Automatic Annotation of Gestural Units in Spontaneous Face-to-Face Interaction

Simon Alexanderson
KTH – Speech, Music and Hearing
Lindstedtsv 24
100 44 Stockholm, Sweden
+46(8)790 6293
simonal@kth.se

David House
KTH – Speech, Music and Hearing
Lindstedtsv 24
100 44 Stockholm, Sweden
+46(8)790 7565
davidh@kth.se

Jonas Beskow
KTH – Speech, Music and Hearing
Lindstedtsv 24
100 44 Stockholm, Sweden
+46(8)790 8965
beskow@kth.se

ABSTRACT

Speech and gesture co-occur in spontaneous dialogue in a highly complex fashion. There is a large variability in the motion that people exhibit during a dialogue, and different kinds of motion occur during different states of the interaction. A wide range of multimodal interface applications, for example in the fields of virtual agents or social robots, can be envisioned where it is important to be able to automatically identify gestures that carry information and discriminate them from other types of motion. While it is easy for a human to distinguish and segment manual gestures from a flow of multimodal information, the same task is not trivial to perform for a machine. In this paper we present a method to automatically segment and label gestural units from a stream of 3D motion capture data.

The gestural flow is modeled with a 2-level Hierarchical Hidden Markov Model (HHMM) where the sub-states correspond to gesture phases. The model is trained based on labels of complete gesture units and self-adaptive manipulators. The model is tested and validated on two datasets differing in genre and in method of capturing motion, and outperforms a state-of-the-art SVM classifier on a publicly available dataset.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *motion*

General Terms

Measurement

Keywords

Gesture recognition; spontaneous dialogue; motion capture

MA3HMI'16 , November 16 2016, Tokyo, Japan

ACM 978-1-4503-4562-0/16/11…$15.00
DOI: http://dx.doi.org/10.1145/3011263.3011268

1. INTRODUCTION

Gestures form an integral part of human communication. In spoken conversation, co-speech gestures are temporally linked to the spoken signal. These gestures often carry complementary information or serve to highlight or emphasize important points.

Today, ever more accurate and affordable sensors and sophisticated computer vision techniques make it feasible to capture gesture data in great detail. This has led to a class of multimodal interface applications where body motion is used as an input modality allowing e.g. gesture control of appliances or computer games. In these applications, each particular gesture typically carries a well-defined meaning and may be seen as forming a gesture vocabulary that often has to be explicitly learnt by the user.

This stands in stark contrast to co-speech gestures, which are a natural, largely sub-conscious part of communication. These gestures, which frequently occur simultaneously with speech, generally cannot be as easily categorized, and will often not have a well-defined semantic interpretation. Being able to process this stream of motion is the key to a number of multimodal interface applications, such as virtual agents or social robots that are able to make sense of – and produce – natural gesturing behavior in order to increase effectiveness and smoothness of the interaction. This paper aims to provide a method for automatic labeling and classification of recorded material. Our work strives towards building agents and robots with natural gesturing and gesture-recognition capabilities, using data-driven approaches and large data-sets of multimodal information.

The problem that we target in this contribution is how to extract boundaries of gestural movements from a continuous motion stream, which is a prerequisite to any subsequent analysis and classification of gestures, regardless of whether the goal is gesture recognition or identification of units for gesture synthesis. We specifically focus on segmenting *gestural units*, which are defined as the periods of gesticulation between two rest positions [7]. A gesture unit may contain a single gesture phrase, or several gesture phrases after each other.

Segmenting gestural units is a non-trivial problem for several reasons. Gestures consists of multiple phases, one or several of which may contain a "hold", i.e. a segment where velocity is zero, but the complete sequence should still be identified as one unit, [8], [12], [7]. Another problem that arises is how to exclude other types of movements that also occur naturally such as those termed manipulators (e.g. face and hair touching) by [3].

During the past few years there has been an increasing interest from researchers in the psycholinguistics community to develop tools for automatic gesture annotation. Gesture annotation from video is a difficult and time consuming task which may take hours or days to perform per minute of data [17]. Also gesture annotation often suffers from lack of annotator agreement. Machine learning algorithms possibly can perform some of the annotation tasks automatically, and generate more consistent and repeatable results than manual annotation. Previous researchers in gesture unit segmentation have used Multi-Layer Perceptrons [20], Support Vector Machines [18], [10], [11], or Hidden Markov Models [16]. While the work of [18], [11] and [16] includes segmentation of both gesture units and phases (preparation, stroke, hold and retraction), other researches have focused solely on stroke [6] or hold detection [1]. A comparative overview of the domains and techniques used in these studies is given in [11].

An important factor for the performance of all machine-learning techniques on time-series data is the modelling of temporal dynamics. A common approach used in [20], [11], [16], is to concatenate observations in a time-slice window centered on each frame of interest. However, when used to model generic co-verbal gestures, this representation has several draw-backs. As there generally is a large temporal variation in different examples of the same gesture phase such as a hold, preparation or retraction, the saliency of the features tend to rapidly decrease from the center of the feature vector. This can lead to confusion between e.g. rest positions and stroke-holds. Also, this representation does not model transitions between gestural phases. As noted before, a gestural unit is the interval between two rest positions of the limbs. A model of valid transition-dynamics may be of great benefit for segmentation.

Another way of modelling temporal dynamics is to automatically find segment transition points by using thresholding techniques [21], [18]. [18] use abrupt changes in acceleration to find possible transition points between gestural phases. A possible problem with such approaches is to find good threshold values, especially for noisy sensors such as the Kinect where the true accelerations are hard to estimate.

A promising alternative is provided by the Hierarchical Hidden Markov model (HHMM) [3], [14]. HHMMs are useful for modelling hierarchical structures where the levels represent concepts on different time scales. An example application for HHMMs is speech recognition [15], where the higher levels represent words, and the lower levels represent phonemes or sub-phonemes. While HHMMs have been proven to work well for recognition of isolated gestures in human computer interaction [19] and musical gestures [5] they have not to our knowledge been used for unconstrained conversational gesticulation.

In this paper, we present a gesture unit segmentation system using HHMMs specially modelled for the domain of conversational co-speech gesticulation. We compare the results from our algorithm with the publicly available dataset (narrative storytelling) used in [20] and [11] and find that our method performs better than both methods in the gestures unit segmentation task. We also run the algorithm on a larger data-set consisting of spontaneous dialogue and find F-score values ranging between 0.80 and 0.94 for leave-one-out cross-subject experiment.

The main novel contribution of this paper lies in the design of the HHMM network specialized to detect and segment conversational gestures in spontaneous, unrestricted dialogue. The model is tested and validated on two datasets differing in genre and in their method of capturing motion. Section 2 of the paper presents and describes the Hierarchical Hidden Markov Model. Section 3 presents the datasets and the methodology used in the two experiments and reports the results of each experiment. Implications of the results are discussed in section 4, and conclusions and future work are presented in section 5.

2. HIERARCHICAL HIDDEN MARKOV MODEL

The HHMM is a hierarchical structure of hidden states, where each state in itself is a HHMM. States emitting observations are called production states. The rest are called internal states, and emit sequences of observations. The states are traversed recursively from top to bottom; when an internal state is activated, it makes a 'vertical' transition and directly activates one of its sub-level states. The process continues until the system reaches a production state and emits an observation. It then makes a 'horizontal' transition and continues until reaching an exit state, where control is given back to the higher level. All levels below an internal state thus have to finish before the internal state may continue transitioning horizontally.

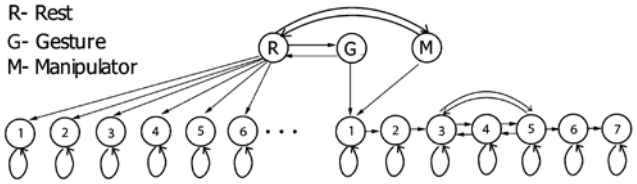

Figure 1. HHMM model with two levels. The top level contains internal states for the main categories: rest, co-speech gesture and manipulator. The bottom level contains production states. Exit states are omitted to avoid cluttering the diagram.

In our model, shown in Figure 1, the top level contains internal states representing a higher level motion state of the hand. We distinguish three main categories: resting, co-speech gestures and manipulators. The second level contains the production states emitting observations. The sub-states to the 'rest' state represent different rest positions, and may only self-transition or exit. The sub-states to the gesture and manipulator states represent motion phases during the excursion of the hand between two rest positions. These states are traversed in a left-to-right manner from 1 to 7, with possible back-transitions between states 3, 4 and 5. The transitions were designed so state 1 and 2 represent preparation, and state 6 and 7 retraction. The states 3, 4 and 5 model the core of the gesture unit. The connections between states 3, 4 and 5 add flexibility for gesture units comprised of several gesture phrases after each other. The number of sub states was chosen by manually testing different values. In this study, we found that 7 sub states for the gestures worked best, and we chose the same number of sub states for the manipulators and rest positions as well.

As can be seen in Figure 1, the high level states have no self-transitions, and there are no transitions between co-speech gestures and manipulators. This implies that a rest state may not change to a different rest state unless there is a gesture or a manipulator in between and that a gesture or a manipulator must be followed by a rest position.

2.1 Training

The model is trained by providing the top level labels to the network, and setting the second level states as missing. The parameters are learned with Maximum Likelihood Estimation using an expectation-maximization (EM) algorithm. We thus do not explicitly tell the system the underlying gesture phases, only that they should follow a specific pattern.

An important step in this process is model initialization, i.e. providing estimates for the mean and covariance for observations in each production state. We initialize each of the sub-states for the rest poses by clustering the data annotated as rest positions with a GMM. The sub states for the gesture prototypes are uniformly initialized with the mean and covariance for these labels respectively.

3. EXPERIMENTS

The 2-level HHMM described above was trained and tested using two independent data sets, thus comprising two separate experiments which will be described in this section. The data set used in the first experiment is the one constructed by Wagner et al. [20] and also employed by Madeo et al. [11]. The complete data set consists of features extracted from videos of seven storytelling sessions ranging from 36 to 61 seconds. In each session, a storyteller was asked to read a comic strip and then requested to retell the story standing in front of a Microsoft Kinect. Three different stories and three different users were included with one user retelling all three stories and two of the users telling two stories each. The data set has been generously made available through the UCI Machine Learning Repository [9]. Using this dataset allowed us to test the HHMM on a previously tested dataset and to compare and benchmark our results with the previous results.

The dataset used in the second experiment is taken from the Spontal corpus of spontaneous dialogue [2]. We selected three five-minute long dialogue sequences involving a total of six participants. We chose sequences which exhibited an abundance of gesture activity. This enabled us to test the HHMM on a larger dataset of unrestricted spontaneous dialogue which is the goal of this study and which we expected to be more challenging as it involves additional dialogue regulating gesture activity such as giving and seeking feedback and turn-taking.

Both experiments were performed using the Matlab implementation of BNT Toolkit [13].

3.1 Experiment 1: Storytelling

3.1.1 Method

For the first experiment, we used the same data set as that used in [11] to facilitate a comparison of the results. The database was recorded using Microsoft Kinect and each session contains approximately 1 minute of data. The data is annotated for gesture units and phases. [11] reports the results from gesture unit and phase segmentation using an SVM classifier and feature vectors consisting of the velocities and accelerations of the hand and wrist data concatenated over an 81 frame window centered on each frame of the test data. The recordings consisted of combinations of the three stories (1, 2 and 3) told by the three storytellers (A, B and C). However, the available data does not include the series B2 and C2. The tests in [11] were divided into three user dependent contexts: training A1 – test A2, training A1 – test A3 and training A1 (initial 70%) – test A1 (final 30%), and one user independent test: training A1 – test B1.

For our experiment we used a feature vector containing the positions and velocities (vector and scalar) of the hand markers for each frame. As a preprocessing step, we transformed the data to be relative to the spine position.

3.1.2 Results

Table 1 shows the performance of our method compared to the previous study. As can be seen in the table, our method outperforms the tests in [11] in all of the user dependent test. For the user independent test, training A1 – test B1, the training set was too small to generalize our model over the change in user gesturing style. However, as more user-independent training data was available in the data set, we complemented the comparison by performing a separate test where we trained our model on all stories from both of the other subjects (training A1, A2, A3, C1, C3 – test B1). As can be seen in the table, our method generated a high F-Score of 0.95 for this test.

Table 1. Test results for the storytelling data set. The table shows the precision, recall and F-score from our method, as well as the F-score reported in [11]

Train	Test	Precision	Recall	F-score HHMM	F-score Madeo et al. [11]
A1	A2	91.0	94.7	**0.926**	0.877
A1	A3	70.0	93.6	**0.801**	0.712
A1 70%	A1 30%	99.0	88.1	**0.937**	0.918
A1	B1	-	-	-	0.731
A1, A2, A3, C1,C3	B1	98.4	91.8	**0.950**	-

3.2 Experiment 2: Spontaneous Dialogue

3.2.1 Method

In this experiment we ran the algorithms on 6 dialog partners taken from the Swedish Spontal corpus [2]. The database contains a rich set of spontaneous dialogue between pairs of speakers, and is comprised of synchronized high-quality audio and video recordings as well as motion capture for body and head movements. During the recordings, the participants were seated in a sound studio and allowed to speak about any topic of their choice for 30 minutes. They remained in a seated position throughout the recording session.

We picked out 5-minute excerpts of three dialogues. The excerpts were selected in dialogues showing turn shifts and a variation of gestures. The three dialogues are referred to in this section as A, B, C and the six participants are labelled 1 and 2 in each dialogue. Figure 2 shows an example of two video frames taken from the data.

The gesture units were annotated into segments of gesture, rest, and non-gestural movement. As opposed to the previous experiment, we annotated and segmented each hand separately. This was done to provide a larger and more consistent dataset to the system, as gesturing styles may change depending on the dominance of the hands for different participants.

Figure 2. Frames from the spontal video corpus.

The Spontal motion data consist of 3D positions of the motion capture markers attached to each subject, sampled at a frame rate of 100 FPS. In this work we down-sampled the data to 33 FPS. This was done both for computational reasons (reduced memory size and computational times), and to obtain results more comparable to the experiment 1 which used Kinect. The marker set contains 12 markers placed on the upper body according to Figure 3. The data was prepared in the following way. First the data for each subject was transformed to a coordinate system with x-axis directed along the vector between the shoulder markers, y-axis towards world up. The origin was set to the mean position of hands as taken from a reference frame, where the subject had the hands in rest between the knees. For cross-subject training and testing purposes, the positions of the hand markers were normalized with the average total distance between the four arm markers, which roughly estimates the length of the arms. As a final step, we mirrored the right hand markers in the y-z plane.

Figure 3. Marker sets with 12 markers per subject.

For each frame we extracted a 9-dimentional feature vector consisting of the coordinates of the positions and velocities of the hand marker, the scalar velocity and acceleration, and the distance from the hand to the head. The process resulted in data sets containing 20000 samples (10000 for each hand), corresponding to 10 minutes of data, for each participant.

Instead of using a GMM for initializing rest states (as described in Section 2.1) we used a k-means algorithm and clustered on hand position. This was done as there was a massive overweight on rest poses at the knees, and the GMM failed to converge.

We then tested the segmentation algorithms for each participant in a "leave-one-out" fashion, i.e. using the data from all the other participants for training.

Table 2. Results from gesture unit segmentation using 'leave-one-out'

Test set	Precision	Recall	F-score
A1	0.89	0.89	0.89
A2	0.96	0.92	**0.94**
B1	0.94	0.90	0.92
B2	1.00	0.76	0.87
C1	0.96	0.88	0.92
C2	0.96	0.68	0.80

3.2.2 Results

The results of the experiment are presented in Table 2, showing the precision, recall and f-score values for gesture segmentation of both hands for each participant. As can be seen in the table, the f-scores ranged between 0.80 and 0.94. Analyzing the characteristics of the dialogues, dialogue A and B were relatively calm compared to dialog C. The poor recall from C2 came from our system classifying small gestures near the knees as rest segments. For example, some of these gestures were only performed with the fingers.

4. DISCUSSION

The aim of this study was to develop a method to automatically detect and segment conversational gesture units from motion data. The results from our experiments indicate that our HHMM network is a promising path to this goal. Some limitations of the method arise due to the fact that we only use one HHMM to model all gestural movements. As seen in the second experiment, the small gestures performed near the knees were incorrectly classified as rest positions. This indicates that our model is too general to handle significantly different types of gesture, and that an increased model complexity would be beneficial. A possible improvement may be to introduce different high-level gesture categories and model these with separate HHMMs. Another challenge is to provide a better model of the manipulators. Such movements were not included in the storytelling dataset, but occurred on a few occasions in the spontaneous dialogue dataset, all of which were incorrectly labeled as gesture units. Also in this case, adding categories for different manipulators may be a good approach to handle such cases.

5. CONCLUSIONS AND FUTURE WORK

In this study we developed a method targeted for segmentation of gesture units from spontaneous dialogue, which varies greatly in the amount and type of gestures exhibited. The method uses a Hierarchical Hidden Markov Model to model gesture dynamics. Our work strives towards building agents and robots with natural gesturing and gesture-recognition capabilities, using data-driven approaches and large data-sets of multimodal information. A key pre-requisite to accomplish this is the ability to bring order to the multimodal data-streams and to chunk up the data into segments for gesture recognition tasks or as possible building blocks for gesture synthesis. An important advantage of our method is the ability to cluster possible gesture-phases in an unsupervised way. In future work we aim to explore the possibilities to use these clusters for gesture synthesis.

6. ACKNOWLEDGMENTS

The work reported here is carried out within the projects: “Timing of intonation and gestures in spoken communication,” (P12-0634:1) funded by the Bank of Sweden Tercentenary Foundation, and “Large-scale massively multimodal modelling of non-verbal behaviour in spontaneous dialogue,” (VR 2010-4646) funded by Swedish Research Council.

7. REFERENCES

[1] Bryll, R., Quek, F., & Esposito, A. 2001. Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Communication.*

[2] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S. and House, D. 2010. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner and D. Tapias [Eds], *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Valetta, Malta: 2992−2995.

[3] Ekman, P. 2004. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation* (pp. 39-50). Springer Netherlands.

[4] Fine, S., Singer, Y., and Tishby, N. 1998. The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, 32(1), 41-62.

[5] Françoise, J., Caramiaux, B., & Bevilacqua, F. 2011. *Realtime segmentation and recognition of gestures using hierarchical markov models.* Mémoire de Master, Université Pierre et Marie Curie–Ircam.

[6] Gebre, B. G., Wittenburg, P., & Lenkiewicz, P. 2012. Towards automatic gesture stroke detection. In *LREC 2012: 8th International Conference on Language Resources and Evaluation* (pp. 231-235). European Language Resources Association.

[7] Kendon, A. 1980. *Gesticulation and speech: Two aspects of the process of utterance. The relationship of verbal and nonverbal communication*, 25, 207-227.

[8] Kita, S., Van Gijn, I., & Van der Hulst, H. 1997. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and sign language in human-computer interaction* (pp. 23-35). Springer Berlin Heidelberg.

[9] Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Madeo, R. C., Lima, C. A., & Peres, S. M. 2013. Gesture unit segmentation using support vector machines: segmenting gestures from rest positions. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 46-52). ACM.

[11] Madeo, R. C. B., Peres, S. M., and Lima, C. A. M. 2016. Gesture phase segmentation using support vector machines. *Expert Systems with Applications*, 56, 100-115.

[12] McNeill, D. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[13] Murphy, K.: Bayes Net Toolbox for Matlab. https://code.google.com/p/bnt/. Accessed 2015 April 10

[14] Murphy, K. P., and Paskin, M. A. 2002. Linear-time inference in hierarchical HMMs. *Advances in neural information processing systems*, 2, 833-840.

[15] Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., and Murphy, K. 2002, May. A coupled HMM for audio-visual speech recognition. In *Acoustics, Speech, and Signal Processing* (ICASSP), 2002 IEEE International Conference on (Vol. 2, pp. II-2013). IEEE.

[16] Okada, S., Bono, M., Takanashi, K., Sumi, Y., & Nitta, K. 2013. Context-based conversational hand gesture classification in narrative interaction. *In Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 303-310). ACM.

[17] Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X. F., Kirbas, C., ... & Ansari, R. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction* (TOCHI), 9(3), 171-193.

[18] Ramakrishnan, A. S., & Neff, M. (2013, May). Segmentation of hand gestures using motion capture data. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems* (pp. 1249-1250). International Foundation for Autonomous Agents and Multiagent Systems.

[19] Yin, Y., & Davis, R. W. (2014, July). Real-time continuous gesture recognition for natural Human-Computer Interaction. In *Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on* (pp. 113-120). IEEE.

[20] Wagner, P. K., Peres, S. M., Lima, C. A. M., Freitas, F. A., and Madeo, R. C. B. 2014. Gesture unit segmentation using spatial-temporal information and machine learning. In *Proceedings of 27th Florida artificial intelligence research society conference.*

[21] Wang, T. S., Shum, H. Y., Xu, Y. Q., and Zheng, N. N. 2001. Unsupervised analysis of human gestures. In *Advances in Multimedia Information Processing—PCM* 2001 (pp. 174-181). Springer Berlin Heidelberg.