning OpenCV: Computer Vision with the OpenCV Linbrary*. O'Reilly.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. In *ACM Transactions on Intelligent Systems and Technology*.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud. 2007. Searching for prototypical facial feedback signals. In *Proceedings of 7th International Conference on Intelligent Virtual Agents*, pages 147–153.

Kristiina Jokinen and Graham Wilcock. 2014. Automatic and manual annotations in first encounter dialogues. In *Human Language Technologies - The Baltic Perspective: Proceedings of the 6th International Conference Baltic HLT 2014*, volume 268 of *Frontiers in Artificial Intelligence and Applications*, pages 175–178.

Bart Jongejan, 2012. *Automatic annotation of head velocity and acceleration in Anvil*, pages 201–208. European language resources distribution agency, 5.

Ashish Kapoor and Rosalind W. Picard. 2001. A real-time head nod and shake detector. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, PUI '01, pages 1–5, New York, NY, USA. ACM.

Adam Kendon. 2004. *Gesture*. Cambridge University Press.

Michael Kipp. 2004. *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Evelyn McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. 2005. Contextual recognition of head gestures. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

P. Paggio and C. Navarretta. 2011. Head movements, facial expressions and feedback in danish first encounters interactions: A culture-specific analysis. In Constantine Stephanidis, editor, *Universal Access in Human-Computer Interaction- Users Diversity. 6th International Conference. UAHCI 2011, Held as Part of HCI International 2011*, number 6766 in LNCS, pages 583–690, Orlando Florida. Springer Verlag.

P. Paggio and C. Navarretta. 2016. The Danish NOMCO corpus multimodal interaction in first acquaintance conversations. *International Journal of Language Resources and Evaluation*, pages 1–32.

P. Paggio, J. Allwood, E. Ahlsén, K. Jokinen, and C. Navarretta. 2010. The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

W. Tan and G. Rong. 2003. A real-time head nod and shake detector using hmms. *Expert Systems with Applications*, 25(3):461–466.

Haolin Wei, Patricia Scanlon, Yingbo Li, David S Monaghan, and Noel E O'Connor. 2013. Real-time head nod and shake detection for continuous human affect recognition. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE.

Victor Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–578.

Z. Zhao, Y. Wang, and S. Fu. 2012. Head movement recognition based on lucas-kanade algorithm. In *Computer Science Service System (CSSS), 2012 International Conference on*, pages 2303–2306, Aug.