Exploring Annotation of Head Gesture Forms in Spontaneous Human Interaction

Spyros Kousidis (spyros.kousidis@uni-bielefeld.de)
Dialogue Systems Group, Bielefeld University

Zofia Malisz (zofia.malisz@uni-bielefeld.de)
Phonetics and Phonology Work Group, Bielefeld University

Petra Wagner (petra.wagner@uni-bielefeld.de)
Phonetics and Phonology Work Group, Bielefeld University

David Schlangen (david.schlangen@uni-bielefeld.de)
Dialogue Systems Group, Bielefeld University

Abstract

Face-to-face interaction is characterised by head gestures that vary greatly in form and function. We present on-going exploratory work in characterising the form of these gestures. In particular, we define a kinematic annotation scheme and compute various agreement measures among two trained annotators. Gesture type mismatches among annotators are compared against kinematic characteristics of head gesture classes derived from motion capture data.

Keywords: Multimodal interaction; Head gestures; Annotation;

Introduction

Head gestures are known to play an important role in face-to-face communication with distinct semantic, discourse and interactive functions (Hadar, Steiner, Grant, & Rose, 1983; McClave, 2000). Mapping the form and function of head gestures is desirable both for understanding human communication as well as in the context of ECAs (Heylen et al., 2011). Developing generative models for ECAs depends on the availability of annotated corpora of multimodal interactions (Lee, Wang, & Marsella, 2010).

However, head gesture annotation is costly as it requires frame-by-frame analysis of video data (Allwood & Cerrato, 2003; Poggi, D'Errico, & Vincze, 2010), a fact that also inhibits statistical analysis over large data sets. Although this limitation could be overcome by automatically detecting gestures, existing detection algorithms focus on a restricted set of portrayed gestures (typically nods and shakes) (Benoit & Caplier, 2005; Morency, Sidner, Lee, & Darrell, 2005; Gunes & Pantic, 2010) while natural interaction shows a greater kinematic variety in head movement (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007).

There are several approaches to coding gesture types which, depending on the study, are based on either the form or the function of head gestures (Heylen, 2008). Definitions of form are typically verbal descriptions of motions (Allwood et al., 2007) which do not capture the actual variability of head movements and rotations. In (Sargin, Yemez, Erzin, & Tekalp, 2008), head movement patterns were learned entirely from data but the number of patterns had to be defined arbitrarily and the information on motion included only rotations and no lateral movement. The work presented here evaluates a purely kinematic scheme using inter-annotator agreement and motion-capture data acquired from high quality in-lab recordings of multimodal dyadic interaction.

Data Collection

The material is a subset (4 dialogues, 60 minutes total in 8 videos) of the *Dream Apartment Corpus*, a collection of dyadic interactions recorded in the MIntLAB (Kousidis, Pfeiffer, Malisz, Wagner, & Schlangen, 2012). Subjects are situated in a comfortable laboratory setting while a rich multimodal recording is acquired. The recording includes HD quality audio and video tracks, motion capture (Microsoft Kinect[1]) as well as head, eye and gaze tracking data (Seeing Machines Facelab[2]). Subjects do not need to be fitted with any equipment in order to be recorded or tracked. This reduces the level of invasiveness, thereby allowing for a higher degree of naturalness in interaction.

The participants are given a task to jointly design an apartment, given a large amount of money is available in order to cover the costs of purchase, furnishing and decoration. The purpose of this task is to elicit spontaneous interaction as a negotiation evolves, with frequent occurrences of iconic and deictic gestures with the head or hands.

The tracking data for the head posture includes a 3D position vector, a quaternion expressing the head orientation and a simple tracking confidence measure. There are 60 frames per second of tracking data.

Annotation Procedure

The report on the annotation procedure below follows the recomendations of Bayerl and Paul (2011). Each video was annotated by two annotators, out of three in total, whose expertise level can be described as "schema developers". Annotation was performed in ELAN (Brugman & Russel, 2004) using a front close-up view of the subject, without audio. The goal was to annotate "any communicative head gesture" which meant any head movement that might be perceived as a signal, excluding inertial movement or movement caused by body posture shift or articulatory movements.

[1] http://www.microsoft.com/en-us/kinectforwindows/

[2] Seeing Machines www.seeingmachines.com

The annotation scheme used(Table 1) is a superset of the one in (Włodarczak, Buschmeier, Malisz, Kopp, & Wagner, 2012) and also similar to that of (Paggio & Navarretta, 2011). Each label is defined purely kinematically and associated with one of six axes of motion. X, Y and Z denote movement of the head left-right, up-down and front-back, respectively. Pitch, Yaw and Roll denote, respectively, rotations of the head up-down, left-right horizontally, and left-right vertically (as in "leaning" the head).

Some gestures are implicitly cyclic (nod, shake, bobble) but were also annotated when only "half" a cycle was present (e.g. nodding downwards without pulling back up). Finally, gestures frequently appear in connected sequences. These were separated only if there was a perceivable "gap" between them, otherwise they were assigned sequences of labels.

Table 1: Head gesture Inventory.

Label	Description	Axis
Nod	Rotation down-up	Pitch
Jerk	'Inverted nod', head upwards	Pitch
Tilt	'Sideways nod'	Roll
Shake	Rotation left-right horizontally	Yaw
Pro	Pushing the head forward	Z
Retr	Pulling the head back	Z
Turn	Rotation left OR right	Yaw
Bobble	Shaking by tilting left-right	Roll
Slide	Sideways movement(no rotation)	X
Shift	Repeated slides left-right	X
Waggle	Irregular connected movement	

Agreement Measures

Since the annotation schema is exploratory, the purpose of computing an agreement measure is to evaluate the schema itself rather than reach high agreement, which in itself is not a reliable measure of correctness (Passonneau, Habash, & Rambow, 2006). We are using mostly percentage measures, as they are more informative, despite the fact that they can be arbitrarily high, compared to stricter measures such as Cohen's kappa. (Bayerl & Paul, 2011).

Event Agreement expresses whether annotators recognised the same events as communicative head gestures. It equals twice the number of annotated intervals that overlap, divided by the sum of all intervals from both annotators. Intervals can be counted more than once in the denominator if they "participate" in more than one overlapping interval. In other words, the error is halved, since neither annotator is "correct". Park, Mohammadi, Artstein, and Morency (2012) present a similar approach which yields slightly lower scores.

Duration Agreement equals twice the sum of durations of the overlaps, divided by the sum of all interval durations. This is equivalent to "time slicing" in (Park et al., 2012) and expresses, jointly, the agreement in marking onsets and offsets.

Label Agreement equals twice the matching unique labels divided by the total number of unique labels. As a second, more strict measure, we use the *Levenshtein distance* between the two sets of labels (Levenshtein, 1966); it minimises the "edit distance" to transform one set to the other, using insertions, deletions and substitutions. All agreement measures yield values between 0 and 1 and are expressed as percentages.

Tracking data analysis

The tracking data is analysed using a collection of python packages for scientific computing provided by EPD[3]. We compute "energy signatures" or "profiles" for each gesture label from the data, in order to explore whether (a) the kinematic properties match the axes defined in Table 1 and (b) confusion in labels occurs due to proximity in the signatures. At this time, only an energy profile is computed while other properties such as *direction* and *periodicity* will be considered in the future.

Energy profiles for gesture instances are calculated as follows. The displacement (or rotation) for each axis is smoothed using the Savitzki-Golay algorithm (Savitzky & Golay, 1964) which also gives a velocity (or radial velocity) vector. The sum of squares of the velocity yields the total energy for each axis. In order to compute an energy percentile per axis, we take into account the ratio of mass to the moment of inertia of the human head (Yoganandan, Pintar, Zhang, & Baisden, 2009). Finally, in order to calculate a profile for each gesture type, we use gestures with (a) a single matching label for both annotators, (b) an average tracking confidence of at least 60% and (c) a length of 200ms or more. Gesture type profiles are simply the mean or median 6-dimensional vectors from all instance profiles per type. As a result of the above constraints, sample size is reduced and an energy profile could be computed for only a subset of the label inventory.

Results and Discussion

In total 3055 events with 4117 labels were analysed. "Waggle" was dropped due to low sample size. The event agreement score (77%) indicates that the instruction to "annotate all communicative gestures" allows for consistent identification of relevant events, despite the relative difficulty in noticing some gestures unless slow-motion playback is used. Conversely, about 20 % of all annotation intervals are "deletions", i.e. they are not matched by an interval from the second annotator. Figure 1 shows the percentage of deletions per label type (blue bars). A higher deletion rate indicates that a movement form is more likely to be (a) perceived as non-communicative, or (b) too subtle or difficult to recognise. It is also possible that annotators are biased towards seeking those gestures that occur more frequently (nod, turn, shake, tilt).

Duration agreement yields a similar score (79%) indicating consistency among annotators in marking gesture boundaries. For reference, a typical (single) head gesture lasts less

[3]Enthought Python Distribution. www.enthought.com

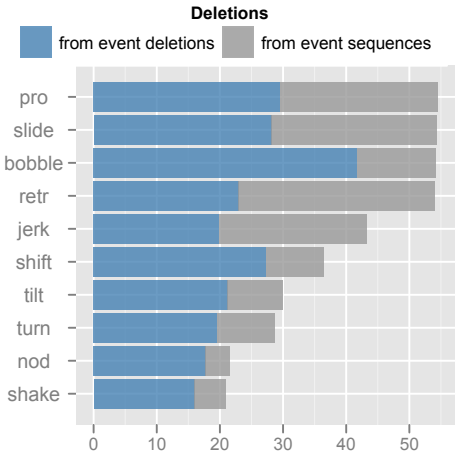

Figure 1: Deletions per label type: Event deletions (blue) and label deletions from event-sequences (grey)

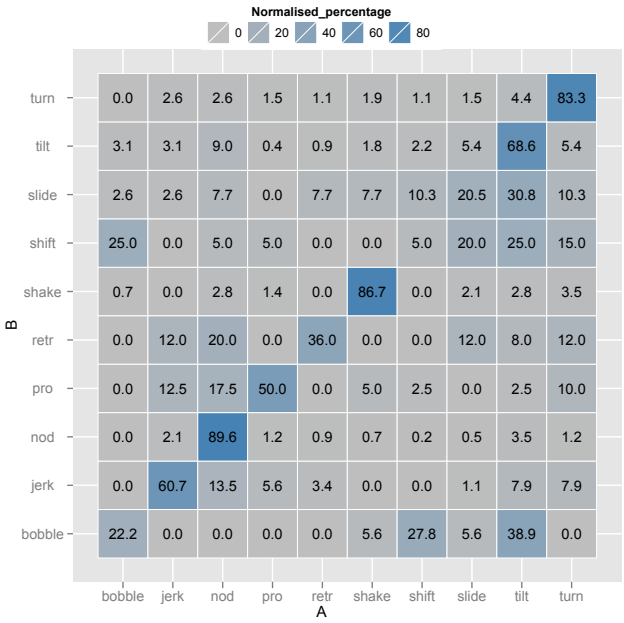

Figure 2: Confusion matrix between head gesture labels

than 500 ms, which translates the score to an average error of 50ms per boundary, but is in fact lower: connecting two single instances into a long sequence has a radical effect on the duration agreement score.

Both label agreement and Levenshtein distance yield 74%. 87% of all single labels are (exact) matches. When taking sequences into account, the percentage drops to 52% (44% for Levenshtein distance). This indicates that confusion occurs mostly in connected sequences of gestures rather than in single gestures. Adding the deletions from annotated gesture sequences (if a sequence is longer than the one it is compared against), raises the deletion rate to 29 %. The distribution of total deletions per label is shown in Figure 1 (grey bars).

Figure 2 shows the confusion between the labels, in both single gestures and sequences, if a definite match or mismatch can be found. Nods, shakes, turns and tilts are consistently agreed upon. Overall label agreement is high as a result of these four being the most common types. Shifts, slides and bobbles are confused with one another and with tilt; all these movement types share (theoretically) prominent axes (X and Roll). Interestingly, turns and shakes share the Yaw axis but are not confused; possibly because shakes are clearly cyclic while turns are uniform, one-direction movements. Retractions and protrusions are both confusable with nods and jerks but not with each other as they move in opposite directions. However jerks are often confused with nod upstrokes.

The energy profiles in Table 2 show that the gestures are not performed uniformly on one axis. However each gesture has a higher percentile than other gestures on its prominent axis. The high energy content on the translation axes is due to the head rotating around the neck rather than its center, hence any rotation is simultaneously a translation and vice versa.

Finally, the distances between gesture energy signatures in Figure 3 show that "opposite" pairs (nod-jerk or pro-retr)

Table 2: Energy profiles of head gestures types

Label	X	Y	Z	Pitch	Yaw	Roll
nod	6	21	35	**25**	10	4
jerk	7	21	43	**12**	10	7
turn	30	5	23	7	**32**	3
tilt	14	9	29	18	13	**17**
shake	19	6	24	15	**31**	5
retr	8	16	**40**	14	11	11
pro	4	12	**44**	27	6	7

have indeed similar signatures. Further information, such as direction, is used to distinguish them. Conversely, nods are "far" from shakes, hence these two can be distinguished fairly easily (by humans or algorithms). When no kinematic classifier (energy, periodicity, direction) can differentiate two gestures, they are often confused. The profile of tilts approximates the profile of nods on its prominent axis (Pitch), leading to some confusion between tilts and nods.

Conclusions

We have explored the validity of an extended head gesture annotation scheme that attempts to cover the kinematic variability of spontaneous face-to-face interaction. Annotator consistency in segmenting and labeling relevant communicative gestures was found, at least for the most common gestures. Tracking data shows that kinematic content maps intuitively to the expectation of what a gesture should look like. Energy signatures can distinguish between head gesture "classes" but more kinematic properties are needed to distinguish between gestures sharing prominent axes.

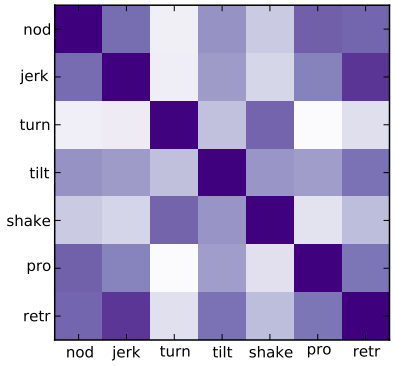

Figure 3: Distance between energy profiles. Darker is closer.

Acknowledgements

This research is partly supported by the Deutsche Forschungsgemeinschaft (DFG) in the CRC 673 "Alignment in Communication". The authors would like to thank Joanna Skubisz, for assisting with annotations.

References

Allwood, J., & Cerrato, L. (2003). A study of gestural feedback expressions. In *First nordic symposium on multimodal communication* (pp. 7–22).

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3), 273–287.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699–725.

Benoit, A., & Caplier, A. (2005). Head nods analysis: interpretation of non verbal communication gestures. In *IEEE intern. conf. on image processing, ICIP 2005.* (Vol. 3, pp. III–425).

Brugman, H., & Russel, A. (2004). Annotating multimedia/multi-modal resources with elan. In *Proc. of the 4th intern. conf. on language resources and evaluation (LREC)* (pp. 2065–2068).

Gunes, H., & Pantic, M. (2010). Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent virtual agents* (pp. 371–377).

Hadar, U., Steiner, T., Grant, E., & Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, *2*(1), 35–46.

Heylen, D. (2008). Listening heads. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling communication with robots and virtual humans.* Berlin: Springer.

Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., & Schröder, M. (2011). Generating listening behaviour. In P. Petta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-oriented systems: The Humaine handbook.* Berlin: Springer.

Kousidis, S., Pfeiffer, T., Malisz, Z., Wagner, P., & Schlangen, D. (2012). Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proc. of the interdisciplinary workshop on feedback behaviours in dialogue.*

Lee, J., Wang, Z., & Marsella, S. (2010). Evaluating models of speaker head nods for virtual agents. In *Proc. of the 9th intern. conf. on autonomous agents and multiagent systems* (pp. 1257–1264).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Forschungsbericht*(8), 707-710.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878.

Morency, L.-P., Sidner, C., Lee, C., & Darrell, T. (2005). Contextual recognition of head gestures. In *Proc. of the 7th intern. conf. on multimodal interfaces* (Vol. 4, pp. 18–24).

Paggio, P., & Navarretta, C. (2011). Learning to classify the feedback function of head movements in a danish corpus of first encounters. In *Talk given at the workshop on multimodal corpora at ICMI.*

Park, S., Mohammadi, G., Artstein, R., & Morency, L.-P. (2012). Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface. In *Proc. of the ACM multimedia 2012 workshop on crowdsourcing for multimedia* (pp. 29–34).

Passonneau, R., Habash, N., & Rambow, O. (2006). Inter-annotator agreement on a multilingual semantic annotation task. In *Proc. of the 5th intern. conf. on language resources and evaluation (LREC)* (pp. 1951–1956).

Poggi, I., D'Errico, F., & Vincze, L. (2010). Types of nods. the polysemy of a social signal. In *Proc. of the 7th intern. conf. on language resources and evaluation (LREC)* (pp. 19–21).

Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2008). Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *30*(8), 1330–1345.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, *36*(8), 1627–1639.

Włodarczak, M., Buschmeier, H., Malisz, Z., Kopp, S., & Wagner, P. (2012). Listener head gestures and verbal feedback expressions in a distraction task. In *Proc. of the interdisciplinary workshop on feedback behaviours in dialogue.*

Yoganandan, N., Pintar, F. A., Zhang, J., & Baisden, J. L. (2009). Physical properties of the human head: Mass, center of gravity and moment of inertia. *Journal of biomechanics*, *42*(9), 1177–1192.