Doctoral Consortium: Desarrollo de sistemas cognitivos para la interacción multimodal humano-robot

Cristina Romero González

Departamento de Sistemas Informáticos. Universidad de Castilla-La Mancha.
Campus Universitario s/n. Albacete. Spain
Cristina.RGonzalez@uclm.es

- **Directores:** Ismael García Varea, Jesús Martínez Gómez
- **Año de inicio de la tesis:** 2013

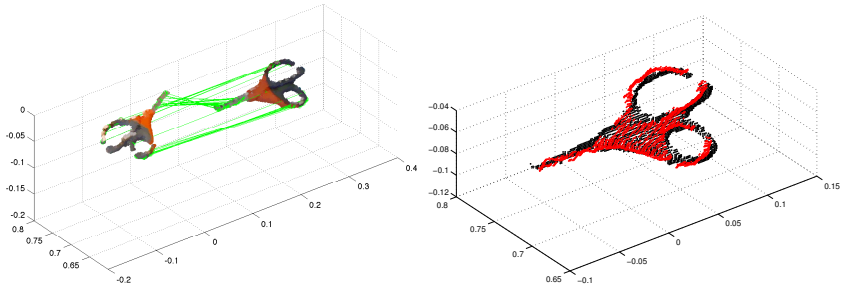
1. Resumen

Tradicionalmente los robots han tenido poca interacción con los humanos. Durante mucho tiempo la mayoría de los robots estaban ubicados en industrias y la única comunicación existente consistía en la programación y supervisión del trabajo de la máquina por parte del operario humano. Sin embargo, dado el cambio de paradigma en la investigación con robots y el auge de la robótica social, es necesario replantearse diferentes estrategias para conseguir mecanismos de interacción humano-robot (HRI) eficientes [3]. La interacción, por definición, requiere de una comunicación entre dos entidades: el ser humano, y la máquina. En este proyecto, se investigará la interacción necesaria cuando ambas entidades se encuentran en un mismo espacio físico, donde es necesario tener en cuenta todos los datos de entrada obtenidos a partir de distintos sensores, tanto visuales como de audio. Este tipo de comunicación es, por tanto, esencialmente multimodal.

En general, uno de los mayores desafíos actuales para un robot doméstico es entender dónde está para comportarse e interactuar con su entorno de forma correcta. Este problema se conoce como clasificación de escenas de interior [12], y puede definirse como el problema de describir el lugar donde el robot se encuentra en ese momento a partir de unas etiquetas predefinidas (por ejemplo, baño, pasillo o cocina). El proceso consiste en capturar una imagen, generar una representación adecuada (descriptor de la imagen), y clasificar (etiquetar) la escena. Normalmente, esto se lleva a cabo usando una representación semántica de la escena, llamada Bag-of-Words (BoW [2]), junto a una pirámide espacial [6] que codifica la distribución geométrica de los objetos en la escena.

Otro de los retos actuales es el registrado de objetos (ver Figura 1), que consiste en encontrar las zonas comunes en dos o más imágenes y calcular la transformación que las uniría. En este proceso se pueden identificar tres pasos críticos. Primero, deben encontrarse las regiones de interés de las imágenes y sus correspondientes puntos clave [9]. En segundo lugar, se deben generar las características o descriptores que codifiquen la información del entorno del vecindad

local de esos puntos clave [1]. Y, por último, se debe calcular la transformación geométrica que, basada en las partes comunes detectadas, una las imágenes superponiendo dichas partes [7]. Esta transformación proporciona información básica sobre la localización de los objetos en la escena, su posición exacta y rotación, que son indispensables para tareas como manipulación de objetos [10]. Siguiendo un procedimiento similar, la generación de descriptores locales se puede utilizar para identificar la categoría del propio objeto. En este caso, hablamos de reconocimiento o clasificación de objetos [5].



(a) (b)

Figura 1. Ejemplo de registrado de objetos. a) muestra las correspondencias de puntos y b) el registrado de los objetos.

También tiene especial importancia para una correcta interacción, que el robot sea capaz de realizar un aprendizaje continuo de los elementos del entorno. Así, el modelo utilizado para clasificar debe ser capaz de reconocer nuevas instancias de objetos o escenas para adaptarse. En otras palabras, debe realizarse un aprendizaje online [11] que incremente el conocimiento que tiene el robot sobre el problema durante su funcionamiento.

Por último, debemos considerar que la principal forma de interacción entre seres humanos está basada en el uso del lenguaje hablado. Debido a esto, en las últimas décadas se han empleado numerosos esfuerzos para conseguir adaptar esta modalidad como interfaz hombre-máquina [8]. En la actualidad, el paradigma seguido para abordar sistemas de reconocimiento del habla es el del reconocimiento de formas utilizando técnicas estadísticas [4]. Así, la transcripción de un fragmento de habla se plantea como la búsqueda de la secuencia de palabras más probable que se puede obtener de una señal acústica de entrada. Por último, la comprensión del habla consiste en analizar la salida del reconocedor del habla con el objeto de extraer la información necesaria para el sistema. Normalmente consiste en un análisis sintáctico para determinar la estructura de la entrada reconocida y uno semántico con el que se busca su significado.

2. Metodología y Plan de Trabajo

Durante el desarrollo de esta tesis doctoral se pretende conseguir los siguientes objetivos:

- **Sistemas de visión:** Dada la importancia de este tipo de sistemas se pretende desarrollar un sistema capaz de reconocer objetos y escenas en tiempo real. Normalmente, esta tarea suele investigarse desde el punto de vista de la comunidad de visión por computador, y aunque cada vez más son relevantes los aspectos de rendimiento y eficiencia, estos resultan fundamentales en el área de robótica. Entre las tareas que se van a desarrollar destacan:
  - Desarrollo de un sistema de detección de puntos clave que tenga en cuenta la información 3D del entorno para detectar las zonas más relevantes en la imagen. Este paso inicial reduce el espacio de búsqueda, de forma que los pasos posteriores se puedan centrar en las partes más representativas de los datos capturados.
  - Generación de descriptores locales que permitan representar el entorno de vecindad de los puntos clave extraídos anteriormente. Este paso suele ser uno de los más costosos del proceso y, aunque se han realizado propuestas orientadas a acelerar el cómputo de los descriptores, la mayoría se centran en el uso de información 2D. Por este motivo, durante el desarrollo de la tesis doctoral se propone el desarrollo de un nuevo descriptor basado en la información 3D y que busque optimizar al máximo los recursos necesarios.
  - Generación de representaciones globales del entorno que permitan la clasificación de los elementos y las escenas que se encuentren en el mismo. A partir de los descriptores locales se debe crear un nuevo descriptor global que codifique toda la información que aparece en la escena para su posterior clasificación. Entre otras técnicas se plantea la generalización de técnicas que han demostrado funcionar bien 2D para su uso con imágenes RGB-D.
  - Uso de los sistemas generados anteriormente para el registrado y clasificación de objetos y la clasificación de escenas.
- **Reconocimiento del habla:** Para esta parte del proyecto de tesis se pretende construir un sistema de reconocimiento capaz de identificar comandos sencillos e información básica sobre el entorno. También es interesante que el robot sea capaz de realizar preguntas para poder así actualizar sus modelos de clasificación en el aprendizaje.
  - Desarrollo de un sistema de reconocimiento y compresión del habla con un vocabulario limitado a instrucciones dentro de un entorno doméstico. En concreto, interesa que el robot sea capaz de reconocer instrucciones sencillas relativas a los objetos y lugares que es capaz de reconocer de forma visual.
  - Desarrollo de un sistema de conversación básico que permita solicitar información adicional sobre las instrucciones que reciba y realizar preguntas sobre el entorno para aumentar su base de datos de conocimiento.

Estas tareas deben generar preguntas para las que se esperan respuestas concretas.

- **Integración multi-modal:** De los sistemas anteriores, el robot recibirá multitud de información sobre su entorno, las personas con las que interactúa y las acciones que se esperan de él. Para que todos esos datos se gestionen de forma adecuada, el robot debe ser capaz de integrar toda la información, tanto visual 2D y 3D, como auditiva. En general, se debe gestionar una base de datos con todo el conocimiento disponible de forma que la interacción humano-robot sea lo más fluida posible.
- **Aprendizaje online:** El uso de un robot móvil en un caso real requiere que éste sea capaz de adaptarse. Para ello se propone el uso de sistemas de clasificación online, de forma que, ante nuevas situaciones, el robot sea capaz de aprender qué es nuevo en su entorno. Además, el sistema conversacional permitirá que el robot realice preguntas al humano sobre los datos desconocidos para poder realizar un aprendizaje activo de los nuevos datos detectados.

Inicialmente, se pretende desarrollar los dos primeros objetivos de forma secuencial: primero se realizará el procesamiento de imágenes para el sistema de visión y después el procesamiento del audio para el reconocimiento del habla. Mientras se desarrollan estos objetivos, se pretende realizar de forma paralela los dos últimos objetivos, ya que dependen de la información que se extraiga de los otros bloques.

3. Relevancia

Este proyecto de tesis pretende desarrollar e integrar varios sistemas imprescindibles para la interacción multimodal humano-robot. En general, se pretende que un robot como el actual Loki (ver Figura 2) sea capaz de comprender qué está viendo y qué está escuchando en un entorno doméstico. Todo esto son temas de especial relevancia en la actualidad, como demuestran por ejemplo las múltiples competiciones en el ámbito de la robótica con esta temática (por ejemplo, RoCKIn[1]).

En general, la robótica social es un campo en auge, en el que se espera que en los próximos años se produzcan avances importantes que permitan su uso cotidiano de robots en ámbitos domésticos. La finalidad de este proyecto es colaborar y aportar avances en las ramas de investigación asociadas a estas tareas.

Referencias

1. Alexandre, L.A.: 3D Descriptors for Object and Category Recognition: a Comparative Evaluation. In: Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (October 2012)

[1] http://rockinrobotchallenge.eu/

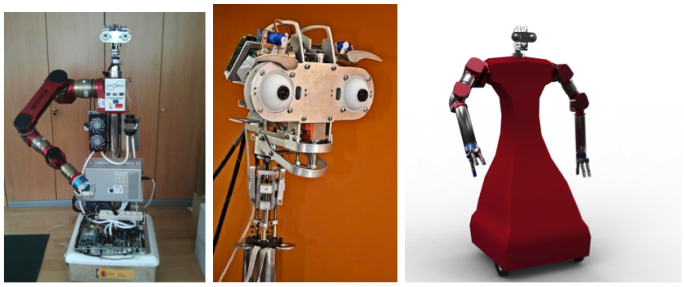


Figura 2. Foto actual de Loki (izquierda). Detalle de la cabeza (centro). Futuro aspecto de Loki con 2 brazos (derecha).

2. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)
3. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems 42(3–4), 143 – 166 (2003), socially Interactive Robots
4. Jelinek, F.: Statistical methods for speech recognition. MIT press (1997)
5. Lai, K., Bo, L., Ren, X., Fox, D.: A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 1817–1824 (May 2011)
6. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. vol. 2, pp. 2169–2178 (2006)
7. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI. vol. 81, pp. 674–679 (1981)
8. McTear, M.F.: Spoken Dialogue Technology: Enabling the Conversational User Interface. ACM Comput. Surv. 34(1), 90–169 (Mar 2002)
9. Salti, S., Tombari, F., Di Stefano, L.: A Performance Evaluation of 3D Keypoint Detectors. In: 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT). pp. 236–243 (May 2011)
10. Stückler, J., Steffens, R., Holz, D., Behnke, S.: Efficient 3D object perception and grasp planning for mobile manipulation in domestic environments. Robotics and Autonomous Systems 61(10), 1106–1115 (2013), selected Papers from the 5th European Conference on Mobile Robots (ECMR 2011)
11. Tao, Y., Triebel, R., Cremers, D.: Semi-supervised Online Learning for Efficient Classification of Objects in 3D Data Streams. In: Proc. IEEE International Conference on Intelligent Robots and Systems (IROS) (2015)
12. Wu, J., Christensen, H.I., Rehg, J.M.: Visual place categorization: Problem, dataset, and algorithm. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009. pp. 4763–4770. IEEE (2009)