A calibration-free head gesture recognition system with online capability

Nils-Christian Wöhler
Faculty of Technology
Bielefeld University, Germany

Ulf Großekathöfer
Ambient Intelligence
Bielefeld University, Germany

Angelika Dierker
Applied Informatics
Bielefeld University, Germany

Marc Hanheide
School of Computer Science
University of Birmingham, UK

Stefan Kopp
Sociable Agents
Bielefeld University, Germany

Thomas Hermann
SFB 673
Bielefeld University, Germany

E-mail: {ugrossek, adierker}@techfak.uni-bielefeld.de

Abstract

In this paper, we present a calibration-free head gesture recognition system using a motion-sensor-based approach. For data acquisition we conducted a comprehensive study with 10 subjects. We analyzed the resulting head movement data with regard to separability and transferability to new subjects. Ordered means models (OMMs) were used for classification, since they provide an easy-to-use, fast, and stable approach to machine learning of time series. In result, we achieved classification rates of 85–95% for nodding, head shaking and tilting head gestures and good transferability. Finally, we show first promising attempts towards online recognition.

1. Introduction

Human head gestures are an important communicational cue. Munhall et. al. [5] showed that head movements even improve the perception of syllables in Japanese. Although there is no such evidence for other languages, it seems useful to equip virtual agents and social robots with appropriate head gestures [1] to make conversation with them more lifelike.

To equip an agent with head gestures performed at appropriate timing it is necessary to investigate these gestures during human-human interaction. For this research, typically video data are annotated manually. This laborious process can be dramatically facilitated by a combination of machine-learning methods with a tracking technique. Possible techniques are detailed by Vatavu et. al. [7] who claim that vision-based approaches (e.g. [4]) have the advantage that the measurements can be done unobtrusively but are dependent on constant lighting conditions and on a full view at the interlocutor's face. Moreover, high processing power is needed. For this reason, our approach is sensor-based. We use motion sensors mounted on the subject's head, which grant lighting independence and almost unrestricted mobility. The practical issue in this context is the easy applicability of our system without any calibration.

To examine the possibilities of such a sensor-based head gesture recognition system, we accomplished a comprehensive study with 10 subjects in comparatively natural setups. Thereby, we abdicated any sensor preparations and adjustments. Subsequently, we analyzed the recorded data with regard to separability and transferability, where we use a new approach to machine learning of time-series and sequences, the so-called ordered means models (OMMs). OMMs can be described as rigorously reduced versions of the well-known and widely-used hidden Markov models (HMMs) [6]. While achieving similar generalization properties, OMMs provide a high level of robustness in terms of fragmented or insufficient data and, additionally, need less computational power [2].

2. Ordered Means Models

Similar to HMMs, an OMM Ω can be characterized as a generative state-space model that emits a sequence of observation vectors $O = \mathbf{o}_1..\mathbf{o}_T$, $\mathbf{o}_t \in \mathbb{R}^d$ out of K hidden states. In opposite to HMMs, OMMs establish some restrictions: (i) OMMs are defined without any transition probabilities. Instead, each path $\mathbf{q}$ through the model, i.e. each combination of states, is equally likely. (ii) The emissions of each state are modeled as probability distributions $b_k(\mathbf{o}_t)$ and assumed to be Gaussian with $b_k(\mathbf{o}_t) = \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_k, \sigma)$. The standard deviation σ is identically in all states and used as a global

Figure 1. Scenario for data acquisition. Both interacting subjects are wearing a head motion sensor.

hyperparameter. (iii) The model topology of OMMs is similar to the so-called left-to-right topology known from HMMs where only self transitions and transitions to subsequent states are allowed. Applying these restrictions, the only parameters left are the location parameters of the emission densities $\boldsymbol{\mu}_k$. Therefore, an OMM is completely defined by a linear array of reference vectors $\Omega = [\boldsymbol{\mu}_1..\boldsymbol{\mu}_K]$. To estimate the parameters of an OMM by a set of observed example sequences $\mathbf{O} = \{O_1, .., O_N\}$, we use a Baum-Welch training procedure similar to HMMs. Note that with $\sigma \rightarrow 0$ the training changes towards the Viterbi algorithm [2].

To use OMMs for classification in a maximum likelihood framework, one model Ω_i is trained for each class i. An unknown sequence O then is assigned to the class $k = \arg\max_i p(O|\Omega_i)$, whose model yields the highest posterior probability (see Rabiner [6]). A more detailed introduction to OMMs and algorithmic details can be found in Großekathöfer et al. [2].

3. Dataset and Experiments

The experimental data was recorded using an Xsens MT9 inertial sensor[1] (using only the gyroscopes with 3 DoF rate-of-turn). With this sensor attached to the top of the subjects' heads, we recorded the head movements of 10 subjects during dyadic conversation (see Figure 1). All subjects were German native speakers and briefly informed about the purpose of the data acquisition. The conversation was video-taped by a scene camera and stopped when it became stagnant. We synchronized the sensor data with the scene camera video and annotated the head movements in ELAN [3]. All

[1] www.xsens.com

	Number of Events for					
S	Nod	Shake	Tilt	Look	length	dur
1	27	17	0	2	0.98s	11 m
2	22	14	0	2	1.1s	11 m
3	33	1	0	0	1.39s	15 m
4	22	1	4	3	1.09s	15 m
5	35	67	2	37	1.4s	33 m
6	27	38	6	26	1.2s	33 m
7	77	81	15	47	1.46s	27 m
8	15	16	1	49	1.04s	27 m
9	122	37	13	65	1.42s	43 m
10	67	33	3	79	1.17s	43 m
$\sum$	447	305	44	310		

Table 1. Number of head movement samples per subject. S: subject number, length: medium length of event (in seconds), dur: overall duration of the measurement (in minutes).

weak annotations (not reliably assignable gestures) were then left out and the head motion data was sliced into segments relating to the four most frequently occurring head movement annotations: *nod*, *shake*, *tilt* (occurred as a gesture of uncertainty) and *look* (sideways). Table 1 gives an overview of the resulting samples. All data were recorded at 33Hz and normalized to zero-mean and unit variance.

To estimate the accuracy of OMMs for head gesture recognition, we tested these four head movement classes with two different evaluations based on the obtained dataset. The first evaluation tested whether the classifier is suitable and robust for this kind of data. For this, we randomly partitioned all available data into equally sized training and test sets. In a second evaluation, we investigated the classifiers' transferability to new subjects. Thus, we used data captured from 9 subjects as training data while the data from the remaining subject was used as test data (test subject). We accomplished this evaluation for each subject. Furthermore, to analyze the mutual influence of head movement classes on the performance, we repeated both evaluations four times, each time with a different set of head gesture classes. Since *nod* and *shake* were the most frequently occurring head movements, every set included these two gestures: (a) *nod*, *shake* (b) *nod*, *shake*, *tilt* (c) *nod*, *shake*, *look* (d) *nod*, *shake*, *look*, *tilt*.

We applied a uniform procedure to all data sets: First, we estimated an appropriate hyperparameter for each method by the means of 5-fold cross valida-

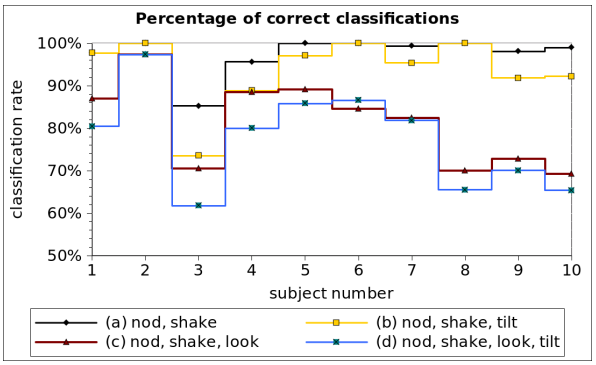

Figure 2. This figure shows the classification rates for all ten subjects from Evaluation 2. Note that the connection of the dots does not indicate interim values.

tion on the training data. We chose 8 different values for the number of OMM states K with $K \in \{5, 10, 15, 30, 45, 70, 90, 110\}$. The set of values for the global standard deviation σ was equal for all data sets with $\sigma \in \{0, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.5, 2, 3\}$. Subsequently, we took the best hyperparameters found in this process to train OMM classifiers with the complete training data set. To obtain the test set classification rate, we applied the resulting classifiers to the prepared test data set.

4. Results

The results of the first evaluation reveal classification success rates between 86.36% and 97.48% (see Table 2). The best rate was achieved when *nod* and *shake* were used as trained head gestures, whereas the lowest rate occurred when all four classes were used. For the case of three trained classes, there are two different results depending on the third added class: the addition of *tilt* results in a classification rate of 95.32%, while the addition of *look* results in a classification rate of 87.90%[2].

The classification rates from the second evaluation, which are weighted averages over all subjects, range from 75.95% to 98.40% (Table 2 last column). Here again, the best rate was achieved with two classes (*nod* and *shake*), and the lowest rate occurred when all four gesture classes were used. In order to examine the mutual influence of all four gesture classes, we generated an overall confusion matrix (see Table 3). This matrix is the sum of the confusion matrices of all 10 runs. Samples on the diagonal are classified correctly. While most samples for *nod*, *shake* and *tilt* are classified correctly, it can be observed that, for the *look* class, the number of correct

[2]Note that random classification yields 25% accuracy with 4 classes, whereas 2 classes reach 50% by chance.

	Classification rates	
Trained head gestures	Evaluation 1	Evaluation 2
(a) *nod*, *shake*	97.48%	98.40%
(b) *nod*, *shake*, *tilt*	95.32%	94.82%
(c) *nod*, *shake*, *look*	87.90%	79.49%
(d) *nod*, *shake*, *look*, *tilt*	86.36%	75.95%

Table 2. Classification rates from Evaluation 1 and 2.

real gesture	classified gesture				performance
	nod	*shake*	*tilt*	*look*	
nod	**405**	6	20	16	90.60%
shake	10	**260**	7	28	85.25%
tilt	2	4	**36**	2	81.82%
look	12	132	27	**139**	44.84%

Table 3. Confusion matrix from Evaluation 2 with all 4 gesture classes accumulated over all 10 subjects. Samples on the diagonal are classified correctly.

classified samples per class is considerably lower. More precisely, 132 of 310 *look* samples have been mistakenly accounted to the *shake* class.

Figure 2 shows the classification rates for the head gesture class sets per subject. The rate ranged from 61.76% to 100%, where the class set (a) with *nod* and *shake* achieved the best performance results again. Similarly to the first evaluation, the system reached very high accuracy with class set (b) for almost all subjects. The classification rates fall off for both class sets that include the *look* class. Subject 3 seems to provide the most difficult data for classification.

5. Online classification capability

To expand the proposed system to online classification some extensions have to be applied. First of all, to process a continuous head motion data stream from the sensor we partition the data via a sliding window approach into fragments. Additionally, we establish a rejection scheme in case no head gesture is performed: based on the posteriori probabilities we define thresholds by which, if under-run, classification is rejected.

Note that the combination of *sliding window* and *rejection scheme* imposes additional parameters. Namely, these are the sliding window's length w and overlap o, and a rejection threshold for each class. In preliminary studies we achieved promising results[3].

[3]cf. http://www.techfak.uni-bielefeld.de/ags/ami/research/hgr/

6. Discussion

The first evaluation tested whether the classifier is suitable and robust for this kind of problem. We found that all four head gesture classes are easily separable although the classification rate slightly decreases for the four classes case.

With the second evaluation we investigated the transferability to new subjects. We observed that all four classes are still easily separable. As a further result, we found good transferability to new subjects for class sets (a) and (b). For the class sets (c) and (d) we can not conclude stable transferability. However, with more than 75% performance, these classifiers still provide good hypotheses.

Overall, especially the *look* class seems to have a negative effect on the classification rate. This finding is further supported by the confusion matrix in Table 3. About half of the *look* gestures were classified as *shake* gestures. A likely reason for this might be the similarity of both movements. We assume the classifiers to assign *look* movements as fragmented *shake* gestures. Note that our dataset is biased in the number of examples per class. This is the result of our comparatively natural acquisition scenario. Instead, we could have asked subjects to perform the four gestures repeatedly but we assumed such resulting gestures much more artificial. We claim that our data acquisition is superior since the nativeness of the recorded gestures should be an advantage for online recognition scenarios.

7. Conclusion & Outlook

This paper introduces and evaluates a calibration-free OMM-based approach for head gesture recognition using 3 DoF gyroscope sensors attached to the subjects' heads. We argued that the benefits of this sensor-based approach compared to vision-based approaches are the independence from lighting conditions and a great mobility for the subject. Another big advantage of the proposed method is the absence of a time-consuming calibration. The sensor can be mounted to the head, is instantly ready to provide data and still yields very good results.

We conducted a comprehensive study with 10 subjects and analyzed the resulting data using OMMs with regard to separability and transferability to new subjects. Although the classification results decrease when the *look* is included in the analyzed class sets (because of the similarity of this gesture to the *shake* gesture) they still provide good hypotheses. For the sets of classes where the *look* gesture is not included, we showed very good results (>95% performance). Evaluation 2 indicates good transferability to new subjects. Finally, the system is capable to be used in online operation, thereby giving an automatic real-time annotation tool for application in human-computer interaction and interaction studies. In our ongoing research we focus on the automatic optimization of classification in online use and develop a more lightweight, wireless version of the sensor.

Acknowledgements

This research was supported by the German Research Foundation (Sonderforschungsbereich 673 *Alignment in Communication* and *Center of Excellence for Cognitive Interaction Technology*).

References

[1] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM New York, NY, USA, 1994.

[2] U. Großekathöfer, T. Lingner, H. Ritter, and P. Meinicke. What is a hidden markov model without transition probabilities? *Neural Computation*, 2010 (submitted).

[3] B. Hellwig and D. Uytvanck. EUDICO Linguistic Annotator (ELAN) Version 2.0. 2 manual. *Nijmegen-NL, Max Planck Institute for Psycholinguistics*, 2004.

[4] L. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environments. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002. Proceedings*, pages 390–395, 2002.

[5] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137, 2004.

[6] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[7] R. Vatavu, Ş. Pentiuc, and C. Chaillou. On natural gestures for interacting in virtual environments. *Advances in Electrical and Computer Engineering*, 24(5), 2005.