Learning joint multimodal behaviors for face-to-face interaction: performance & properties of statistical models

Gérard Bailly, Alaeddine Mihoub, Christian Wolf, Frédéric Elisei
Learning joint multimodal behaviors for face-to-face interaction: performance & properties of statistical models

Gérard Bailly(1)  Alaeeddine Mihoub(1,2)  Christian Wolf(2)  Frédéric Elisei(1)

(1) GIPSA-Lab, Université Grenoble Alpes/CNRS
11, rue des Mathématiques, St Martin d’Hères, France
cristian.wolf@liris.cnrs.fr

(2) Université de Lyon/CNRS
INSA-Lyon, LIRIS, UMR5205 F-69621, France
christian.wolf@liris.cnrs.fr

ABSTRACT
We evaluate here the ability of statistical models, namely Hidden Markov Models (HMMs) and Dynamic Bayesian Networks (DBNs), in capturing the interplay and coordination between multimodal behaviors of two individuals involved in a face-to-face interaction. We structure the intricate sensory-motor coupling of the joint multimodal scores by segmenting the whole interaction into so-called interaction units (IU). We show that the proposed statistical models are able to capture the natural dynamics of the interaction and that DBNs are particularly suitable for reproducing original distributions of so-called coordination histograms.

Theme: behavior coordination between animals, humans and robots
Keywords: human-human interaction; multimodal interaction; behavioral models; machine learning; statistical modeling.

1 INTRODUCTION
Deictic expressions [1] such as the famous “put that there” explored by Bolt [2] implies a tight coordination between gaze, head/torso/arm and finger pointing and speech [3, 4]. This context-dependent intermodal coordination is further affected by interleaving multimodal cues provided by the recipient of the information: responsive gaze cues, mimics as well as acoustic backchannels pace the effective encoding and decoding of the intended message. Several seminal works such as those of Richardson et al on swinging [5], MacFarland on respiration [6] and Bailly et al on gaze [7] have shown that the patterns of multimodal coordination are sensitive to cognitive requirements. Coordinated behaviors are traditionally described with an epistemic approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cope with such a complex orchestration of multimodal conversational agents, numerous rule-based systems have been approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others’ speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for streams: see notably the Ymir model proposed by Thórisson [8] or proposed to cop
Once trained on parallel perception and action streams (see Figure 1), sensory-motor HMMs are easily split into a recognition model that infers the optimal state sequence on the sole basis of the observed perception stream and a synthesis model that generates the most likely motor observations given the decoded sequence of states (see [21] for further explanation). Incremental Hidden Semi-Markov Models (HSSMs) further combine the possibility to model state duration (also called residence time) and to infer states and generate actions with limited look ahead. The bounded version of the Short-Time Viterbi (STV) algorithm [22] slightly impairs the estimation of the IUs but with no substantial degradation of the generated actions [23].

2.2 Dynamic Bayesian Networks (DBNs)

Classical HMMs are characterized by a fixed dependency graph modeling the conditional independence relationships between random variables. Extensions to the standard HMM formalism have been proposed to cope with direct input/output dependencies [24]. By representing observations and hidden states as random variables, DBNs are particularly suitable for modeling the dynamics of multimodal behaviors in face-to-face interactions [25]. As their graphical structure can be arbitrary. These dependency structures between variables can be provided by an expert but numerous methods have been introduced to learn the network’s structure automatically from data both for intra- and inter-frame conditional dependencies. An example of such a dependency graph obtained for the data described below is displayed Figure 3. We nicely recover the causal relations between the gaze, pointing gesture, speech of the instructor and the final grasping gesture of the manipulator.

3 MULTIMODAL INTERACTIVE DATA

We trained our statistical models on interactive data collected during a “put-that-there” game (see Figure 2). Models are trained to capture the behavior of the instructor. The perceptual and action streams are gathered by respectively analyzing the visual field of the instructor via a head-mounted camera and monitoring his speech, head, gaze and hand movements (see the deictic gesture in Figure 2) by motion capture devices. The scenario consists in a repetitive task that samples the working space: the instructor asks the manipulator to reproduce various cube arrangements on a chessboard according to a layout he is the only one to know. The task of the statistical model is to predict the gaze (FX) and hand movements (GT) of the instructor given his speech (SP) and the perceived gestures (MP) of the manipulator. All observations are discretized, e.g. alternative points of interest for gaze fixations (cubes, locations, hands, and manipulator’s face), gesture strokes (grasp, pick-up, transport, and release), speech (cube name, source and target locations), etc. We distinguish between 6 interaction units: get instruction, seek cube, point to source location, indicate target location, check manipulation, and validate manipulation. Note that HMM sensory-motor states are fully connected: seeks, mutual attentions, errors and repetitions often result in state looping and acyclic graphs that may solicit several times the same hidden state within one interaction unit.

4 COORDINATION HISTOGRAMS

Observations are here discrete: modal events are thus generated each time the statistical model observes or generates a transient between successive observations. A coordination histogram (CH) for a couple of modalities is computed as follows: for each event, we search for the nearest event of the other modality and record the time lag between these two events. The CH for ground truth fixations with reference to input streams are compared in Figure 4 with CH for data generated by various statistical models. While DBN significantly outperforms HSMM – which significantly outperforms HMM – in terms of prediction performance [26], coordination patterns are also significantly closer to original ground truth data.

Figure 2. The “put-that-there” game.

Figure 3: Dependency graph between interaction units, perception and action streams automatically inferred from “put-that-there” data for the DBN. Arrows of different colors (distinguishing intra- vs. inter-modality, intra- vs. inter-frame dependencies) cue significant dependencies between units and observations. Contrary to HMM, no latent states have been added here.

CONCLUSIONS

We present statistical multimodal behavioral models trained on multimodal data collected on dyads involved in a situated collaborative face-to-face interaction. We compared the performance of Hidden Markov Model (HMM), Hidden Semi-Markov Model (HSMM) and Dynamic Bayesian Network (DBN). We introduce the concept of coordination histogram that characterizes how different modalities synchronize between each other. DBN leads to the best performances in both interactive units recognition and behavior generation. It also displays a faithful coordination between generated trajectories compared to the ground truth. We suggest incorporating such behavioral characteristics for model assessment and coordination studies.

We plan to further implement the models on social robots in order to gather subjective evaluations and performative assessments.
Figure 4: Coordination histograms of the instructor’s gaze with the input streams (his own speech onsets and the onsets of the manipulator’s strokes) for the ground truth (a) and the predicted gaze by HMM (b), HSMM (c) and DBN (d).

ACKNOWLEDGMENTS
This research is financed by the Rhône-Alpes ARC6 research council and the ANR-14-CE27-0014 SOMBRERO.

5 REFERENCES