Online Motion Concept Learning: A Novel Algorithm for Sample-Efficient Learning and Recognition of Human Actions

Extended Abstract

Miguel Vasco
INESC-ID, Instituto Superior Técnico,
Universidade de Lisboa
Lisbon, Portugal
miguel.vasco@gaips.inesc-id.pt

Francisco Melo
INESC-ID, Instituto Superior Técnico,
Universidade de Lisboa
Lisbon, Portugal
fmelo@inesc-id.pt

David Martins de Matos
INESC-ID, Instituto Superior Técnico,
Universidade de Lisboa
Lisbon, Portugal
david.matos@inesc-id.pt

Ana Paiva
INESC-ID, Instituto Superior Técnico,
Universidade de Lisboa
Lisbon, Portugal
ana.paiva@inesc-id.pt

Tetsunari Inamura
National Institute of Informatics
Tokyo, Japan
inamura@nii.ac.jp

KEYWORDS

Learning agent capabilities (agent models, communication, observation); Other

ACM Reference Format:
Miguel Vasco, Francisco Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. 2019. Online Motion Concept Learning: A Novel Algorithm for Sample-Efficient Learning and Recognition of Human Actions. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Humans have the remarkable ability to interact with their environment through rich and diverse actions. In order to actuate on shared environments with humans, artificial agents should have the ability to recognize the human's actions. However the creation of an artificial agent that is capable of recognizing all possible human actions from prior training is unrealistic, given the diversity of potential actions and ways to perform them. If we aim at deploying competent artificial agents in such scenarios, this hurdle should be overcome.

A common approach is to program agents which are able to learn representations of the human's actions through demonstration [1]. In such scenario, it is improbable to expect that the human will willingly provide the agent with a large amount of demonstrations of novel actions for training purposes. Indeed, contrary to several methodologies based on deep-learning [2, 7], which obtain impressive recognition results yet require large amounts of training data, the agent should be able to learn and recognize novel actions in a sample-efficient manner, i.e. requiring a minimum amount of human-provided demonstrations.

As such, the learning procedure should consider the multimodal data provided by the human in order to create rich representations of novel actions. Yet, the conventional methodology to learn human action representations considers exclusively the motion pattern of the demonstrator. Several motion-data-based representations have been proposed to model human actions, such as view-invariant representations [10], feature-based representations [11], interpretable representations [6] and topological-based representations [9]. The creation of action representations based solely on motion data neglects the context of the demonstration, which often is decisive in the distinction of actions with similar motion patterns. If we wish to attain efficient learning of action representations through

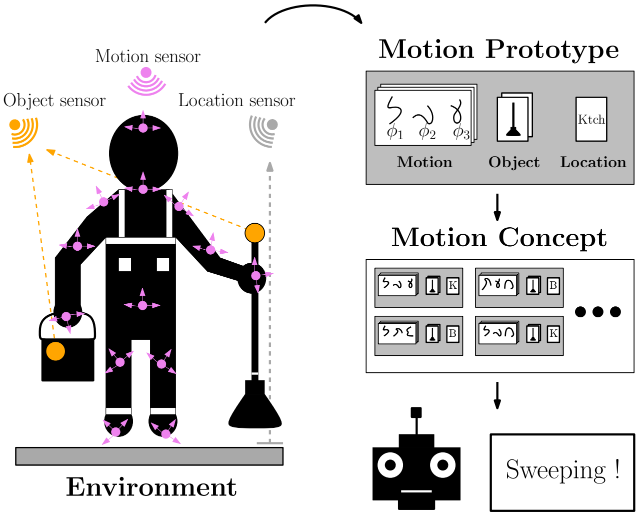

Figure 1: Scenario of the learning by demonstration approach for the creation of multimodal action representations, discussed in this work: The human performs a demonstration of a given action in an environment engineered with *motion*, *object* and *location* sensors. The single demonstration information is compactly represented in the *motion prototype*. The motion prototype is then employed for the creation, update or recognition of the correspondent *motion concept*, the high-level description of an action class introduced in this work.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

demonstration, we must consider the rich contextual background of the demonstration itself.

2 MOTION CONCEPT

We address the question of learning and recognizing human actions in a household environment, from few demonstrations provided by the human. The general setup of the learning from demonstration scenario described is presented in Figure 1. We consider that a human demonstrates an action class, which may involve interaction with objects present in the environment. Moreover, the demonstration takes place in any of a number of previously defined locations of interest that compose the environment. The environment itself contains sensors that provide information regarding the *location* where the human is performing the action, the *motion* of the human during the demonstration, as well as the *objects* that the human interacts with.

The demonstration of the action by the human user is compactly represented by the created *motion prototype*. Motion prototypes capture in a probabilistic way the motion and contextual information (object and location) provided by the human. For each motion sensor, the motion information is encoded as a sequence of motion primitives, widely used to describe animal motion and to represent robot motion [4, 5]. The motion primitives are selected from a library of available motion primitives, composed using the XOKDE++ algorithm [3], in order to maximize the likelihood of the observed trajectory. Similarly, the object and location information are described probabilistically as distributions of Bernoulli and categorical variables, respectively.

Moreover, as a single action may be performed in multiple ways, we introduce *motion concept* as a higher-level representation of an action. A motion concept contains a list of previously observed motion prototypes of that action class. Moreover, this representation also incorporates information obtained through interacting directly with the human demonstrator (the name of the action class) and weight parameters that allow the agent to reason about the importance of each contextual modality for the recognition of that specific action type.

3 LEARNING MOTION CONCEPTS

We evaluate the online learning of motion concepts in a virtual-reality household environment, with the presence of a human avatar. To do so we resort to the Online Motion Concept Learning (OMCL) algorithm, which is responsible for the creation of novel motion concepts through interaction with humans and the recognition of previously observed motion concepts [8]. Given a provided motion prototype of an unknown action class, the algorithm compares this representation with the motion concepts previously stored in its library. If the comparison cost between the provided motion prototype and a given motion concept is below a predefined threshold, then we use it to update the corresponding motion concept. Otherwise, we use the provided motion prototype to create a novel motion concept of the performed action class. For a complete description of the algorithm and the evaluation scenario please refer to [8].

We evaluate the performance of our learning system on two different aspects: the recognition of previously observed actions, whose motion concept was already built, and the identification of novel, previously unobserved, actions. The evaluation is performed on a *tabula rasa* scenario, i.e., a scenario in which the system, prior to the evaluation, is not trained on any data. We asked 12 participants to perform a sequence of demonstrations of 10 action classes. Each participant performs one example of 5 action classes and two examples of the remaining 5 action classes. In each example, the system processes the demonstration data and assesses its nature: a novel action class, previously unobserved, or a previously observed action class along with its denomination. The participant subsequently evaluates the assessment of the system and, accordingly to the participant's response, the system creates, or updates, the corresponding motion concept. The selection of the 10 classes and of the subset of classes with two demonstrations, along with the order of the actions to perform, are randomly selected.

We evaluated a total of 168 interactions corresponding to 14 action demonstrations of 12 participants, discarding the initial interaction of every participant which the system always recognizes as a novel motion concept. We define a *successful* interaction when the participant evaluates the system's assessment of the demonstration as correct, and *unsuccessful* otherwise. The results show that a majority of the interactions (80.4%) are successful. We can decompose the total 168 interactions into two different categories: 108 interactions due to the demonstration of a novel action class (*novel* interactions) and 60 interactions due to the performance of a previously observed action class (*recognition* interactions). The system is able to correctly evaluate the novelty of a action class previously not demonstrated in 85% of novel interactions. In 72% of recognition interactions, the system is also able to recognize previously observed action classes. In 18% of these interactions the system also recognizes the correct action class yet, due to significant differences in the motion pattern, location or objects interacted during the performance, it classifies the demonstration as an example of a novel action class. Nonetheless, the results attest to the system's potential for sample-efficient action learning and recognition in the presence of a human demonstrator.

4 CONCLUSIONS

In this work we introduce motion concepts, a novel multimodal representation for human actions in a household environment. In an online motion concept learning task, we demonstrate the potential of the discussed framework for action learning scenarios. We showed that the system is able to access the novelty of a provided unknown sample and to build the associated motion concept through interaction with the human. We plan to develop further work on the extension of the representation framework to task learning as well to actions performed by multiple agents.

ACKNOWLEDGMENTS

This work was supported by the NII internship program and JST CREST(No. JPMJCR15E3). This work was also supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019. The first author also acknowledges the FCT PhD scholarship SFRH/BD/139362/2018. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] Sonia Chernova and Andrea L Thomaz. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8, 3 (2014), 1–121.

[2] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.

[3] Jaime Ferreira, David Martins de Matos, and Ricardo Ribeiro. 2016. Fast and Extensible Online Multivariate Kernel Density Estimation. *arXiv preprint arXiv:1606.02608* (2016).

[4] Tamar Flash and Binyamin Hochner. 2005. Motor primitives in vertebrates and invertebrates. *Current opinion in neurobiology* 15, 6 (2005), 660–666.

[5] Jens Kober and Jan Peters. 2009. Learning motor primitives for robotics. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2112–2118.

[6] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2014. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation* 25, 1 (2014), 24–38.

[7] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*. 568–576.

[8] Miguel Vasco, Francisco S. Melo, David Martins de Matos, Ana Paiva, and Tetsunari Inamura. 2019. Learning multimodal representations for sample-efficient recognition of human actions. (2019). arXiv:arXiv:1903.02511

[9] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 588–595.

[10] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. 2012. View invariant human action recognition using histograms of 3d joints. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 20–27.

[11] Xiaodong Yang and Ying Li Tian. 2012. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 14–19.