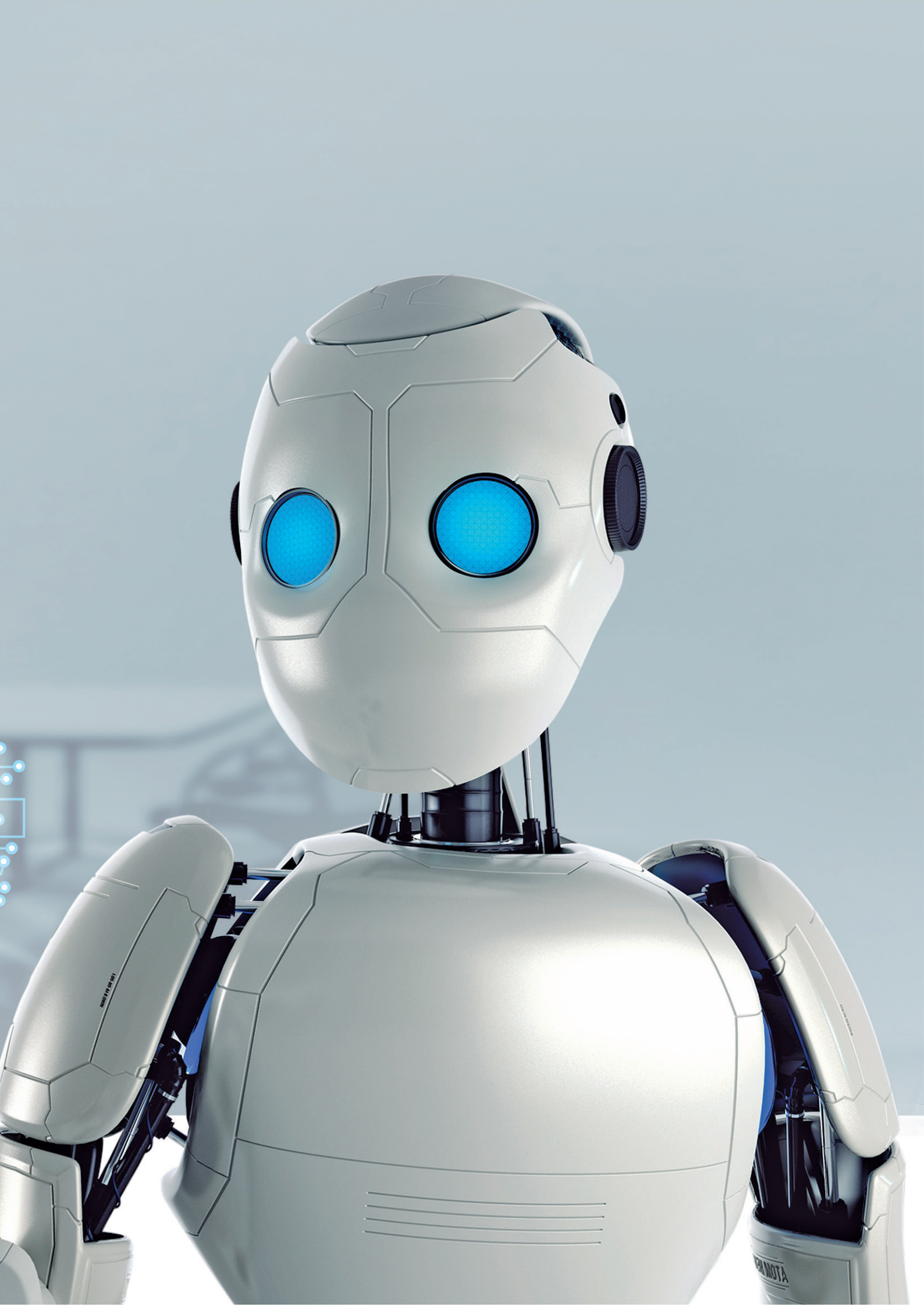
¿CÓMO EDUCAR ROBOTS?

INTEGRACIÓN MULTI-MODAL DE VOZ Y GESTOS PARA ESCENARIOS INTERACTIVOS ROBÓTICOS

Francisco Cruz

Doctor en Ciencias de la Computación



INTRODUCCIÓN

La interacción humano-robot (HRI, Human-Robot Interaction) se ha convertido en un área creciente de interés y estudio entre científicos que trabajan en robótica evolutiva y cognitiva. Esto debido a que el aprendizaje en robots puede ser acelerado con el uso de entrenadores similares a los padres quienes entregan consejos útiles permitiendo a los robots aprender tareas específicas en menos tiempo en comparación a un robot explorando el entorno de forma autónoma. En este sentido, el entrenador similar a un padre guía al robot aprendiz con acciones que permiten mejorar su rendimiento de la misma manera que un asistente puede apoyar a un infante en el cumplimiento de una tarea dada, teniendo en cuenta que la ayuda que se provee frecuentemente decrece en el tiempo. Este técnica de enseñanza es conocida como andamiaje parental (*parental scaffolding*).

Al interactuar con sus asistentes, los infantes son sometidos a diferentes estimulos ambientales, los cuales pueden estar presentes en diferentes modalidades. En términos generales, es posible considerar algunos de esos estímulos como guías que el entrenador entrega al agente aprendiz. Sin embargo, cuando más modalidades sensoriales son consideradas, pueden emerger problemas relacionados a la interpretación y la integración de la información multi-modal, especialmente cuando múltiples fuentes son conflictivas o ambiguas. Como consecuencia, la guía o consejo puede no ser claro o malentendido, y por lo tanto, puede conllevar al agente aprendiz a una disminución en su desempeño al resolver una tarea.

INTEGRACIÓN MULTI-MODAL

Las personas estamos constantemente sometidas a diferentes estímulos de percepción a través de diferentes modalidades tales como visión, audición y tacto entre otras. Tales modalidades son usadas para percibir información y procesarla independientemente, en paralelo o integrando la información recibida para proveer una percepción coherente y robusta. Similarmente, los robots humanoides trabajan con muchas de esas modalidades sensoriales y la forma de procesar e integrar la información, que viene de variadas fuentes, es actualmente un importante problema a investigar en los robots autónomos. En escenarios del tipo HRI, los robots pueden tomar ventaja de esta información multi-sensorial con el fin de mejorar sus capacidades cuando alguna modalidad sensorial esta no disponible, limitada, o es inexistente.

Por ejemplo, en un trabajo desarrollado por Andre et al. se propuso una integración multi-modal de audio y gestos para interacción humano-computador usando un guante táctil para identificar gestos en las manos y un conjunto de micrófonos para el reconocimiento de audio. La funcionalidad del sistema estuvo limitada solo a la manipulación de objetos geométricos en mapas topográficos. En escenarios robóticos, Kimura & Hasegawa usaron una red neuronal incremental para integrar información en tiempo real con el fin de estimar atributos de objetos desconocidos. El método uso una cámara de tipo RGB-D, un micrófono estereo y la presión o sensores de peso para procesar diferentes modalidades. Ozasa et al. Propusieron valores de confianza para la integración de imágenes y el reconocimiento de audio para mejorar precisión en el reconocimiento de objetos desconocidos por medio de una regresión logística. En este enfoque, los valores de confianza integrados no consideran el caso en el que las etiquetas predichas son contradictorias. Además, con el fin de obtener un mejor reconocimiento es necesario estimar adecuadamente los coeficientes de la regresión logística.

Sin embargo, en escenarios domésticos y ambientes dinámicos, los robots para asistencia aún necesitan entender e interpretar las instrucciones recibidas de manera más rápida y más eficientemente, además de integrar la información multi-sensorial disponible con diferentes niveles de confianza de una manera consistente.

ENFOQUE PROPUESTO

En nuestra arquitectura, un entrenados similar a un padre interactúa con un robot aprendiz usando su voz y gestos como guía o instrucciones. En este trabajo estamos particularmente enfocados en la integración de entradas audiovisuales multi-modales. Una vista general de nuestra arquitectura, incluyendo el procesamiento de la voz y los gestos, se muestra en la *Figura 1*, donde λ y γ son la etiqueta y el valor de confianza respectivamente. En primer lugar, las entradas sensoriales de audio y visión son procesadas individualmente usando el sistema de reconocimiento de audio *DOCKS* y una variación de *HandSOM* para el reconocimiento de gestos. Luego, las salidas (etiquetas predichas y valores de confianza) se convierten en las entradas para el sistema de integración multi-modal.

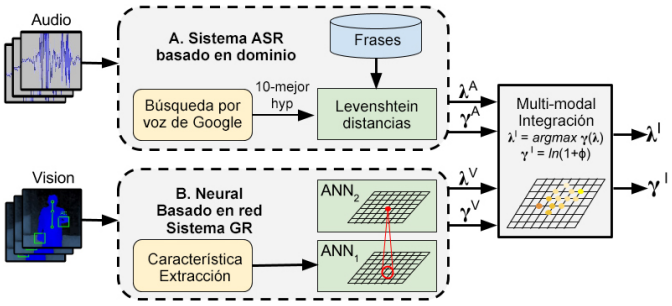


FIGURA 1

Vista

general de la arquitectura del sistema. Un sistema de reconocimiento automático de voz basado en un dominio específico procesa la entrada de modalidad de audio para obtener una etiqueta para el consejo de voz λ^A y un valor de confianza para el audio γ^A (parte superior del diagrama). Un sistema de reconocimiento de gestos basado en una red neuronal procesa la entrada de modalidad visual para obtener una etiqueta asociada a la instrucción gestual λ^V y un valor de confianza para la visión γ^V (parte inferior del diagrama). Luego, las salidas anteriores se convierten en las entradas del sistema multi-modal integrador para obtener una etiqueta de consejo integrada λ^I y un valor de confianza integrado γ^I.

En este trabajo proponemos una función matemática que relacione los pares de clases de consejo predichas y los valores de confianza obtenidos desde la entradas uni-sensoriales denotados como (λ^A, γ^A) para audio y (λ^V, γ^V) para visión. La etiqueta integrada predicha es calculada de acuerdo al mayor valor de confianza:

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

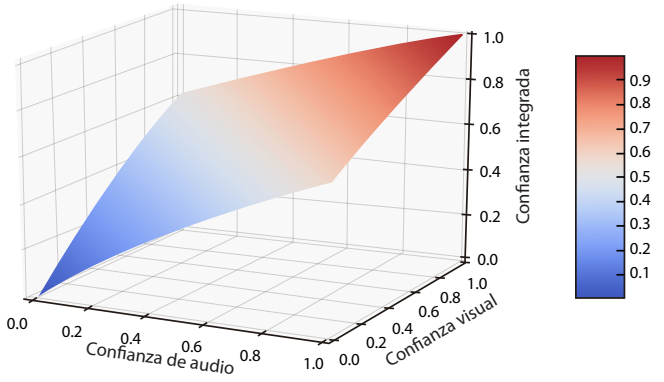
En otras palabras, si la etiqueta visual y de audio son diferentes, entonces la etiqueta integrada toma el valor desde la modalidad que posee el valor de confianza más alto. Por otra parte, el valor de confianza es calculado por la función:

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

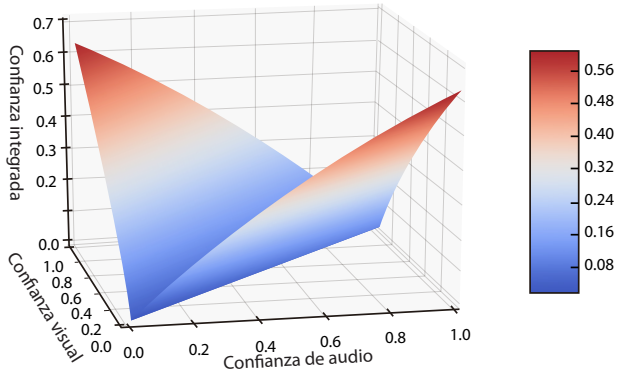
donde es un parámetro variante en el tiempo el cual depende de cada etiqueta y valor de confianza . Llamamos a parámetro de similitud y es obtenido de acuerdo a la siguiente ecuación:

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

Por lo tanto, si las etiquetas λ^A y γ^V son las mismas, entonces el valor de confianza γ^I es calculado usando $\phi = \gamma^A + \gamma^V$ con el fin de fortalecer el nivel de confianza integrado sobre la clasificación hecha por ambas modalidades. Por el contrario, si las etiquetas y λ^A y λ^V son diferentes, entonces el valor de confianza integrado γ^I es calculado usando $\phi = |\gamma^A - \gamma^V|$ con el objeto de reducir el nivel de confianza dadas las diferencias en las clasificaciones uni-modales.



(a) Valores de confianza integrados con etiquetas uni-modales predichas iguales.



(b) Valores de confianza integrados con etiquetas uni-modales predichas diferentes.

FIGURA 2

Valores de confianza obtenidos luego de la integración. En (a) las etiquetas de salida correspodientes a las modalidades de audio y visión son las mismas, mientras que en (b) son diferentes.

La función de integración propuesta produce valores de confianza integrado $\gamma' \in [\ln(1), \ln(3)] = [0,1.0986]$, por lo que utilizamos una normalización unitaria de base para reescalar el rango de valores de confianza entre 0 y 1.

La *Figura 2* muestra los valores de confianza integrados cuando la predicción de audio y de visión son iguales (a) y cuando son distintas (b).

DISCUSIÓN Y TRABAJOS FUTUROS

En el presente trabajo, se ha propuesto una integración multi-modal de consejos audiovisuales dinámicos. La arquitectura mostrada procesa de manera individual el consejo de entrada para clasificarlo con su correspondiente valor de confianza asociado. Posteriormente, el consejo es integrado en una sola etiqueta y un solo valor de confianza de manera consistente. En este sentido, se ha propuesto una función que permite fortalecer o disminuir el consejo integrado para un robot aprendiz usando múltiples fuentes de información para un procedimiento de aprendizaje más natural. Diferencias en los valores de confianza pueden conllevar a que el robot actúe de manera diferente de acuerdo a la tarea específica que se intenta resolver.

Trabajos futuros consideran experimentos en escenarios del tipo HRI tomando en cuenta interacciones en línea con el fin de testear el método propuesto de manera más efectiva. Además, futuros experimentos deberían considerar también entrenadores similares a padres que posean diferentes características de enseñanza.

MÁS INFORMACIÓN

El presente reporte está basado en los siguientes artículos previamente publicados: *Multi-modal Integration of Dynamic Audiovisual Patterns for an Interactive Reinforcement Learning Scenario, publicado en los Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) [9] y Multi-modal Integration of Speech and Gestures for Interactive Robot Scenarios, publicado en IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics*. Para mayores detalles, por favor consultar los artículos mencionados donde además es posible encontrar experimentos adicionales aplicados a un escenario robótico de aprendizaje por refuerzo interactivo.

BIBLIOGRAFÍA

- *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.*
- *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.*
- *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.*
- *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.*
- *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.*
- *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.*
- *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.*
- *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.*
- *Francisco Cruz, German I. Parisi, Johannes Twiefel, and Stefan Wermter. "Multi-modal Integration of Dynamic Audiovisual Patterns for an Interactive Reinforcement Learning Scenario". In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 759-766, Daejeon, Korea, 2016.*
- *Francisco Cruz, German I. Parisi, and Stefan Wermter. "Multi-modal Integration of Speech and Gestures for Interactive Robot Scenarios". IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics, Daejeon, Korea, 2016.*