# On-the-fly behavior coordination for interactive virtual agents – A model for learning, recognizing and reproducing hand-arm gestures online

# (Extended Abstract)

Ulf Großekathöfer, Nils-Christian Wöhler, Thomas Hermann, Stefan Kopp
Center of Excellence Cognitive Interaction Technology, Bielefeld University, Germany
{ugrossek, nwoehler, thermann, skopp}@techfak.uni-bielefeld.de

## ABSTRACT

In human conversation, verbal and nonverbal behaviors are coordinated by the interlocutors on the fly. To participate in this, artificial conversational agents must be able to create, adopt, and adjust behaviors flexibly and autonomously. We present a novel approach to learning behavioral patterns online, Ordered Means Models (OMMs), that meets the demands of dynamic behavior coordination in interaction. We describe how OMMs enable the virtual agent VINCE to engage in playing Rock-Papers-Scissors games, in which he learns, adapts to, and recognizes every human opponent's gestures on-the-fly such that he becomes unbeatable after only a few rounds. An evaluation study demonstrating this is presented.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Algorithms

## Keywords

Ordered Means Models, Vince, Anytime Classification

## 1. INTRODUCTION

When interacting with one another, humans rely on a variety of expressive behaviors including words, gestures, facial expressions, or gaze. Being able to interact with others thus requires to perceive, recognize, and interpret such behaviors, as well as to generate and employ them purposefully to fulfill communicative intentions. To act autonomously in communication settings hence implies a number of requirements for virtual conversational agents: (1) fast and on-the-fly recognition, i.e. the ability to create first hypotheses even from partial, ongoing observation; (2) online learning and adaptation, i.e. the ability to learn a behavior from few presentations, and to adjust learned models incrementally to further observations of this behavior; (3) reproduction and generation, i.e. the ability to perform a learned or adapted behavior, e.g., to align to an observed behavior [2, 3, 4, 5]. In this paper we present an approach to endow interactive virtual agents with these abilities for the case of hand-arm gestures. The basic idea is to treat gestures as multivariate time series, e.g. location coordinates of body parts that evolve over time, and to develop a machine learning approach that meets these critical requirements of *on-the-fly recognition capabilities*, *online adaptability*, and *reproduction capabilities*.

**Appears in:** *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

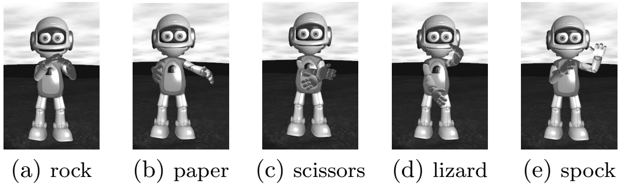
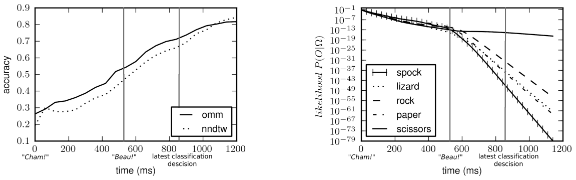

(a) rock (b) paper (c) scissors (d) lizard (e) spock

**Figure 1: Screenshots of VINCE performing the learned OMM prototypes for the five classes.**

## 2. ORDERED MEANS MODELS

We use a novel approach to on-the-fly classification of time series, which we refer to as Ordered Means Models (OMMs). OMMs are generative probabilistic state space models that emit a sequence of observation vectors out of $K$ fully connected, hidden states. Thereby, and as opposed to HMMs, OMMs do not include any transition probabilities between states, leading to a simple model architecture. The network of model states follows a left-to-right topology, i.e. OMMs only allow any transition to states with equal or higher indices as compared to the current state. The emissions of each state are modeled as probability distributions and are assumed to be Gaussian. The standard deviation parameter is identical for all states and is used as a *global hyperparameter*. In order to estimate particular model parameters from a set of observations we maximize the complete data log-likelihood by means of an EM-algorithm with respect to the mean vectors. The process of parameter estimation as well as the computation of production likelihood can be achieved efficiently by dynamic programming. This dynamic programming scheme also allows an on-the-fly and incremental evaluation of time series, i.e., sample-by-sample as they are observed. To use OMMs for classification, i.e. to assign a new gesture trajectory to one of $J$ classes, we assume that $J$ class-specific models have been estimated before from the available data. An unknown gesture then is assigned to the class associated with the model that yields the highest production likelihood of all models.

Given the above-mentioned model architecture, an OMM is completely defined by an ordered sequence of reference vec-



(a) Change of recognition accuracies of $\mathrm{OMM}_{\text{on-the-fly}}$ and $\mathrm{NN}_{\mathrm{DTW}}$ classifiers during observation of gestures.

(b) Likelihoods of class-related $\mathrm{OMM}_{\text{on-the-fly}}$ models during observation of a *spock* gesture.

tors, i.e. the expectation values of the emission distributions. Since these values are elements of the same data space as the observed data examples, the series of reference vectors is fully interpretable and reproducible as a time series prototype.

# 3. SCENARIO & EVALUATION

In order to evaluate the proposed approach, we realized an extended version of the rock-paper-scissors game for a human player and the virtual agent VINCE [1], adding two extra gestures to the game, a "lizard" and a "spock" gesture: *rock-paper-scissors-lizard-spock (RPSLS)*. In our setup, a Microsoft Kinect$^{\text{TM}}$ sensor captures the scene in 3D, in which we extract a human skeleton for a present user by means of the OpenNI[1] library. The agent does an initial counting phase to sync with the player, but instead of performing a pre-chosen gesture, VINCE tries to recognize as rapidly as possible the gesture of the human player and then to perform a corresponding winning gesture. For recognizing the gestures, VINCE uses an $\mathrm{OMM}_{\text{on-the-fly}}$-based classifier which returns a classification decision if the likelihood ratio between the most-likely and the second-most-likely OMM exceeded a value of 2 or, latest, $310ms$ after "Beau!". In result, the user gets the impression that VINCE presents his gesture without noticeable delay.

After each trial, the user has to name the gesture she just performed. Using this information, the agent learns and adapts to the particular user-specific way of performing the RPSLS-gestures. The game begins with an unlearned classifier and, thus, the human or the agent will win by chance. The classifier is then re-trained after each turn from all data collected so far. Hence, Vince's abilities to predict the gestures of the player improve rapidly during the course of the game. Further, VINCE uses the learned OMM prototypes to generate gestures during the game himself, thus reproducing the observed behaviors and coordinating with the user. Before a model is available, i.e. before the user presented a gesture to VINCE, we use pre-recorded gesture trajectories.

We conducted an evaluation study with 11 participants who played the game with VINCE until either player reaches a score of 20. We collected a data set containing 439 gestures in five classes and recorded the wrist positions of both arms as location coordinates relative to the user's body center for later analysis. Figure 2(a) gives the results of the comparison of the recognition accuracies achieved with $\mathrm{OMM}_{\text{on-the-fly}}$ and Nearest-Neighbor with dynamic time warping distance measure ($\mathrm{NN}_{\mathrm{DTW}}$) classifiers in online classification. As can be seen, recognition accuracy of both classifiers increases with each additional sample available. For complete gesture performances, $\mathrm{NN}_{\mathrm{DTW}}$ classifiers reach a slightly higher accuracy of $\approx 0.84$ in contrast to a recognition rate of $\approx 0.82$ for $\mathrm{OMM}_{\text{on-the-fly}}$ classifiers. However, for partial gesture performances, $\mathrm{OMM}_{\text{on-the-fly}}$ classifiers yield up to $\approx 10\%$ (on average $\approx 5\%$) higher recognition rates. This indicates that $\mathrm{OMM}_{\text{on-the-fly}}$ classifiers are well suited for *on-the-fly* recognition of behavior patterns. Figure 2(b) shows how the production likelihoods of the $\mathrm{OMM}_{\text{on-the-fly}}$ models for the five different gestures evolve during observation of an example *spock* gesture. After $\approx 600ms$ (approximately on "Beau!") the model related to class *spock* stably stays on a likelihood level of $\approx 10^{-13}$ while the likelihood associated with the other models decreases to a minimum of $\approx 10^{-79}$. In this case, a recognition of this particular gesture performance is possible $\approx 600ms$ after the gesture performance begins, i.e. in synchronization with the gesture presentation. The learning curves show that, at the first turn with unlearned nor adapted OMMs, VINCE is almost completely unable to recognize a gesture the human player performs (recognition accuracy of $\approx 0\%$). Over the first 10-15 rounds, the recognition rate increases up to an average of $\approx 85\%$ demonstrating the rapid learning ability of the used $\mathrm{OMM}_{\text{on-the-fly}}$ classifiers. In total, VINCE managed to win all 11 games with a lead of at least 3 points.

# 4. CONCLUSIONS

We have proposed Ordered Means Models as a specific kind of probabilistic state-based model that can provide rapid learning, efficient processing, on-the-fly classification, and prototype reproduction. The results from the Rock-Paper-Scissors game scenario demonstrate that OMMs can meet in fact the before-mentioned requirements for fast interpersonal behavior adaptation and coordination. Future work will test how OMMs perform when confronted with natural, more variable communicative gestures and will use OMMs for hierarchical clustering of behavioral patterns.

## Acknowledgments

This work was supported by the DFG in the Cluster of Excellence Cognitive Interaction Technology (CITEC).

# 5. REFERENCES

[1] U. Großekathöfer, A. Sadeghipour, T. Lingner, P. Meinicke, T. Hermann, and S. Kopp. Low Latency Recognition and Reproduction of Natural Gesture Trajectories. In *ICPRAM (Int.Conf. on Pattern Recognition Applications and Methods)*, 2012.

[2] T. Inamura, I. Toshima, and Y. Nakamura. Acquiring motion elements for bidirectional computation of motion recognition and generation. In B. Siciliano and P. Dario, editors, *Experimental Robotics VIII*, volume 5, pages 372–381. Springer-Verlag, 2003.

[3] D. Kulić, W. Takano, and Y. Nakamura. Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden markov chains. *The International Journal of Robotics Research*, 27(7):761, 2008.

[4] J. Kwon and F. Park. Natural movement generation using hidden markov models and principal components. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 38(5):1184–1194, 2008.

[5] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(9):884–900, 1999.

[1] http://www.openni.org