Introduction to the Special Issue on Machine Learning for Multiple Modalities in Interactive Systems and Robots

HERIBERTO CUAYÁHUITL, Heriot-Watt University, United Kingdom
LUTZ FROMMBERGER, University of Bremen, Germany
NINA DETHLEFS, Heriot-Watt University, United Kingdom
ANTOINE RAUX, Lenovo Research, United States of America
MATHEW MARGE, Carnegie Mellon University, United States of America
HENDRIK ZENDER, Nuance Communications, Germany

This special issue highlights research articles that apply machine learning to robots and other systems that interact with users through more than one modality, such as speech, gestures, and vision. For example, a robot may coordinate its speech with its actions, taking into account (audio-) visual feedback during their execution. Machine learning provides interactive systems with opportunities to improve performance not only of individual components but also of the system as a whole. However, machine learning methods that encompass multiple modalities of an interactive system are still relatively hard to find. The articles in this special issue represent examples that contribute to filling this gap.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning — Interactive Learning, Supervised Learning, Reinforcement Learning, Multiclass Learning, Unsupervised Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing — Conversational Interfaces; I.2.9 [**Artificial Intelligence**]: Robotics — Human-Robot Interaction

General Terms: Theory, Algorithms, Design, Experimentation, Performance

Additional Key Words and Phrases: interactive robots, interactive systems, human-machine interaction, machine learning from interaction, intrinsic evaluation, extrinsic evaluation.

ACM Reference Format:
Heriberto Cuayáhuitl, Lutz Frommberger, Nina Dethlefs, Antoine Raux, Matthew Marge, Hendrik Zender, 2014. Introduction to the Special Issue on Machine Learning for Multiple Modalities in Interactive Systems and Robots. *ACM Trans. Interact. Intell. Syst.* 4, 3, Article A (August 2014), 5 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. AIM AND SCOPE

Interactions between humans and artificial systems, may they be robots, virtual agents, machines, or simply computer interfaces, have become increasingly complex and ubiquitous in recent years. The users do not only operate them, instead they are in interactive contact and communication with them. As complexity of the systems increases, this communication may employ different modalities, such as speech, touch, gestures, and vision. Modern interactive systems often use a combination of these modalities to communicate meaningfully. For example, a robot may coordinate its speech with its actions, taking into account visual feedback during their execution. Alternatively, a multimodal system can adapt its input and output modalities to the user's goals, workload, and surroundings. Machine learning methods provide interactive systems with opportunities to improve performance not only of individual components but also of the system as a whole. At the same time, interactive

Author's addresses: Heriberto Cuayáhuitl, Heriot-Watt University, Edinburgh, EH14 4AS, UK; Lutz Frommberger, University of Bremen, P. O. Box 330 440, 28334 Bremen, Germany; Nina Dethlefs, Heriot-Watt University, Edinburgh, EH14 4AS, UK; Antoine Raux, 425 Market St, San Francisco, CA 94105-2403, USA; Matthew Marge, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA; Hendrik Zender, Kackertstr. 10, 52072 Aachen, Germany.

DOI: http://dx.doi.org/10.1145/0000000.0000000

systems provide machine learning with opportunities to develop new theories and algorithms driven by challenging tasks that involve interaction with humans and their environments. This special issue highlights some advances and challenges in the field of machine learning for multiple modalities in interactive systems.

2. MULTIMODAL INTERACTIVE LEARNING SYSTEMS: WHAT AND WHY?

An **interactive learning system** is an entity that learns through continuous interaction with the physical world, humans and/or other systems [Cuayáhuitl et al. 2013]. These kinds of systems can perceive the other end of the interaction, such as a human or another interactive system, (e.g. via speech and vision), and can act and communicate as part of the interaction (e.g. via speech and gestures). While many interactive systems and robots can be scripted or programmed to behave just as expected, the rich nature of interaction with the physical world and humans demands systems to deal with dynamic and partially known environments. Such systems therefore require some adaptation to dynamic environments (e.g., different degrees of noise or objects in constantly changing locations), and some learning due to partially known environments (e.g., unseen conversational behaviors or unseen objects/gestures).

A typical machine learning system learns a scoring function from inputs (e.g., features representing the state of the environment) to outputs (e.g., actions, class labels, or preferences). Important aspects of such a system include the task to be supported (e.g., object or gesture recognition—possibly combined with verbal interpretation), and the type and source of feedback given to the system (e.g., correct answers, or delayed rewards). Traditionally in the research community, machine learning has been applied on inputs and outputs from isolated datasets, which can have weak relationships to real world problems [Wagstaff 2012]. On the other hand, a system learning from interaction requires inputs and outputs from datasets that match the targeted real environment (often based on multiple modalities).

Machine learning frameworks differ in the way they treat data and the way they process feedback. Some machine learning frameworks addressed in this special issue are briefly described as follows:

— **Supervised learning** can be used whenever it comes to the task of classifying data, where the data consists of instances (pairs of features and class labels). The task of a supervised learning algorithm is to induce a function that maps the unlabelled instances to labels. This function is known as a classifier when the labels are discrete and as a regressor when the labels are continuous. All articles in this special issue make use of classifiers to predict events during human-machine interactions [Ngo et al. 2014; Benotti et al. 2014; Keizer et al. 2014; Cuayáhuitl et al. 2014].

— In contrast to supervised learning that makes use of direct feedback, **reinforcement learning** makes use of indirect feedback typically based on numerical rewards given during the interaction, and the goal is to maximize them in the long run. A policy is defined by the behaviour of the learning agent for a given domain, and can be found through a trial and error search in which the agent explores different action strategies. Each strategy is evaluated by the cumulative rewards that it yields over time. This framework can be seen as a very weak form of supervised learning, where the instances themselves are not rated, but the impact of actions that lead to the overall goal (e.g. fetching and delivering an object or playing a game). This form of learning is applied in this special issue to guide the behaviour of interactive robots as described in [Keizer et al. 2014] and [Cuayáhuitl et al. 2014].

— While both supervised and reinforcement learning assume some supervision at learning time, either in the form of labels or rewards, **unsupervised learning** does not require such information. Since an unsupervised learner does not receive any form of feedback, it has to find patterns in the data solely based on its observable features. The task of an unsupervised learning algorithm is thus to uncover hidden structure in unlabelled data. Examples of this form of learning include clustering and non-parametric Bayesian learning, which are applied to multimodal interaction in this special issue by [Keizer et al. 2014].

- Compared to the types of learning described above, **active learning** includes a human directly within the learning procedure. An active learner assumes three data sets: labelled examples, unlabelled examples, and chosen examples. The latter are built in an interactive fashion by an active learning algorithm who queries a human annotator for labels of specific examples, e.g. those it is most uncertain of. The learner can therefore actively ask the teacher for the knowledge that would help it learn the fastest rather than acting as a passive recipient of knowledge. This form of learning is applied in this special issue to a human-robot interaction scenario by [Ngo et al. 2014].

While a single form of learning can be integrated and evaluated in interactive systems and robots, combining multiple forms of machine learning can be used to address perception, action and communication as a whole. The next section describes concrete examples of these kinds of systems.

3. ARTICLES IN THIS SPECIAL ISSUE

3.1. Efficient Interactive Multiclass Learning from Binary Feedback

In this article, Ngo, Luciw, Nagi, Förster, Schmidhuber, and Vien [2014] present a new algorithm for interactive learning that only requires binary (correct/incorrect) feedback instead of full (class label) feedback. Their approach, *upper confidence weighted learning*, adds a multi-class classifier to an online learning framework. The model updates whenever it receives feedback. Unlike the traditional machine learning paradigm, training and testing phases are entwined. In this approach, the best class prediction is not always selected. Instead, the model can select suboptimal responses that can ultimately be more informative – the authors label this a type of “artificial curiosity” that can improve learning. *Upper confidence weighted learning* strikes a balance between collecting binary feedback and conducting interactive learning. Since the model only receives binary feedback, teachers simply label observations as either “right” or “wrong”. However, in the human-robot domain, this sort of feedback can be intuitive for a human teacher to provide. *Ngo et al.* evaluate the algorithm over three multi-class datasets, one with direct application to the human-robot domain and two others that assess generalizability. They report that their framework outperforms the existing state-of-the-art without increasing computational complexity. In fact, their method outperforms some multi-class models that use class label feedback. They suggest extending their model to incorporate a mix of binary and class label feedback in future work.

3.2. Interpreting Natural Language Instructions Using Language, Vision, and Behavior

In this article, Benotti, Lau, and Villalba [2014] present an interactive system that learns to interpret *natural language* in the context of action sequences uttered between humans in a virtual 3D world. The system learns in a minimally supervised fashion in that it draws *feedback directly from observable cues*, such as a human user’s reaction to an instruction or a visible change of the environment. The domain knowledge gained in this way is subsequently utilized as an automatically annotated dataset for training classifiers that predict reactions to unseen instruction sequences. In a fully automated way, this system attains a prediction accuracy of 77%, which increases to over 90% when human feedback is allowed.

As a representative of an interactive system that learns from the environment and the humans within it, this research promises to alleviate our dependence on large sets of human-designed rules or manually annotated corpora to build sophisticated systems. Instead, the learner observes human behavior, imitates and iteratively improves it over time to reach near-human performance. Future work in the field will need to take an ambitious step forward and abandon the controlled conditions of virtual or artificial environments. Instead, systems will need to be deployed into the wild, where learning is driven through observation, imitation and adaptation to new users, environments and circumstances. To learn natural language, this requires a holistic interactive system that learns from visual and haptic real-world signals from humans that communicate genuine actions and plans.

3.3. Machine Learning for Social Multi-Party Human-Robot Interaction

In this article, Keizer, Foster, Wang, and Lemon [2014] identify *socially appropriate robot behavior* as an important prerequisite to successful human-robot interaction (HRI). In HRI it is typically not sufficient for a robot to achieve only its *task-based goals*; it is equally important to take the human and the social implications of human behavior into account when deciding *how* and *when* to act in order to achieve such goals. To this end, any intelligent machine interacting with human users needs to be able to recognize the user's intentions through continuous sensor-based observations, which, in the real world, are inherently noisy and uncertain. Keizer *et al.* refer to this as *social state recognition*. And, furthermore, it needs to be able to adapt its own decision making and action execution based on these observations, which the authors refer to as *social skills execution*.

The article discusses several machine-learning based approaches to dealing with social signals for robust (inter-)action planning and execution in *multimodal, multi-party HRI*. First, the authors present a *data-driven, supervised-learning based* approach to social state recognition, followed by a *reinforcement-learning based* approach to social skills execution. The approaches have been implemented and combined into an integrated robotic system, and evaluated in a bartender scenario with human customers. After that, the authors describe an alternative approach that combines social state recognition and social skills execution in an *unsupervised-learning based* framework. The approach makes use of *hierarchical Dirichlet processes* and *infinite POMDPs*. The models have been trained from realistic human-human as well as human-robot interaction data.

3.4. Non-Strict Hierarchical Reinforcement Learning for Interactive Systems and Robots

In this article, Cuayáhuitl, Kruijff-Korbayová, and Dethlefs [2014] present an interactive robot that learns dialogue policies for playing a quiz game. The humanoid robot achieves a balance between learning a scalable and flexible policy–which represents a trade-off in reinforcement learning (RL) interactive systems. Systems are usually either flexible (using flat RL) or scalable (using hierarchical RL), but rarely both. For example, hierarchical RL has been shown to increase the scalability of learnt policies by using a divide-and-conquer approach. As a drawback, however, hierarchical algorithms impose a rigid structure on the interaction with human users, leaving the user little flexibility in taking an initiative in the interaction.

The authors in this article aim to achieve the best of both worlds and introduce hierarchical reinforcement learning with flexible state transitions. The latter means that the learning agent is allowed to transition to new states within a subtask as well as across subtasks. This breaks up the pre-defined interaction structure and allows the user to converse flexibly. In addition, the authors introduce function approximation to generalise to the unseen states that flexible transitions may cause. The flexibly trained policy is compared against a conventional hierarchical RL policy in an experiment with human users. Results show that the proposed technique leads to more flexible interactions and is preferred by human users.

4. FUTURE DIRECTIONS

Machine Learning for Multimodal Interaction is still a young research area and many directions still need to be explored. In particular,

— Most work in this field is grounded in practical systems, whose specific issues and characteristics vary widely. Just in this special issue, the interactive tasks range from learning to label hand gestures to understanding natural language instructions, to optimizing social behavior. As the community investigates more types of tasks related to multimodal interaction, we should be able to understand better what issues are specific to multimodal systems, yet shared among different such systems. Only then will we be able to develop truly general models and algorithms. The article of Ngo et al. in this issue represents a nice attempt at developing a general ML paradigm that is particularly suited to learning interactive behavior. Future work in that direction should address questions specific to multimodality, such as: What are good models/model structures for integrating different modalities? Are there general ML principles that exist or can be derived specifically to address the issue of multimodal interaction?

- Machine Learning requires data. However, interaction data is more expensive to collect in large amounts than traditional static data (e.g. text corpora). Resorting to virtual environments as Benotti et al. did in their contribution can make it simpler to collect data through crowdsourcing. However, it is not yet well known how this data compares to real world data, particularly when it comes to multimodal interaction, since virtual environments necessarily have to simulate some or all of the modalities. How can researchers obtain data of enough quality and quantity to develop new models and algorithms of practical relevance?
- Last but not least, interactive learning systems need to be trained from real world environments. The articles in this special issue still rely on lab-based experiments and recruited participants. Training and evaluation from real environments (e.g. homes, shopping centers, assisted living environments and public spaces) remains to be demonstrated in multiple domains. This is perhaps the path to more advanced interactive learning systems that may target the optimization of multiple modalities as a whole rather than independent systems. ACM TiiS welcomes this kind of research [Jameson and Riedl 2011].

ACKNOWLEDGMENTS

We thank the chief editors of ACM TiiS, Anthony Jameson, John Riedl and Krzysztof Gajos, for letting this special issue become a reality and for their high commitment with this journal. Special thanks go to all authors that submitted articles to this special issue, and to the anonymous reviewers for their support in maintaining the high standard of this issue.

REFERENCES

Luciana Benotti, Tessa Lau, and Martín Villalba. 2014. Interpreting Natural Language Instructions Using Language, Vision, and Behavior. *ACM Trans. Interact. Intell. Syst.* 4, 3 (2014).

Heriberto Cuayáhuitl, Ivana Kruijff-Korbayová, and Nina Dethlefs. 2014. Non-Strict Hierarchical Reinforcement Learning for Interactive Systems and Robots. *ACM Trans. Interact. Intell. Syst.* 4, 3 (2014).

Heriberto Cuayáhuitl, Martijn van Otterlo, Nina Dethlefs, and Lutz Frommberger. 2013. Machine Learning for Interactive Systems and Robots: A Brief Introduction. In *IJCAI Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Perception, Action and Communication (MLIS '13)*. ACM International Conference Proceedings Series.

Anthony Jameson and John Riedl. 2011. Introduction to the Transactions on Interactive Intelligent Systems. *ACM Trans. Interact. Intell. Syst.* 1, 1 (Oct. 2011).

Simon Keizer, Mary Ellen Foster, Zhuoran Wang, and Oliver Lemon. 2014. Machine Learning for Social Multi-Party Human-Robot Interaction. *ACM Trans. Interact. Intell. Syst.* 4, 3 (2014).

Hung Ngo, Matthew Luciw, Jawas Nagi, Alexander Förster, Jürgen Schmidhuber, and Ngo Anh Vien. 2014. Efficient Interactive Multiclass Learning from Binary Feedback. *ACM Trans. Interact. Intell. Syst.* 4, 3 (2014).

Kiri Wagstaff. 2012. Machine Learning that Matters. In *International Conference on Machine Learning (ICML)*.