Pattern Recognition Letters 74 (2016) 82–89

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Graphical models for social behavior modeling in face-to face interaction☆

Alaeddine Mihoub[a,b,c,*], Gérard Bailly[a], Christian Wolf[b,c], Frédéric Elisei[a]

[a] GIPSA-Lab, Université de Grenoble-Alpes/CNRS, Speech & Cognition Department, France
[b] Université de Lyon/CNRS, France
[c] INSA-Lyon, LIRIS, UMR5205, F-69621, France

ARTICLE INFO

Article history:
Received 19 January 2015
Available online 11 February 2016

Keywords:
Face-to-face interaction
Behavioral model
DBN
Structure learning
Multimodal generation

ABSTRACT

The goal of this paper is to model the coverbal behavior of a subject involved in face-to-face social interactions. For this end, we present a multimodal behavioral model based on a dynamic Bayesian network (DBN). The model was inferred from multimodal data of interacting dyads in a specific scenario designed to foster mutual attention and multimodal deixis of objects and places in a collaborative task. The challenge for this behavioral model is to generate coverbal actions (gaze, hand gestures) for the subject given his verbal productions, the current phase of the interaction and the perceived actions of the partner. In our work, the structure of the DBN was learned from data, which revealed an interesting causality graph describing precisely how verbal and coverbal human behaviors are coordinated during the studied interactions. Using this structure, DBN exhibits better performances compared to classical baseline models such as hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs). We outperform the baseline in both measures of performance, i.e. interaction unit recognition and behavior generation. DBN also reproduces more faithfully the coordination patterns between modalities observed in ground truth compared to the baseline models.

1. Introduction

Face-to-face communication is considered as one of the most basic and classic forms of communication in our daily life [29]. Its apparent simplicity and intuitive use conceals a complex and sophisticated bidirectional multimodal phenomenon in which partners continually convey, perceive, interpret and react to the other person's verbal and co-verbal signals and displays [33]. Studies on human behavior have confirmed for instance that co-verbal cues – such as body posture, arm/hand gestures, head movement, facial expressions, and eye gaze – strongly participate in the encoding and decoding of linguistic, paralinguistic and non-linguistic information. Several researchers have notably claimed that these cues are largely involved in maintaining mutual attention and social glue [16].

Human interactions are paced by multi-level perception-action loops [1]. Thus, a multimodal behavioral model should be able to orchestrate this complex closed-loop system. In particular, the model is facing the complex task of predicting multimodal behavior given a scene analysis while monitoring the intended goals of the conversation. Our challenge in this paper is to build statistical multimodal behavioral models that are trained by exemplars of successful human–human (H/H) interactions i.e. that map perception to action. The end goal of this research is to build user-aware social robots that are able to engage efficient and believable face-to-face conversations with human partners. In this work, this problem is solved in a data driven way through a dynamic Bayesian network (DBN) whose graphical structure and its parameters are learned from observed training data. We will show that automatically learning the model's structure as well as parameters leads to faithful predictions of multimodal scores that reproduce how humans coordinate their own modalities (intra) and between each other (inter).

The paper is organized as follows: the next section briefly reviews the state-of-the art of multimodal behavior analysis and modeling. In Section 3, we present our face-to-face interaction scenario, the experimental setting and the collected signals. In Section 4, the DBN model is presented and the structure of intra-frame and inter-frame dependencies is discussed. Two other models (HMMs/HSMMs) are used as baselines. In Section 5, comparative performances of these models are given and discussed. We show that the DBN model outperforms the other

☆ This paper has been recommended for acceptance by A. Petrosino.

* Corresponding author at: GIPSA-Lab, Université de Grenoble-Alpes/CNRS, Speech & Cognition Department, France. Tel.: +33 646771357.
E-mail address: alaeddine.mihoub@gmail.com (A. Mihoub).

http://dx.doi.org/10.1016/j.patrec.2016.02.005
0167-8655/

statistical models both in terms of performance and reproduction of coordination patterns.

2. Related work

This research is a part of the general field of social signal processing (SSP) [36], a new emerging domain spanning research not only in signal and image processing but also in social and human science. In recent years, it is becoming an attractive research area and there is an increasing awareness about its technological and scientific challenges. SSP essentially deals with the analysis and synthesis of multimodal behavior in social interactions.

One of the goals of SSP is automatic conversation and scene analysis [10]. The challenge is here to retrieve high-level information such as cognitive activities (e.g. addressing, turn taking, backchannel), social emotions (e.g. happiness, anger, fear), social relations (e.g. roles) as well as social attitudes (e.g. degree of engagement or interest, dominance, personality) [36] from the exchanged signals. Several computational models have been proposed to cope with these problems. Pentland et al. [31] have characterized face-to-face conversations using wearable sensors. They have built a computational model based on coupled hidden Markov models (CHMMs) to describe interactions between two people and characterize their dynamics in order to estimate the success of the intended goals. Otsuka et al. [30] 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 estimates gaze patterns while the third one estimates conversation regimes. The objective of Otsuka and colleagues is to evaluate the interaction between regimes and behaviors during multi-party conversations. For social affect detection, Petridis and Pantic [32] presented an audiovisual approach to distinguish laughter from speech and showed that this approach outperforms the unimodal ones. The model uses a combination of AdaBoost and Neural Networks, where AdaBoost is used as a feature selector rather than a classifier. The model achieved a 86.9% recall rate with 76.7% precision. A decision tree is used in [2] 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 (SVM) 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 [10,36].

The second scope of SSP is the generation of relevant social behavior. The behavioral models should here predict the most appropriate sequence of multimodal signals for conveying given linguistic, paralinguistic or non-linguistic information. One possible application is to integrate these models into social agents [14] to make them capable of displaying social actions, social emotions and social attitudes via an appropriate animation of their artificial bodies. Several models have proposed to model and synthesize human behavior. We here focus on data-driven approaches, which automatically infer the behavioral models from data using machine learning techniques. For instance, Morency et al. [26] showed how sequential probabilistic models, i.e. HMMs (hidden Markov models) and CRFs (conditional random fields) can directly estimate listener backchannels from a dataset of human-to-human interactions using multimodal output features of the speaker, in particular spoken words, prosody and eye gaze. They notably addressed the problem of automatically selecting relevant features and their optimal representation for probabilistic models. Lee and Marsella [17] also opted for a probabilistic approach to predict speaker head nods and eyebrow movements for a virtual agent application. The authors explored different feature sets (syntactic features, dialog acts, paralinguistic features, etc.) and different learning algorithms, namely HMM, CRF and latent-dynamic CRF (LDCRF). Quantitative evaluation showed that the LDCRF models achieved the best performance, underlying the importance of learning the dynamics between different gesture classes and the orchestration of the gestures. In our previous work [24], we proposed statistical models that, for a given interaction scenario (i.e. a sentence-repeating game), estimate the cognitive state of a subject – given his verbal activity and the multimodal behavior of his interlocutor – and then generate his gaze. We showed that sequential models (HMMs) are better than frame-based classifiers (SVMs and decision trees) in both tasks. Moreover, Huang and Mutlu [12] used dynamic Bayesian networks (DBNs) to model the coordination of speech, gaze, and gesture behaviors in narration. Given input speech features, the most probable coverbal behavior – gesture type and gaze target – were computed. The evaluation of their model shows that this learning-based approach achieves similar performance compared to conventional rule-based approaches while reducing the effort involved in identifying hidden behavioral patterns. More generally, these learning approaches frequently use probabilistic graphical models because of their capacity to capture subtle covariations between modalities and coordination between multimodal events that often escape to human expertise. Other interesting properties of statistical models include their ability in discovering and exploiting hidden states and latent variables that are not directly observed. That is why, in this work, the proposed behavioral models are data-driven and confronted to multimodal observation spaces.

In the next section we describe the scenario we designed to collect multimodal data of H/H face-to-face social interactions. This data is then used to train and compare statistical models of joint behaviors.

3. Face-to-face interaction

3.1. Scenario

The objective of the proposed face-to-face interaction is to collect multimodal behaviors observed in a collaborative task called "put that there" [4] involving an instructor and a manipulator. This task – simple as it can appear at first sight – is a very interesting benchmark for studying and learning human strategies used to maintain mutual attention and coordinate multimodal deixis towards objects and locations.

More concretely, the task consists in reproducing a particular arrangement of cubes. Each game involves an instructor and a manipulator, the latter following orders of the former. The objective of the statistical model is to learn and reproduce the instructor's behaviors. The long-term goal is to transfer this model to a humanoid robot that will instruct a human manipulator. Credible scenarios where the HRI leads robots to instruct human partners are not so uncommon: robots may serve as coaches for physical or mental training [8,11] or rehabilitation, education [5,9] as well as instructors for gaming or shopping recommendation [34].

In our scenario, the instructor has to reproduce a target arrangement of cubes with the help of the manipulator who is the only agent allowed to touch and move the cubes. Conversely, the target arrangement is only known to the instructor. The instructor and the manipulator must therefore cooperate (i.e. share knowledge and coordinate their sensory–motor capabilities) to perform this collaborative task. The game involves 16 cubes. Each cube is marked by a colored symbol drawn on its upper side, i.e. a unique combination of one symbol (square, cross, circle and dot) and one