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

Gérard Bailly, Alaeddine Mihoub, Christian Wolf, Frédéric Elisei

To cite this version:

Gérard Bailly, Alaeddine Mihoub, Christian Wolf, Frédéric Elisei. Learning joint multimodal behaviors for face-to-face interaction: performance & properties of statistical models. Human-Robot Interaction. Workshop on Behavior Coordination between Animals, Humans, and Robots, Mar 2015, Portland, United States. hal-01110290

HAL Id: hal-01110290
https://hal.archives-ouvertes.fr/hal-01110290
Submitted on 27 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Learning joint multimodal behaviors for face-to-face interaction: performance & properties of statistical models

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

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 Thorisson [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 developed for handling multimodal communication and interaction [9-11].

More recently, *epigenetic* approaches have been proposed to learn patterns of coordination from sensory-motor experience. These data-driven approaches automatically infer the behavioral models from embodied and grounded interaction thanks to machine learning techniques and various learning strategies (observation, demonstration, etc.). We introduce below two of these techniques.

2. STATISTICAL MODELS

Statistical models have been used for years for scene analysis, i.e. inferring semantic information from signals: speech and speaker recognition, visual scene analysis as well as inferring human activity, emotions or social features of the conversational partners. More recently, statistical models have been used to cope with behavior generation, e.g. see [12] for speech synthesis and [13] for gesture synthesis. Multi-space probability distribution models [14] and various statistical models capturing the cross-correlations between time series have been further used to generate signals from observed ones. Otsuka et al. [15] proposed a Dynamic Bayesian Network (DBN) to estimate addressing and turn taking (“who responds to whom and when?”) from speech and head gestures and generate appropriate gaze patterns. Moréncy et al [16] 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 (spoken words, prosody and eye gaze). For more insights see also [17, 18]. More generally, these approaches use probabilistic graphical models because they provide a flexible and trainable representation of the dynamics of human behavior and capture spatiotemporal relationships between multimodal observations under uncertainty.

Figure 1. Mapping perception to action using HMM. Hidden sensory-motor states are trained using parallel perception and action streams. The structuring of these hidden states into *interaction units* facilitates the alignment of states and observations at training stage. At run time, states are inferred on the sole basis of the observed perception stream and generate the corresponding motor stream (see text for further explanation).

2.1 Hidden Semi-Markov Models (HSMMs)

Sensory-motor scores (i.e. perceptual input and motor output trajectories) are here supposed to be generated by unobserved (hidden) states of a Markov process. In a hierarchical model, these hidden sensory-motor states are organized into *interaction units* (IUs) [19] (such as thinking, informing, listening, taking turn, glazing over, etc). Each IU is thus responsible for modeling several parts of the sensory-motor scores that share common semantic or pragmatic information. The number, extend and ordering of these IUs depend on the task. The sequencing of the IUs, i.e. their syntax, provides a sort of behavioral grammar that chains elementary interaction units. A given IU can be seen as an instance of the *joint cognitive states* of the interacting dyads. A similar concept was used in [20] in which the authors propose a gaze model driven by the cognitive operations of a virtual agent. Note that disjoined or desynchronized cognitive states can be modeled via coupled or product HMMs.
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 (HSMMs) 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] [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.

Figure 2. The “put-that-there” game.

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 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.](image)

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


