Modeling Perception-Action Loops: Comparing Sequential Models with Frame-Based Classifiers

Alaeddine Mihoub, Gérard Bailly, Christian Wolf

Modeling Perception-Action Loops: Comparing Sequential Models with Frame-Based Classifiers

Alaeddine Mihoub
GIPSA-Lab & LIRIS
Grenoble, France
alaeddine.mihoub@gipsa-lab.fr

Gérard Bailly
GIPSA-Lab
Grenoble, France
gerard.bailly@gipsa-lab.fr

Christian Wolf
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205
F-69621, France
christian.wolf@liris.cnrs.fr

Abstract
Modeling multimodal perception-action loops in face-to-face interactions is a crucial step in the process of building sensory-motor behaviors for social robots or users-aware Embodied Conversational Agents (ECA). In this paper, we compare trainable behavioral models based on sequential models (HMMs) and classifiers (SVMs and Decision Trees) inherently inappropriate to model sequential aspects. These models aim at giving pertinent perception/action skills for robots in order to generate optimal actions given the perceived actions of others and joint goals. We applied these models to parallel speech and gaze data collected from interacting dyads. The challenge was to predict the gaze of one subject given the gaze of the interlocutor and the voice activity of both. We show that Incremental Discrete HMM (IDHMM) generally outperforms classifiers and that injecting input context in the modeling process significantly improves the performances of all algorithms.

Keywords
Social behavior model, HMMs, SVMs, cognitive state recognition, gaze generation

INTRODUCTION
The design of social robots/agents able to engage efficient and believable face-to-face conversations with human partners is still an open issue. Although this kind of communication is considered as one of the most basic and classic forms of communication in our daily life [23], it is a complex and sophisticated bi-directional multimodal phenomenon in which partners continually convey, perceive, interpret and react to the other person's verbal and co-verbal displays and signals [26]. Studies on human behavior has confirmed for instance that co-verbal features – such as body posture, arm/hand gestures , head movement, facial expressions, eye gaze– strongly participate in the encoding and decoding of linguistic, paralinguistic and non-linguistic information. Several researchers have notably claimed that these features are largely involved in maintaining mutual attention and social glue [18].

Human interactions are paced by multi-level perception-action loops [2]. Thus, social robots/agents aiming at monitoring a multimodal and natural communication should mimic the very aspects of this complex close-loop system. In concrete terms, the robot has to couple two principal tasks: (1) scene analysis and (2) behavior generation. A multimodal behavioral model is responsible for computing behavior generation given the scene analysis and the intended goals of the conversation.

Our goal is to train statistical multimodal behavioral model that learns by observation of human-human interactions i.e. that maps perception to action. In this context, we present and compare three different candidate models: the first one is based on Hidden Markov Models (HMMs) and models the evolution of joint perception/action features over time. The two others are standard classifiers (Support Vector Machines and Decision Trees) that perform direct mapping without any explicit sequential modeling.

The paper is organized as follows: The next section reviews the state-of-the art of trainable multimodal generation systems. The three models are introduced in section 3. Section 4 illustrates the application of our models on data collected in a previous experiment [1]. We analyze the impact of contextual data in section 5. Finally, we conclude in section 6.

RELATED WORK
The analysis of multi-party interaction is an interdisciplinary domain spanning research not only in signal and image processing but also in social and human science involving sociology, psychology and anthropology [24]. In recent years, it is becoming an attractive research area and there is an increasing awareness about its technological and scientific challenges. Actually, automatic conversation scene analysis copes with several issues, including turn taking, addressing, activity recognition, roles detection, degree of engagement or interest, state of mind, personal traits and dominance. Several computational models have been proposed to predict or generate observed multimodal human behavior.

For instance, Otsuka et al. [22] proposed a Dynamic Bayesian Network (DBN) to estimate addressing and turn taking ("who responds to whom and when?"). The DBN

framework is composed of three layers. The first one perceives speech and head gestures; the second layer generates gaze patterns while the third one estimates conversations regimes. While the first layer is observable, the others are latent and should be estimated. In order to recognize individual and group actions, Zhang et al. [30] suggested a two layered HMM. The first layer estimates personal actions taking as input raw audio-visual data. The second one infers group actions taking into account the estimations of the first layer. A Decision Tree is used in [3] for automatic role detection in multiparty conversations. Based mostly on acoustic features, the classifier assigns roles to each participant including effective participator, presenter, current information provider, and information consumer. In [13], Support Vectors Machines have been used to rate each person's dominance in multiparty interactions. The results showed that, while audio modality remains the most relevant, visual cues contribute in improving the discriminative power of the classifier. More complete reviews on models and issues related to nonverbal analysis of social interaction can be found in [10] [9][29].

For multimodal behavior generation, several platforms have been proposed for virtual agents and humanoid robots. Cassel et al. [6] notably developed the BEAT system ("Behavior Expression Animation Toolkit") which processes textual input and generates convenient and synchronized behaviors with speech such as intonation, eye gaze and iconic gestures. The synthesized nonverbal behavior is assigned on the basis of a contextual and linguistic analysis that relies on a set of rules inspired from research on conversational social human behavior. Later, Krenn [17] introduced the NECA project ("Net Environment for Embodied Emotional Conversational Agents") which aims to develop a platform for the implementation and the animation of conversational emotional agents for Web-based applications. This system hosts a complete scene generator and has the advantage of providing an ECA with communicative attitudes (e.g. head nodes, eye brow raising) as well as non communicative attitudes (e.g. moving/walking in the scene, physiological breathing). Another major contribution of the NECA platform is Gesticon [16]. It consists of repository of predefined co-verbal animations and gestures that can drive both virtual and physical agents. "MAX", the "Multimodal Assembly eXpert" developed by Kopp and colleagues [14], interacts with humans in a virtual reality environment and collaborates with them in order to achieve some tasks. MAX is able to ensure reactive and deliberative actions via synthetic speech, facial expressions, gaze, and gestures. Most mentioned platforms have many similarities: multimodal actions are selected, scheduled and integrated according to rules-based configurations. The SAIBA framework [15] has been developed to establish a unique platform, unify norms and accelerate advancements in the field. It is organized into three main components: "Intent planning", "Behavior planning" and "Behavior realization". It's worth noticing that SAIBA offers only a general framework for building multimodal behavioral models. In fact, the modeling within each component and its internal processing is treated as a "black box" and it is to researchers to fill the boxes by specifying their own models. One missing aspect of SAIBA is the perception dimension. In [26] a specific representation of perceptual cues was introduced to fill this gap. Many systems have adopted the SAIBA framework, particularly the GRETA platform [19] and the SmartBody system [28].

In the next section we will present our proposed models that, unlike pre-mentioned rule-based models (BEAT, SAIBA, etc), rely on machine learning and statistical modeling to intrinsically associate actions and percepts and to organize sequences of percepts and actions into so-called joint sensory-motor behaviors.

SOCIAL BEHAVIOR MODELING

This section presents statistical/probabilistic approaches for modeling jointly multimodal sensory-motor behaviors. Thus, these models should enable an artificial agent (1) estimate its cognitive state from perceptual observations (e.g. speech activity/gaze fixations of the partner), this state should reflect the joint behaviors of the conversation partners at that moment; (2) generate suitable actions (e.g. its own gaze fixations) that should reflect its current cognitive state and its current awareness of the evolution of the shared plan.

Each situated conversation is controlled by a specific syntax that defines a particular sequencing of joint cognitive states by a sort of behavioral grammar. As matter of fact, we chose HMMs because they have intrinsic sequential and temporal modeling capabilities. We compare here their performance with those of two well-known powerful classifiers (SVMs and Decision Trees).

HMMs

For each dyad, we model each cognitive state with a single Discrete Hidden Markov Model (DHMM) and the whole interaction with a global HMM, that chains all single models with a task-specific grammar. The hidden states of these HMMs model the perception-action loop by capturing joined behaviors. In fact, the observations vectors are composed by two streams: the first stream contains the perceptual observations and the second stream observes actions. The "hidden" states are then sensory-motor. In the training stage, all data are available while in testing only perceptual observations are available. After training, two sub-models are thus extracted: a recognition model that will be responsible of estimating sensory-motor states from perceptual observations and a generation model that will generate actions from these estimated states. In our model, these two phases of decoding and generation are performed incrementally using a modified version of the Short-Time Viterbi algorithm [5]. Since observations here have discrete values, we called this model IDHMM (for Incremental

Discrete HMM). For more details about the IDHMM model see [21].

SVMs and Decision Trees

SVMs and Decision Trees are among the most used and powerful classifiers. In our context, we will train two distinct classifiers: the first one will estimate the most likely cognitive state from perceptual observations while the second one will directly determine the most likely actions from perceptual observations.

APPLICATION TO A FACE-TO-FACE INTERACTION

Experimental setting

The dataset used has been collected by Bailly et al. [1]. The setting is shown in Figure 1. It consists of speech and gaze data from dyads playing a speech game via a computer-mediated communication system that enabled eye contact and dual eye tracking. The gaze fixations of each one are estimated by positioning dispersion ellipsis on fixation points gathered for each experiment after compensating for head movements. The speech game involved an instructor who reads and utters a sentence that the other subject (respondent) should repeat immediately in a single attempt. Dyads exchange Semantically Unpredictable Sentences (SUS) that force the respondent to be highly attentive to the audiovisual signals. The experiment was designed to study adaptation: one female main speaker LN interacted with eight subjects (females) both as an instructor for ten sentences and as a respondent for another set of ten sentences.

Data and models

For each dyad, we have two observations streams: voice activity (*v1/v2* with 2 modalities: on/off) and gaze fixations (*g1/g2* with 5 regions of interest ROI: face/mouth/left eye/right eye/else) of both speakers. Seven cognitive states (CS) [4] have been labeled semi-automatically ('Read', 'Prephon', 'Speak', 'Wait, 'Listen', 'Think' and 'Else'). For SVMs and Decisions Trees, a first classifier is used to estimate the CS of the principal subject LN from (*v1, v2, g2*). Then a second classifier is used to estimate her gaze (*g1*) from the same data. Similarly for the IDHMM, the recognition model is used to estimate the CS from (*v1,v2,g2*) and the eye fixations (*g1*) are synthesized using the generation model.

Gaze data have been monitored by two Tobii® eyetrackers operating at 25Hz. Voice activity detection has been sampled at the same rate.

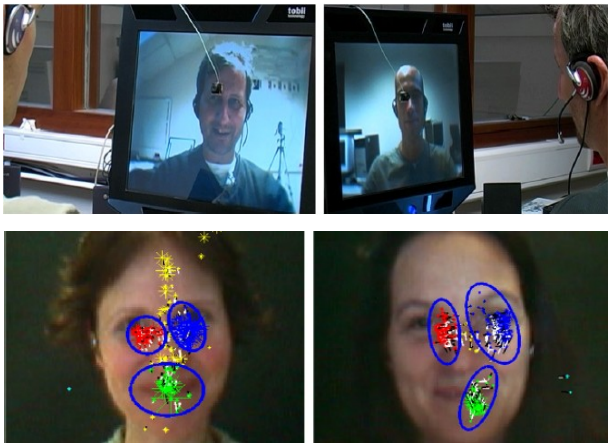


Figure 1: Experimental setting (only female subjects are included in our dataset)

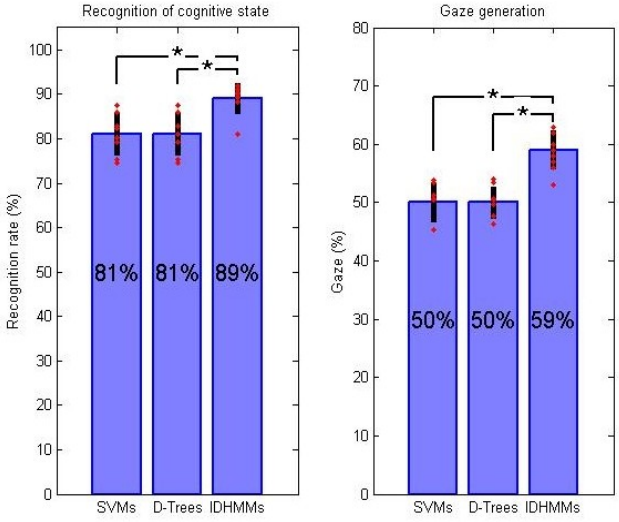


Figure 2: Results of the three models: SVMs, Decision Trees and IDHMMs

Results and comparison

DHMMs are trained with HTK [12], the IDHMM model was implemented in Matlab using PMTK3 toolkit [8]. For SVMs/Decision Trees, the Weka java package [11] has been used for both training and testing. For all models, 8-fold cross validation was applied: 7 subjects have been used for training while the eighth for testing.

Accuracy rates are used to evaluate cognitive state recognition, where the Levenshtein distance [20] is adopted for the evaluation of gaze generation because it allows more adequate comparison between generated and original signals. In fact, The Levenshtein distance is a metric for measuring the difference between two sequences; it computes the minimum number of elementary operations (insertions, deletions and substitutions) required to change one sequence into the other. From this optimal alignment, recall, precision and their harmonic mean (the F-measure)

can be directly computed. In this paper all generation rates represent F-measures.

Figure 2 clearly shows that there is no significant variation between the two classifiers. However, the IDHMM model outperforms the two classifiers and the improvement provided by this model is quite significant ($p<0.05$). The IDHMM model has a rate of 89% for cognitive state detection and 59% for eye gaze generation. Moreover Figure 5 shows that the IDHMM model is more efficient in detecting the structure of the interaction. We can see that the estimated path of cognitive states reflects correctly the predefined syntax of the task. In comparison, the SVMs have more difficulty in capturing the organization of the real path (see Figure 5) and discard short transition states: we can see that the estimated states are principally « Speak », « Wait » and « Listen ». This is in not in contradiction with the 81% recognition rate because these three cognitive states alone represent 85% of the ground truth. This performance gap is mainly due to the sequential constraints afforded by HMMs. This lack of sequential organization impairs the performance of SVMs and Decision Trees that should exclusively exploit bottom-up information provided by the observations.

MODELS WITH CONTEXTUAL ATTRIBUTES

New models

In order to build a generation model of demonstrative pronouns in dialogues of a collaborative situated task, Spanger et al. [27] proposed an SVM classifier that uses actual and historical information about the interaction. This idea is also used by [7] in order to generate beat gestures from the acoustic signal. In fact, classifier performance can be improved by adding memory (historical values) to each observation. In the previous section, at a time t, the initial models use only the data of that moment. In the new configuration, we added the same three attributes (v1,v2,g2) but from a previous instant t-T, T being the size of the memory. Moreover we have varied this sole instant T from 1 frame to 80 frames to find the optimal delay.

Results and comparison

Our tests revealed that there is no significance difference between SVMs and Decision Trees, thus, in the rest we will focus on comparative performance of SVMs vs. IDHMMs. Figure 3 shows that the optimal delay for this task is T= ~55 frames (~ 2 seconds). We got the same value for Decision Trees. This optimal delay corresponds exactly to [25] in which authors demonstrate that, if a speaker looks at an object, 2 seconds after the listener will most likely be looking at the same object. From Figure 4, we can see that the addition of past observations results in better performance ($p<0.05$) for both SVM recognition (91%) and generation (59%). This memory injection leads also to a better modeling of the interaction structure. In fact, in Figure 5 we can obviously notice the improvement of the recognition of cognitive states.

Likewise, we added this past observation to the sensory stream of the IDHMM. As a result, we also observe a significant improvement in the gaze generation (59% to 63%) while the recognition rate remains the same at a 95% confidence level.

In the initial configuration, we concluded that IDHMM model was the most efficient due to the sequential property of Markov Models. In the second configuration, the results are generally improved; while the IDHMM is still better in gaze generation (63% vs. 59%), the SVM model leads to a higher rate (91% vs. 87%) for a 95% confidence level. Hence, supplying the SVM model with memory has relatively addressed the missing temporal aspect.

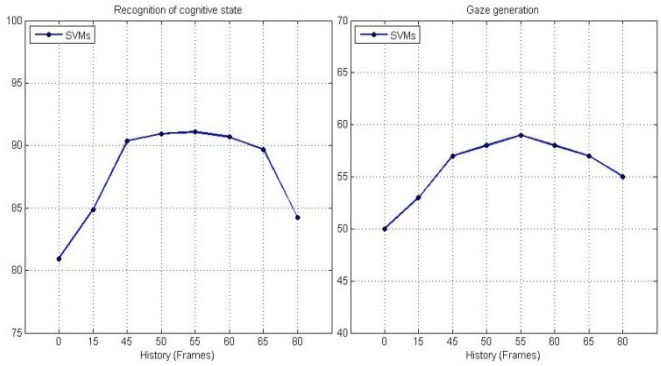


Figure 3: Optimal memory instant for the SVM

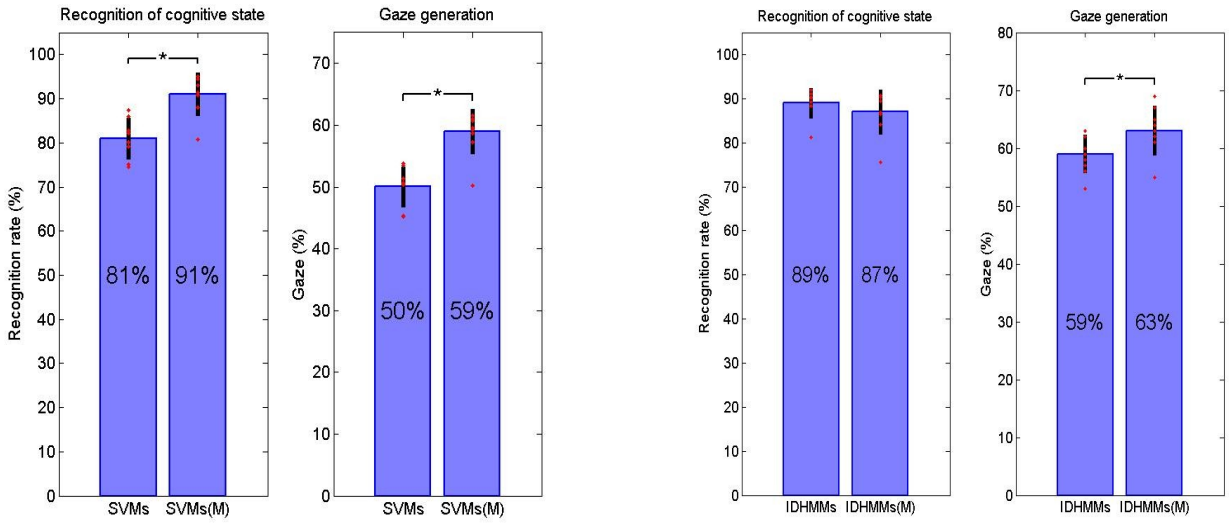
CONCLUSIONS

In this paper, we presented a comparative study of three behavioral models designed for social robots/agents (SVMs, Decision Trees and IDHMMs). These models have been tested in two different configurations: with & without history features. Comparison results showed that, in both settings, the IDHMM, thanks to its sequential modeling properties, remains a robust model for cognitive state recognition and eye gaze generation, and that classic classifier like SVMs could result in high performance if a certain memory (~2 seconds in our case) was included in the input observations.

Currently, we are studying a new scenario of a face-to face interaction that allows generating not only gaze but also deictic gestures. For the IDHMMs, we are also studying the influence of the number of hidden sensory-motor-states on the performance of each cognitive state and thus the impact on the generation figures.

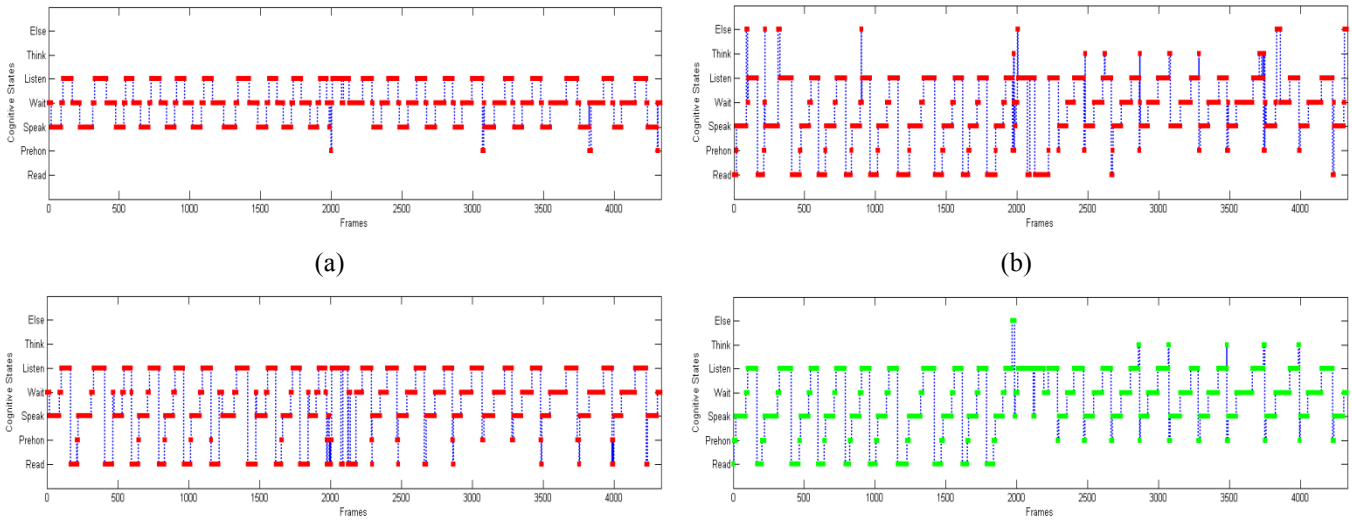
ACKNOWLEDGMENTS

This research is financed by the Rhône-Alpes ARC6 research council.



(a) (b)

Figure 4: No memory / Memory (M=55) (a) for SVMs (b) for IDHMMs



(a) (b)

(c) (d)

Figure 5: Estimation of the cognitive state (CS) for a specific subject (a) using SVMs (b) using IDHMM (c) using SVMs and memory attributes (d) the real CS path

REFERENCES

1. Bailly, G., Raidt, S., and Elisei, F. Gaze, conversational agents and face-to-face communication. *Speech Communication 52*, 6 (2010), 598–612.
2. Bailly, G. Boucles de perception-action et interaction face-à-face. *Revue fran\ccaise de linguistique appliquée 13*, 2 (2009), 121–131.
3. Banerjee, S. and Rudnicky, A.I. Using simple speech–based features to detect the state of a meeting and the roles of the meeting participants. (2004).
4. Baron-Cohen, S. *Mind Reading: The Interactive Guide to Emotions*. Jessica Kingsley Publishers, London u.a., 2004.
5. Bloit, J. and Rodet, X. Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, (2008), 2121–2124.
6. Cassell, J., Vilhjalmsson, H., and Bickmore, T. *BEAT: the Behavior Expression Animation Toolkit*. 2001.
7. Chiu, C.-C. and Marsella, S. Gesture Generation with Low-dimensional Embeddings. *Proceedings of the*

2014 International Conference on Autonomous Agents and Multi-agent Systems, International Foundation for Autonomous Agents and Multiagent Systems (2014), 781–788.

8. Dunham, M. and Murphy, K. *PMTK3: Probabilistic modeling toolkit for Matlab/Octave, http://code.google.com/p/pmtk3/*. .
9. Gatica-Perez, D. Analyzing group interactions in conversations: a review. *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, (2006), 41–46.
10. Gatica-Perez, D. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing 27*, 12 (2009), 1775–1787.
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1 (2009), 10–18.
12. HTK, The Hidden Markov Model Toolkit. *http://htk.eng.cam.ac.uk/*. .
13. Jayagopi, D.B., Hung, H., Yeo, C., and Gatica-Perez, D. Modeling dominance in group conversations using nonverbal activity cues. *Audio, Speech, and Language Processing, IEEE Transactions on 17*, 3 (2009), 501–513.
14. Kopp, S., Jung, B., Lessmann, N., and Wachsmuth, I. Max - A Multimodal Assistant in Virtual Reality Construction. *KI 17*, 4 (2003), 11.
15. Kopp, S., Krenn, B., Marsella, S., et al. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. *INTERNATIONAL CONFERENCE ON INTELLIGENT VIRTUAL AGENTS*, (2006), 21–23.
16. Krenn, B. and Pirker, H. Defining the gesticon: Language and gesture coordination for interacting embodied agents. *Proc. of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters*, (2004), 107–115.
17. Krenn, B. The NECA project: Net environments for embodied emotional conversational agents. *Proc. of Workshop on emotionally rich virtual worlds with emotion synthesis at the 8th International Conference on 3D Web Technology (Web3D), St. Malo, France*, (2003).
18. Lakin, J.L., Jefferis, V.E., Cheng, C.M., and Chartrand, T.L. The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior 27*, 3 (2003), 145–162.
19. Le, Q.A. and Pelachaud, C. Generating Co-speech Gestures for the Humanoid Robot NAO through BML. In E. Efthimiou, G. Kouroupetroglou and S.-E. Fotinea, eds., *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*. Springer Berlin Heidelberg, 2012, 228–237.
20. Levenshtein, V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady 10*, 8 (1966), 707–710.
21. Mihoub, A., Bailly, G., and Wolf, C. Social Behavior Modeling Based on Incremental Discrete Hidden Markov Models. In *Human Behavior Understanding*. Springer International Publishing, 2013, 172–183.
22. Otsuka, K., Sawada, H., and Yamato, J. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances. *Proceedings of the 9th international conference on Multimodal interfaces*, ACM (2007), 255–262.
23. Otsuka, K. Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings. (2011), 171–179.
24. Otsuka, K. Conversation Scene Analysis [Social Sciences]. *IEEE Signal Processing Magazine 28*, 4 (2011), 127–131.
25. Richardson, D.C., Dale, R., and Shockley, K. Synchrony and swing in conversation: coordination, temporal dynamics, and communication. In I. Wachsmuth, M. Lenzen and G. Knoblich, eds., *Embodied Communication in Humans and Machines*. Oxford University Press, 2008, 75–94.
26. Scherer, S., Marsella, S., Stratou, G., et al. Perception markup language: towards a standardized representation of perceived nonverbal behaviors. *Intelligent Virtual Agents*, (2012), 455–463.
27. Spanger, P., Yasuhara, M., Iida, R., and Tokunaga, T. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. *Proceedings of PreCogSci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*, (2009).
28. Thiebaux, M., Marsella, S., Marshall, A.N., and Kallmann, M. Smartbody: Behavior realization for embodied conversational agents. *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, (2008), 151–158.
29. Vinciarelli, A., Pantic, M., Heylen, D., et al. Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing 3*, 1 (2012), 69–87.
30. Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. Modeling individual and group actions in meetings with layered HMMs. *Multimedia, IEEE Transactions on 8*, 3 (2006), 509–520.