In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics (ML-HLRC), Daejeon, Korea (2016).

Multi-modal Integration of Speech and Gestures for Interactive Robot Scenarios

Francisco Cruz, German I. Parisi, and Stefan Wermter

I. INTRODUCTION

Human-Robot Interaction (HRI) has become an increasingly interesting area of study among developmental roboticists since robot learning can be speeded up with the use of parent-like trainers who deliver useful advice allowing robots to learn a specific task in less time than a robot exploring the environment autonomously [1]. In this regard, the parent-like trainer guides the apprentice robot with actions that allow to enhance its performance in the same manner as external caregivers may support infants in the accomplishment of a given task, with the provided support frequently decreasing over time. This teaching technique has become known as parental scaffolding [2].

When interacting with their caregivers, infants are subject to different environmental stimuli which can be present in various modalities. In general terms, it is possible to think about some of those stimuli as guidance that the parent-like trainer delivers to the apprentice agent. Nevertheless, when more modalities are considered, issues can emerge regarding the interpretation and integration of multi-modal information, especially when multiple sources are conflicting or ambiguous [3]. As a consequence, the advice may not be clear and misunderstood, and hence, may lead the apprentice agent to a decreased performance when solving a task [1].

II. MULTI-MODAL INTEGRATION

People are constantly subject to different perceptual stimuli through different modalities such as vision, hearing, and touch among others. Such modalities are used to perceive information and process it independently, in parallel, or integrating the received information to provide a coherent and robust perceptual experience. Similarly, humanoid robots work with many of these sensory modalities and the way of processing and integrating the information coming from various sources is currently an important research issue in autonomous robotics. In HRI scenarios, robots can take advantage of such multi-sensory information in order to improve their capabilities when any sensory modality is limited, lacking, or unavailable.

For instance, early work by Andre et al. [4] proposed a multi-modal integration of speech and gestures for human-computer interaction using a tactile glove to identify hand gestures and a microphone array for speech recognition. The system functionality was limited to manipulate geometric objects on topographical maps. In robotic scenarios, Kimura & Hasegawa [5] used an incremental neural network to integrate real-time information in order to estimate attributes for unknown objects. The method used an RGB-D camera, a stereo microphone, and pressure and weight sensors to process different modalities. Ozasa et al. [6] proposed the integration of image and speech recognition confidence values to improve the recognition accuracy of unknown objects using logistic regression. In their approach, the confidence integration does not consider the case in which predicted labels are in contradiction. Moreover, in order to obtain improved recognition, it is also necessary to estimate proper logistic regression coefficients.

Nevertheless, in domestic scenarios and dynamic environments, assistive robot companions still need to understand and interpret instructions faster and more efficiently, yielding the integration of available multi-sensory information with different confidence levels in a consistent mode.

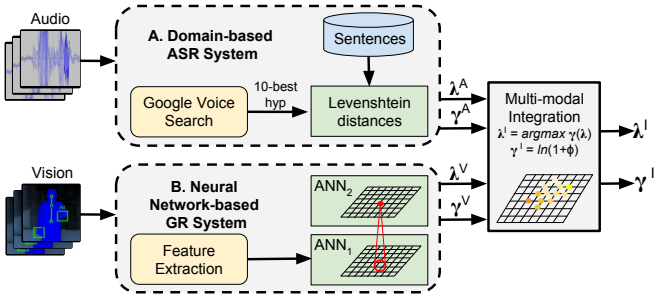


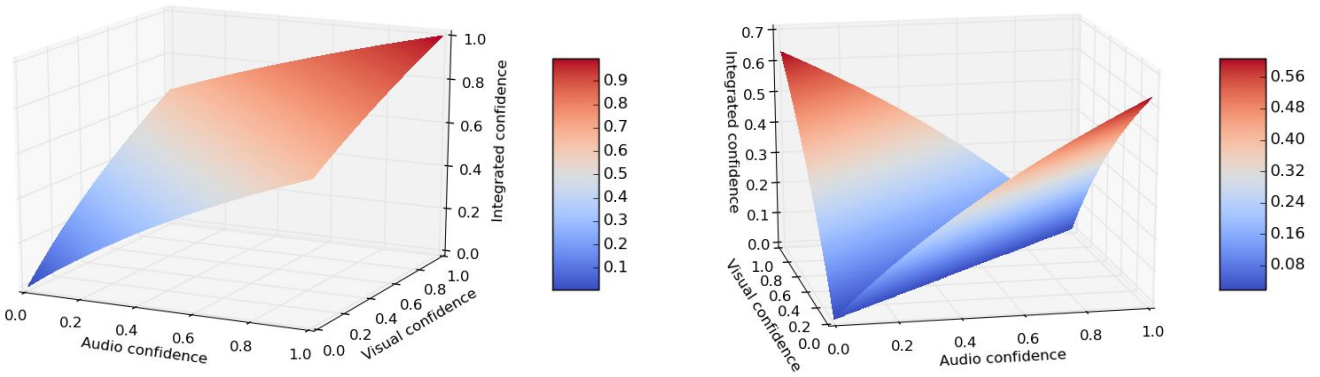
Fig. 1. Overall view of the system architecture. A domain-based automatic speech recognition system (on top) processes the audio input modality to obtain an audio advice label λ^A and an audio confidence value γ^A and a neural network-based gesture recognition system (at bottom) processes the visual input modality to obtain a visual advice label λ^V and a visual confidence value γ^V. Afterwards they become the input of the multi-modal integrative system to obtain the integrated advice label λ^I and the integrated confidence value γ^I.

III. OUR APPROACH

In our architecture, a parent-like trainer interacts with an apprentice robot using speech and gestures as guidance. In this work, we are particularly focused on the integration of multi-modal audiovisual inputs. A general overview of the architecture including the speech and gesture processing is depicted in Fig. 1, where λ and γ are the label and the confidence value respectively. First, the audio and visual

The authors gratefully acknowledge partial support by the Universidad Central de Chile, CONICYT scholarship 5043, the German Research Foundation DFG under project CML (TRR 169), and the Hamburg Landesforschungsförderungsprojekt.

Francisco Cruz, German I. Parisi, and Stefan Wermter are with the Knowledge Technology Institute, Department of Informatics, University of Hamburg, Germany. See: http://www.informatik.uni-hamburg.de/wtm/.



(a) Integrated confidence with equal uni-modal predicted labels

(b) Integrated confidence with different uni-modal predicted labels

Fig. 2. Obtained confidence values, in (a) the corresponding output labels for audio and visual modalities are the same, in (b) they are different.

sensory inputs are individually processed using the *DOCKS* speech recognition system [7] and a variation of *HandSOM* for gesture recognition [8]. Then, the outputs, i.e. predicted labels and confidence values, become inputs for the multi-modal integration system.

We propose a mathematical function which relates the predicted advice classes and confidence pairs from uni-sensory input denoted as (λ^A, γ^A) for audio and (λ^V, γ^V) for vision. The integrated predicted label λ^I is calculated according to the highest confidence value:

$$\lambda^I = \underset{\lambda}{\operatorname{argmax}}\ \gamma(\lambda) \tag{1}$$

In other words, if the audio and visual labels λ^A and λ^V are different, then the integrated label λ^I takes the value from the modality which has the biggest confidence value.

On the other hand, the integrated confidence value is computed by the function:

$$\gamma^I = \ln(1 + \phi), \tag{2}$$

where ϕ is a time-varying parameter which depends on each label λ and confidence value γ. We call this parameter the *likeness parameter* and it is obtained according to the following equation:

$$\phi = \begin{cases} \gamma^A + \gamma^V & \text{if } \lambda^A = \lambda^V \\ |\gamma^A - \gamma^V| & \text{if } \lambda^A \neq \lambda^V \end{cases} \tag{3}$$

Therefore, if the labels λ^A and λ^V are the same, then the confidence value γ^I is calculated using $\phi = \gamma^A + \gamma^V$ in order to strengthen the integrated confidence level over the classification made from both devices. On the contrary, if the labels λ^A and λ^V are different, then the integrated confidence value γ^I is calculated using $\phi = |\gamma^A - \gamma^V|$ in order to diminish the confidence level given the differences in the classification.

The proposed integration function yields an integrated confidence value $\gamma^I \in [\ln(1), \ln(3)] = [0, 1.0986]$. We use a unity-base normalization to rescale the range of confidence between 0 and 1. Fig. 2 shows the integrated confidence values when the predicted audio and visual labels are the same (a) and different (b).

IV. DISCUSSION AND FUTURE WORK

We have proposed a multi-modal integration of dynamic audiovisual input advice. The shown architecture processes individually the input advice to classify them with a correspondent associated confidence value. Afterwards, we integrate the input advice into one single label and confidence value. In this regard, we have shown an integration function that allows to strengthen or diminish the integrated advice for a learning robot using multiple sources of information for a more natural trainer-like learning procedure. The higher (or lower) confidence value of the integrated signal can lead the robot to act differently according to the specific task that it is intended to solve.

Future work directions consider experiments in HRI scenarios accounting for online interactions in order to effectively test the proposed method.

REFERENCES

[1] F. Cruz, S. Magg, C. Weber, and S. Wermter. Training agents with interactive reinforcement learning and contextual affordances. Accepted to IEEE Transactions on Autonomous Mental Development, 2016, doi: 10.1109/TCDS.2016.2543839.

[2] E. Ugur, Y. Nagai, H. Celikkanat, and E. Oztop. Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills. In Robotica, vol. 33, pp. 1163–1180, 2015.

[3] J. Bauer, J. Dávila-Chacón, and S. Wermter. Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes. In Connection Science, vol. 27, no. 4, pp. 358–376, 2015.

[4] M. Andre, V. G. Popescu, A. Shaikh, A. Medl, I. Marsic, C. Kulikowski, and J. Flanagan. Integration of speech and gesture for multimodal Human-Computer interaction. In International Conference on Cooperative Multimodal Communication, pp. 28–30, 1998.

[5] D. Kimura and O. Hasegawa. Estimating multimodal attributes for unknown objects. In Proceedings of the International Joint Conference on Neural Networks IJCNN, pp. 1–8, 2015.

[6] Y. Ozasa, Y. Ariki, M. Nakano, and N. Iwahashi. Disambiguation in unknown object detection by integrating image and speech recognition confidences. In Computer Vision – ACCV 2012, pp. 85–96, 2012.

[7] J. Twiefel, T. Baumann, S. Heinrich, and S. Wermter. Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In Proceedings of the 28th AAAI Conference on Artificial Intelligence AAAI, pp. 1529–1535, 2014.

[8] G.I. Parisi, D. Jirak, and S. Wermter. HandSOM - Neural clustering of hand motion for gesture recognition in real time. In Proceedings of IEEE International Symposium on Robot and Human Interactive Communication RO-MAN, pp. 981–986, 2014.