Convocatoria de ayudas de Proyectos de Investigación Fundamental no orientada

MEMORIA TÉCNICA PARA PROYECTOS TIPO A o B

1. RESUMEN DE LA PROPUESTA (Debe rellenarse también en inglés)

INVESTIGADOR PRINCIPAL: Antonio Bandera Rubio

TÍTULO DEL PROYECTO: ViPS: Sistema de percepción visual para interacción hombre-robot y navegación de robots móviles

RESUMEN
(breve y preciso, exponiendo solo los aspectos más relevantes y los objetivos propuestos)

Una de las areas básicas de la robótica actual está relacionada con el desarrollo de los denominados robots sociales o de servicio, cuya finalidad es la de ayudar a las personas que los rodean a realizar sus tareas diarias. Ahora que los robots de servicio buscan compartir con las personas un mismo espacio, deberan aprender a interactuar con éstas de una manera intuitiva, incluso mientras llevan a cabo otras tareas. Para conseguirlo, resulta interesante que el robot sea capaz de percibir el mundo que lo rodea de una manera similar a como las personas lo hacen, generando para ello, a partir de la información recibida, un conjunto de representaciones internas que sean consistentes con las representaciones humanas.

En este proyecto se pretende estudiar e implementar el sistema de percepción visual de un robot autónomo. El elemento central de este sistema será un mecanismo atencional que deberá ser capaz de discriminar, de toda la información captada por los sensores del robot, los elementos de mayor relevancia para finalizar correctamente las tareas en curso. Normalmente, este proceso es llevado a cabo generando un mapa de importancia para cada característica percibida, que asociará valores elevados a las regiones de la escena que son interesantes y valores bajos al resto. En nuestro caso, estos mapas serán pesados por las tareas actualmente en proceso. De esta forma, la importancia de cada característica será determinada no sólo por su valor dentro de la escena, sino también por las tareas a realizar por el robot. Como novedad, en este proyecto todo este sistema pre-atentivo será implementado usando lógica borrosa. Si a esta etapa pre-atentiva se unen los módulos atencionales que se proponen en este proyecto, el sistema perceptivo desarrollado permitirá al robot generar dos representaciones internas sumamente importantes. Por un lado, este sistema detectará marcas visuales perceptivamente uniformes en un entorno inicialmente desconocido, generando a partir de éstas un mapa topológico del entorno en el que las marcas métricas se fusionarán en nodos que representarán un mismo espacio físico. Por otra parte, el sistema permitirá detectar y seguir el movimiento de la parte superior de la persona interesada en interactuar con el robot, proporcionando información sobre quien es la persona o qué gesto está llevando a cabo. Ambas fuentes de información permitirán al robot percibir su entorno, y serán adquiridas y actualizadas a distintas velocidades, en función de la secuencia de tareas llevadas a cabo. Finalmente, un robot de servicio no sólo necesita generar estas representaciones para ser útil, sino que, además, resulta igualmente importante que estas representaciones internas se correspondan con las representaciones que, del entorno o los gestos percibidos, tienen las personas que los rodean. En este proyecto se proponen procesos de interacción hombre-robot que permitirán incorporar a las representaciones internas que genera de forma autónoma el robot, conceptos e información semántica dada por un instructor. En particular, este proceso de interacción se ha extendido, en lo referente a la percepción orientada a la persona, con un módulo que permite al robot imitar los gestos percibidos. Esta herramienta permitirá, en el futuro, que el robot sea capaz de expresarse con las personas que lo rodean usando mecanismos de comunicación no-verbales.

PROJECT TITLE: ViPS: Visual perception system for human-robot interaction and mobile robotic navigation

SUMMARY
(brief and precise, outlining only the most relevant topics and the proposed objectives)

A key area in robotics research is concerned with developing social robots for assisting people in everyday tasks. Now that robots are moving into public places or homes, they must interact with people in an intuitive manner at the same time that they are developing other tasks. To achieve this intuitive interaction, it is interesting that the robot will be able to perceive the real world in a similar way that people do, extracting from the sensed data a set of internal representations which will be consistent with the human representations.

In this project, we focus on the study and develop of the robot's visual perception system. The central part of this system will be an attention mechanism which must be able to discriminate, from all the information provided by the robot's sensors, the most relevant elements needed to carry out the currently executed tasks. Typically, this is conducted generating a map for each sensed feature, which contains high values for interesting regions and lower values for other regions. In our case, feature maps will be also weighted by the currently executed tasks. Thus, the importance of a sensed feature not only depends on its own value into the corresponding map, but also on the tasks to carry out. These weights will be established by a fuzzy system. The perception system will allow the robot to detect distinguished visual landmarks in an initially unknown environment, generating a hierarchical map which fuses these landmarks into a high-level topological map and to detect and capture the upper-body motion of people interested in interact with the robot, providing information about who is the person and what gesture is the person doing. Both mid-level sources of information will allow the robot to perceive the surrounding environment, and it will be acquired and updated at different rates depending on the sequence of executed tasks. However, mobile cognitive social robots not only need to have these internal representations to be useful. Thus, for a social robot it is also crucial that its internal representations will be consistent with the human representations. In this project, a human-robot interaction process will be conducted to incorporate to the internal representations of the robot the information that is given by a human intructor. Finally, this last issue will be extended to include a module which will allow the robot to imitate the perceived upper-body gestures.

2. INTRODUCCIÓN
(máximo 5 páginas)

Las personas muestran una habilidad extremadamente robusta para extraer la información más relevante del entorno. Aunque esta habilidad les permite, por ejemplo, desplazarse en entornos desconocidos, está especialmente orientada a los procesos de interacción social o interpretación del comportamiento de otras personas. La implementación de sistemas artificiales de percepción, que exhiban esta misma habilidad, se ha convertido en un paso previo necesario en el desarrollo de robots que cooperen con las personas que los rodean como ayudantes útiles, capaces de entender órdenes expresadas de forma natural, y para los que la interacción es un proceso intuitivo, pero que, además, son capaces de simultanear esta percepción orientada a la interacción con la capacidad para realizar otras tareas, como navegar en un entorno inicialmente desconocido o coger un determinado objeto. Actualmente, sin embargo, no es todavía habitual que un robot demuestre este tipo de habilidades. Aunque se asume que el sistema perceptivo del robot debería imitar la habilidad de los sistemas de visión de los seres vivos más avanzados para seleccionar la información más relevante del conjunto de datos visuales, este sistema selectivo es normalmente un proceso que no depende del conjunto de tareas a llevar a cabo. Por ello, suele ser necesario que el sistema mueva secuencialmente el denominado foco de atención por todas las regiones relevantes detectadas, para así tratar de extraer de la escena toda la información de interés, entre la que se encontrará la necesaria para finalizar con éxito las distintas tareas en curso. Al no haber relación con las tareas a realizar, este proceso puede ser lento, visitándose muchas zonas que, en realidad, no son interesantes, y, por ello, aunque finalmente el robot encuentre la información buscada, la velocidad del proceso obliga a dejar de lado otras cuestiones, como la de interactuar de una forma socialmente correcta con las personas, que el robot encontrará en su entorno mientras esté resolviendo estas otras tareas. Esta cuestión se convierte en un problema cuando los robots se diseñan y crean específicamente para cooperar con las personas como compañeros socialmente interactivos (Fong et al, 2002), pues a la dificultad de tener que manejar una gran cantidad de estímulos, se une la necesidad de que el robot responda a estos estímulos perceptivos a una velocidad similar a la humana, pues las personas que lo rodean así se lo van a exigir.

El presente proyecto propone estudiar y desarrollar un sistema de percepción visual para un robot de servicio o social, que le permitirá desarrollar tareas relacionadas con la navegación y la interacción hombre-robot. Un robot de servicio o robot social es un agente autónomo, que no sólo es capaz de navegar o resolver otras tareas típicamente exigidas a un robot móvil, sino que es capaz de simultanearlas con la comunicación e interacción con otros agentes socialmente activos, personas o robots, de su entorno. Esto implica, por un lado, que el robot social deberá percibir de manera simultánea una gran variedad de estímulos, que le llegarán, fundamentalmente, por los canales visual o auditivo, y, por otro, que la respuesta a dichos estímulos deberá ser a velocidades similares a las que se espera de una persona. Además, para conseguir interactuar de forma intuitiva con las personas, será interesante que el robot sea capaz de percibir el mundo que lo rodea de forma similar a como lo hacen las personas, extrayendo de la información que proporcionan los sensores, un conjunto de representaciones internas que sean coherentes, en la medida de lo posible, con las representaciones que extraería una persona. Así, un robot socialmente interactivo interpretará los mismos fenómenos que las personas observan (Dautenhahn and Nehaniv, 2002). El elemento central de nuestra propuesta es un mecanismo atencional artificial que será capaz de discriminar, de toda la información de bajo nivel proporcionada por los sensores del robot, aquellos datos que son relevantes para satisfacer las tareas que, en ese preciso instante de tiempo, éste debe llevar a cabo. En este sentido, la primera contribución importante del sistema atencional propuesto es la inclusión, en el proceso de selección de información, de una etapa semi-atentiva, que tendrá en cuenta las tareas a realizar. De esta forma, la importancia del conjunto de datos adquiridos no depende sólo de características de bajo nivel que no están relacionadas con las tareas a llevar a cabo, sino que también dependerá de éstas. La combinación de la información de bajo nivel con la información relativa a las tareas será llevada a cabo usando un sistema de lógica borrosa. Una vez que la información relevante del entorno ha sido extraída, será usada por una etapa atentiva para construir representaciones jerárquicas tanto del entorno como de las personas detectadas. Estas representaciones, que se pueden considerar como de nivel medio, serán la salida final del mecanismo atencional propuesto y, para facilitar la interacción con las personas, incluirán información semántica y no sólo relaciones de tipo espacial. La segunda contribución importante del presente sistema será la integración de las capacidades perceptivas del robot con procesos de interacción hombre-robot, que permitirán a éste anotar las representaciones internas percibidas con información semántica de manera supervisada. La importancia de este tipo de procesos ha sido positivamente valorada en la literatura. Así, esta información semántica permite razonar la funcionalidad de objetos y entornos, o proporcionar una entrada adicional a los módulos encargados de llevar a cabo tareas relacionadas con la navegación o la localización. En cualquier caso, es importante proporcionar al robot datos que le permitan comunicarse con las personas usando un mismo conjunto de términos y conceptos (Galindo et al, 2005). En los siguientes apartados de esta introducción se presentan los principales objetivos del proyecto, ofreciéndose un esbozo del estado de los conocimientos científicos en este campo de la investigación.

El mecanismo atencional como elemento central del sistema perceptivo visual

Uno de los modelos teóricos más influyentes desarrollados para explicar la atención visual es la denominada teoría de la integración de características, un modelo biológicamente inspirado que se propone explicar el mecanismo innato de la atención en la visión humana (Treisman and Gelade, 1980). Según este modelo, el proceso atencional se lleva a cabo en dos etapas principales. En primer lugar, en una etapa pre-atentiva, en la que las tareas a llevar a cabo no influyen, se

estiman un determinado número de características de bajo nivel de forma paralela. Las características extraídas se integran en un único mapa de importancia, que codifica el interés de cada región percibida en la escena. De este mapa se extraerán las regiones más importantes. Entonces, en una segunda etapa atentiva que si está modulada por las tareas en proceso, se moverá secuencialmente el foco de atención visual sobre cada región seleccionada, para así analizarla. Para evitar mover dicho foco de atención sobre una región ya visitada recientemente, las regiones analizadas se inhiben durante un tiempo, proceso que se lleva a cabo incluyéndolas en el denominado mapa de inhibición. De esta forma, mientras que la etapa atentiva debe ser redefinida en función de la aplicación, la etapa pre-atentiva se puede considerar como general e independiente de la aplicación final. Aunque los modelos implementados siguiendo estos fundamentos teóricos funcionan correctamente en entornos estáticos, presentan problemas para manejar escenarios dinámicos, pues no tienen en cuenta el movimiento o las oclusiones de los objetos presentes en la escena. Para resolver este problema, el mecanismo atencional deberá integrar información sobre profundidad y movimiento, para así poder seguir el desplazamiento de los objetos dinámicos (Backer y Mertsching, 2003). En esta línea de investigación, Maki et al (2000) propusieron un mecanismo atencional que incorporaba ambas fuentes de información entre las características evaluadas para generar el mapa de importancia. Backer y Mertsching (2003) también emplean la profundidad como característica de bajo nivel, pero proponen usar redes neuronales dinámicas para seguir el movimiento de las regiones extraídas del mapa de importancia. La estructura atencional que propone este trabajo resulta muy interesante, aunque el gasto computacional que implican las herramientas propuestas impide su aplicación en sistemas interactivos rápidos (el sistema tarda unos 30 segundos por imagen procesada). Finalmente, el problema de tratar con entornos dinámicos ha sido analizado por nuestro grupo de investigación en el marco de los proyectos TIN2005-01359 y VISOR, subvencionados por el Ministerio de Educación y Ciencia y la Red de Excelencia EURON (*European Robotics Network*), respectivamente. El mecanismo atencional propuesto en estos trabajos previos incluía una etapa semi-atentiva, ubicada entre las etapas pre-atentiva y atentiva, en la que las regiones previamente extraídas del mapa de importancia eran seguidas para así evitar el catalogarlas como nuevas regiones de interés. De esta forma, esta etapa genera y actualiza una lista de regiones importantes, que estará disponible para el resto de módulos del sistema (Vázquez-Martín et al., 2005; Marfil et al, 2006). Sin embargo, estos trabajos previos presentan la importante limitación de que fueron abordados de forma que, finalmente, todo el sistema atencional sólo puede satisfacer una única tarea específica: extraer marcas visuales naturales del entorno o capturar el movimiento de la parte superior del cuerpo de una persona, pero no ejecutar ambas de forma simultánea.

En este proyecto se propone el desarrollo de un mecanismo atencional que extienda estos trabajos previos para permitir al robot generar unas representaciones internas, tan completas como sea posible, del mundo percibido. Particularmente, este sistema perceptivo estará especialmente diseñado para reconocer cuando y por cuanto tiempo está una persona interesada en establecer un proceso de interacción y para permitir al robot localizarse y, simultáneamente, crear un mapa de un entorno inicialmente desconocido. Para ello, los módulos que forman la etapa atentiva del sistema propuesto tendrá en cuenta tanto elementos estáticos – marcas visuales del entorno- como dinámicos – partes del cuerpo de la persona-, para generar representaciones jerárquicas del mapa del entorno y de las personas detectadas que puedan ser, posteriormente, empleadas por los módulos de más alto nivel de la arquitectura de control. De forma resumida, se puede entender que el mecanismo atencional mezcla un proceso que integra información de bajo nivel para generar conceptos de nivel medio, determinando las regiones de interés a partir de características de bajo nivel (etapa pre-atentiva), con otro proceso que usa conceptos de alto nivel para catalogar la importancia de estos conceptos de nivel medio, empleando modelos previos de los objetos demandados por las representaciones que se están generando, para así filtrar los datos percibidos y seguir sólo los que se consideren de interés (etapa semi-atentiva). El algoritmo de seguimiento manejará objetos dinámicos y tendrá en cuenta posibles oclusiones. Para llevar a cabo este proceso, las representaciones jerárquicas incluirán modelos que se asociarán a la apariencia y movimiento de los objetos, y que han sido recientemente propuestos en nuestro grupo (Marfil et al, 2006b; Marfil et al, 2007). De esta forma, se puede entender que el sistema propuesto sigue tres pasos principales previos a la propia etapa atentiva final: cálculo en paralelo de mapas de características, integración de características y seguimiento simultáneo del conjunto de regiones de interés más importantes. Como una aportación novedosa respecto a nuestros trabajos previos, la integración de características será llevada a cabo usando lógica borrosa. La lógica borrosa ha sido previamente aplicada a la robótica, tanto en el diseño y coordinación de comportamientos de control en navegación, como en la construcción de mapas del entorno o en la integración de capas de control deliberadas y reactivas. En nuestro caso, el sistema difuso generará el valor final de importancia de cada región de interés seleccionada en la etapa semi-atentiva. Este valor no sólo dependerá, por tanto, de las características de bajo nivel, sino que también estará determinado por las tareas a ejecutar en ese momento por el robot. Finalmente, las regiones seleccionadas en la etapa semi-atentiva serán empleadas por la tercera y última etapa del mecanismo atencional: la etapa atentiva. Como se ha comentado, en este proyecto esta etapa actualizará, por cada imagen de entrada, dos representaciones internas. La estructura global de una de ellas será proporcionada a priori: la representación o modelo de la apariencia humana. En este caso, el procesado de la secuencia de imágenes de entrada permitirá que dicho modelo almacene y filtre, de forma rápida, el movimiento de la parte superior del cuerpo de la persona (Molina-Tanco et al, 2005). La otra representación interna será inicialmente desconocida: el mapa del entorno. Este mapa se organizará en dos niveles de abstracción. El nivel bajo codificará la información métrica asociada al conjunto de marcas visuales naturales, y el nivel superior contendrá una representación topológica del entorno. Finalmente, para permitir al robot comunicarse con otras personas usando un mismo conjunto de términos, se anotarán dichas representaciones con información semántica. Este proceso será abordado a través de la interacción del robot con un supervisor humano. En la parte del proceso perceptivo que está relacionada con la detección y seguimiento del movimiento de las personas, este

proceso de interacción será extendido, incluyendo un módulo que permitirá al robot imitar los gestos percibidos. De esta forma, el robot no sólo memorizará conceptos semánticos relativos al nombre con el que se designa una habitación o un determinado gesto, sino que también memorizará como llevar a cabo un gesto específico. En los dos siguientes apartados se analizan con más detalles los dos procesos de atención descritos y relativos a la captura y reconocimiento de bajo nivel de la actividad de la persona interesada en interactuar con el robot y a la representación jerárquica del entorno.

Etapa atentiva: captura del movimiento de la persona y reconocimiento e imitación de gestos

El desarrollo de un robot social implicará la generación de interfaces naturales de interacción hombre-robot. Sin embargo, dado que estos robots desarrollaran su actividad en entornos en los que distintas personas se estarán moviendo en torno a él, será necesario proporcionar al robot los mecanismos que le permitan, de forma autónoma, percibir cuando una determinada persona está interesada en establecer una interacción. Además, cuando este proceso se establezca, el robot deberá comprender qué expresan los términos no-verbales que suelen complementar cualquier proceso humano típico de comunicación verbal. Estos términos no-verbales no sólo incluyen expresiones faciales, sino que también comprenden gestos y posturas del cuerpo. En particular, este proyecto propone un sistema de percepción visual que permitirá al robot detectar cuando una persona tiene interés real en interactuar con él, así como reconocer e imitar gestos descritos por movimientos de la parte superior del cuerpo de la persona.

Para satisfacer estos objetivos, será inicialmente necesario que el robot social sea capaz de detectar la presencia de personas en su entorno y de seguir sus movimientos, usando para ello sólo los sensores ubicados sobre el propio robot. Además, la resolución de estas tareas se complica dado que, compartiendo un mismo entorno, pueden encontrarse varias personas, cuyos movimientos y trayectorias se cruzarán y ocluirán. Entre los distintos trabajos que recientemente se han desarrollado sobre esta temática, los ya mencionados proyectos TIN2005-01359 y VISOR permitieron a nuestro grupo abordar estas tareas. En el marco de trabajo de estos proyectos se desarrolló y verificó, como un elemento fundamental de los mismos, un sistema de percepción visual que permite la captura del movimiento de la parte superior de una persona situada frente al robot, usando para ello sólo la información proporcionada por un sistema de visión estéreo (Bandera et al, 2006; Bandera et al, 2007). Así, este sistema de captura funciona sin emplear marcadores o dispositivos especiales, basando su rapidez en la hipótesis de que, para poder interpretar los movimientos de la persona, basta con seguir el movimiento de la cabeza y las manos, pues son éstos los elementos no-verbales más importantes en cualquier proceso de interacción entre personas. La presencia de la persona se detecta integrando la información proporcionada por un detector del color de la piel, un detector de caras y un mapa de profundidad. Estas características eran combinadas por la etapa pre-atentiva de un mecanismo atencional, que era el responsable final de encontrar la cara y manos de la persona (Marfil et al, 2006). Además, y dado que la información proporcionada por el proceso de seguimiento podía ser ocasionalmente inestable o incompleta por oclusiones o fallos en la estimación de la profundidad, se incorporó al sistema un modelo o representación interna que permitía filtrar los datos de posiciones (Molina-Tanco et al, 2005; Molina-Tanco et al, 2006). En nuestra nueva propuesta, las etapas pre-atentiva y semi-atentiva serán las responsables de determinar si una persona está interesada en interactuar con el robot. Aunque se podrían integrar en el sistema otras fuentes de información, p. ej. la localización de la fuente sonora o el reconocimiento de voz, esta propuesta sólo tendrá en cuenta información visual. Así, las características que se manejan para determinar la presencia de una persona con interés en interactuar serán las empleadas en la versión previa, esto es, detectores del color de la piel y de caras, y mapas de profundidad. Las regiones de la imagen más importantes según estas características, asociadas a posibles caras, serán seguidas por la etapa semi-atentiva. Finalmente, en la etapa atentiva, cuando el foco de atención se mueve a una posible cara, se completará la percepción de la persona mediante la detección de otras partes de su anatomía, como las manos o la cintura, y se lanzará un algoritmo que seguirá los gestos que dicha persona lleve a cabo. Este algoritmo analizará el movimiento seguido por distintas partes del cuerpo para reconocer gestos, comparándolos con los almacenados en una base previamente creada.

A la hora de reconocer los gestos observados, el movimiento de las distintas partes del cuerpo puede ser caracterizado por sus correspondientes trayectorias 3D en coordenadas cartesianas. Así, este tipo de descripción ha sido satisfactoriamente empleado en un sistema de aprendizaje que permite al robot reconocer y aprender gestos realizados por la persona con sus dos manos (Bandera et al, 2006). Sin embargo, cuando la longitud del gesto o el conjunto de elementos a seguir aumenta, el descriptor puede alcanzar un tamaño excesivo. Este problema se puede resolver representando las trayectorias completas por un conjunto reducido de parámetros significativos. En la literatura, esta selección se puede llevar a cabo usando parámetros que caractericen las trayectorias desde un punto de vista global, definidos en relación a una referencia externa, o usando parámetros locales de las trayectorias, basados normalmente en medidas diferenciales (curvatura o torsión). Si los parámetros globales son más robustos al ruido, los locales son mejores para discriminar detalles más finos de las trayectorias. Por otra parte, el sistema de reconocimiento de gestos también deberá afrontar el problema de comparar las trayectorias percibidas con las previamente memorizadas. Esta etapa de comparación deberá tener en cuenta las características propias de las trayectorias 3D, tales como las distintas tasas de muestreo, la presencia de datos provenientes de fallos en el proceso de seguimiento, o las diferencias de tamaño de las secuencias a comparar. Actualmente, se puede considerar que el uso de campos ocultos de Markov (*hidden Markov models*, HMMs) constituye la técnica de modelado más empleada en reconocimiento de gestos, y así lo demuestra su uso por distintos grupos de investigación (Calinon and Billard, 2004; Asfour et al, 2006). Sin embargo, su uso también acarrea determinados problemas. Así, el número y precisión de los gestos que pueden ser modelados usando HMMs estará limitado por la complejidad de los algoritmos de entrenamiento y estimación de probabilidades de transición. Además, la creación de nuevos campos implicará

volver a estimar tanto la correspondencia entre estados y observaciones como todo los HMMs para todos los otros gestos, siendo por tanto necesario incluir inicialmente todas las posibles observaciones si el sistema pretende aprender nuevos gestos de forma interactiva (Calinon and Billard, 2004). Otra forma de comparar trayectorias 3D, que han sido profusamente empleadas, son las técnicas basadas en programación dinámica. Estas técnicas se implementan como funciones para estimar, de forma rápida, la distancia entre dos trayectorias. Particularmente, en este proyecto se propone describir cada una de las trayectorias que componen un determinado gesto por una secuencia de puntos significativos, que serán seleccionados en función de una estimación local de la curvatura. En la etapa de comparación, se evaluarán y compararán distintas funciones de distancia basadas en algoritmos de programación dinámica, para así elegir la más adecuada para reconocer gestos usando la representación propuesta. Finalmente, en caso de que sea necesario, la función de distancia podrá ser complementada por medidas de parecido entre parámetros globales extraídos de las trayectorias a comparar.

Finalmente, como se ha comentado previamente, resulta interesante que el robot social no sólo sea capaz de generar una representación interna de la persona con la que está interactuando y de la cuál puede reconocer qué gesto está haciendo, sino que será igualmente importante que este conocimiento sea consistente con el que podría percibir una persona que estuviera en la situación del robot. En esta propuesta, esta consistencia será adquirida a través de un proceso de interacción hombre-robot, en el que la persona podrá enseñar al robot la base interna de posibles gestos, anotándolos semánticamente. Además, este proceso de interacción se reforzará con la inclusión de un módulo que permitirá al robot imitar los gestos percibidos (Bandera et al, 2006; Bandera et al, 2007). De esta forma se pretende que el robot pueda mostrar a la persona, a su vez, el gesto que ha aprendido, memorizando en una misma estructura los procesos de reconocimiento y ejecución.

Etapa atentiva: Representación jerárquica del entorno basada en percepción visual

La problemática de la navegación en el campo de la robótica móvil necesita de una representación interna del entorno. Aunque los métodos convencionales han utilizado habitualmente mapas métricos basados en características, donde se identifica cada localización en el entorno con un conjunto de características y su distribución espacial, esas representaciones están limitadas por el incremento de la complejidad computacional y el coste de almacenamiento que suponen el aumento del número de características. Puesto que la complejidad del mapa global asociado a grandes entornos no puede ser evitada, algunos investigadores proponen utilizar una jerarquía de mapas para representar esos entornos a diferentes resoluciones o niveles de abstracción. Normalmente, se usan dos niveles de abstracción: un mapa de bajo nivel y uno de alto nivel o topológico. El mapa de alto nivel podrá ser empleado para representar grandes entornos, como por ejemplo como un grafo que conecte habitaciones y corredores en un edificio. Este tipo de mapas se pueden usar para trazar caminos abstractos que permitan navegar de una habitación a otra, sin tener en cuenta los detalles métricos exactos de estas habitaciones. El mapa de bajo nivel podrá emplearse para navegar de manera precisa de una habitación a la siguiente o para definir posiciones a alcanzar dentro de una habitación particular, sin tener en cuenta para ello al resto de habitaciones. Por otra parte, será también interesante que los elementos que componen esta representación jerárquica del entorno se asocien a elementos espaciales con significación, que pueden se empleados por el robot en procesos posteriores de interacción con las personas. Esto sería coherente con la hipótesis de desarrollar robots que sean capaces de percibir el mundo real de forma parecida a como lo hacen las personas, pues existen evidencias de que las personas también describen internamente el entorno usando representaciones jerárquicas (McNamara, 1986).

El trabajo de investigación en este campo se concentra en el uso de datos de bajo nivel obtenidos de sensores (apariencia, características visuales o medidas de rango o profundidad) para la representación jerárquica, el uso de objetos para el reconocimiento espacial y el empleo de interacción humana para definir conceptos. En los últimos años se han desarrollado algunas propuestas con objeto de construir una representación jerárquica del espacio agrupando datos obtenidos mediante visión. Por ejemplo, el grupo de Sistemas Autónomos Inteligentes (*Intelligent Autonomous Systems*) de la Universidad de Amsterdam ha estudiado el uso de técnicas basadas en la apariencia para agrupar imágenes. De esta forma, el mapa de bajo nivel se construye como un grafo, donde un conjunto filtrado de las imágenes capturadas constituyen un conjunto de nodos y la similitud entre imágenes definen arcos entre esos nodos. El objetivo de esta técnica es que el grafo contenga la distribución espacial del entorno dada por paredes u otro tipo de elementos. Por tanto, puede utilizarse un algoritmo para agrupar imágenes, obteniéndose una representación del entorno de alto nivel (Zivkovic et al, 2007). Sin embargo, una propuesta más intuitiva en cuanto a interpretación del entorno es modelar el mundo real en términos de objetos y la forma en que se relacionan entre ellos. Según este esquema, los nodos del grafo de bajo nivel son objetos reconocibles, que pueden agruparse para proporcionar nodos para grafos de alto nivel. Otro tipo de enfoque es el empleado en el laboratorio de Sistemas Autónomos (*Autonomous Systems Lab*) de la Escuela Politécnica Federal de Lausanne o en el LAAS-CNRS (Cottret and Devy, 2006), donde se emplean sistemas de detección de objetos para la caracterización espacial. Más concretamente, este último trabajo propone usar un mecanismo de atención visual para proporcionar estos objetos. Siguiendo una estrategia similar, nuestro grupo desarrolló un mecanismo de atención visual para detectar marcas naturales en entornos de oficinas o exteriores (Vázquez-Martín et al, 2005; Vázquez-Martín et al, 2006). Finalmente, dado que los robots se mueven cada vez más en lugares públicos y hogares, las personas deben tenerse en cuenta. Esto modifica la tarea de construir una representación del entorno, siendo necesario añadir información semántica a los datos obtenidos de los sensores. De esta forma, esto ayuda a conseguir una mejor representación (evitando problemas de solapamiento), y hace posible el comunicarse con humanos sobre su entorno.

En nuestra propuesta, se emplearán estas directrices para desarrollar una representación topológica/métrica del entorno. Como salidas, este enfoque proporcionará una representación jerárquica del entorno y la posición y orientación del robot dentro de ese mapa. Asumiendo que el robot está siempre localizado en un área del entorno, el sistema propuesto estimará la posición y orientación del robot y la estructura del mapa local a nivel métrico mediante la creación de un mapa estocástico basado en características, utilizando un filtro extendido de Kalman (EKF) (Vázquez-Martín et al, 2006b; Núñez et al, 2008). En este nivel, distintas regiones diferenciadas serán extraídas de la imagen de entrada por la etapa semi-atentativa y se emplearán como marcas naturales. Esas marcas serán asociadas a regiones perceptivamente uniformes de la imagen de entrada, cuyas formas o colores serán caracterizadas mediante variables lingüísticas. Estos descriptores permitirán asociar estas marcas visuales a las características almacenadas en el mapa de una forma más robusta. Por otro lado, también se implementará una técnica para detectar, de forma no supervisada, cambios perceptivos en la secuencia de imágenes adquiridas. Por último, en el mapa topológico, se empleará un modelo del entorno basado en grafos donde los nodos se relacionarán con áreas significativas del entorno, como habitaciones o pasillos. La información semántica dada por un supervisor humano será incorporada a este mapa de alto nivel. Siguiendo trabajos previos en este ámbito (Topp et al, 2006; Núñez et al, 2006), la interacción será directa, pues se asume que un "recorrido guiado" es la forma más apropiada para proporcionar al usuario la posibilidad de personalizar la representación topológica del entorno creada por el robot.

Referencias

(Fong et al, 2002) T. Fong, I. Nourbakshsh y K. Dautenhahn, A survey of social robots. Robotics and Autonomous Systems 42, 143–166, 2002.

(Dautenhahn y Nehaniv, 2002) K. Dautenhahn y C. Nehaniv, Imitation in animals and artifacts. MIT Press, Cambridge, USA, 2002.

(Galindo et al, 2005) C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.A. Fernández-Madrigal y J. González, Multi-Hierarchical semantic maps for mobile robotics. En Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Systems, 2005.

(Treisman y Gelade, 1980) A.M. Treisman y G. Gelade, A feature integration theory of attention. Cognitive Psychology 12(1), 97-136, 1980.

(Backer y Mertsching, 2003) G. Backer y B. Mertsching, Two selection stages provide efficient object-based attentional control for dynamic vision. In Proc. Int. Workshop Attention and Performance in Computer Vision, 2003.

(Maki et al, 2000) A. Maki, P. Nordlund y J.O. Eklundh, Attentional Scene Segmentation: Integrating Depth and Motion. Computer Vision and Image Understanding 78(3), 351-373, 2000.

(Vázquez-Martín et al, 2005) R. Vázquez-Martín, J.C. del Toro, A. Bandera y F. Sandoval. Data- and model-driven attention mechanism for autonomous visual landmark acquisition. In Proc. of the IEEE International Conference on Robotics and Automation, 2005.

(Marfil et al, 2006) R. Marfil, R. Vázquez-Martín, L. Molina-Tanco, A. Bandera y F. Sandoval, Fast attentional mechanism for a social robot. Workshop on Vision-Based Human-Robot Interaction, inside EUROS2006, 2006.

(Marfil et al, 2006b) R. Marfil, L. Molina-Tanco, A. Bandera, J.A. Rodríguez y F. Sandoval, Pyramid segmentation algorithms revisited. Pattern Recognition 39, 1430-1451, 2006.

(Marfil et al, 2007) R. Marfil, L. Molina-Tanco, J.A. Rodríguez y F. Sandoval, Real-time object tracking using bounded irregular pyramids. Pattern Recognition Letters 28(9), 985-1001, 2007.

(Molina-Tanco et al, 2005) L. Molina-Tanco, J.P. Bandera, R. Marfil y F. Sandoval, Real-time human motion analysis for human-robot interaction. In Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Systems, 2005.

(Molina-Tanco et al, 2006) L. Molina-Tanco, J.P. Bandera, J.A. Rodríguez, R. Marfil y F. Sandoval, A grid-based approach to the body correspondence problem in robot learning imitation. International Workshop on Vision-Based Human-Robot Interaction (HRI 2006), inside EUROS 2006, 2006.

(Calinon y Billard, 2004) S. Calinon y A. Billard, Stochastic gesture production and recognition model for a humanoid robot. In Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems, 2004.

(Bandera et al, 2006) J.P. Bandera, R. Marfil, L. Molina-Tanco, J.A. Rodríguez, A. Bandera y F. Sandoval, Robot learning of upper-body human motion by active imitation. In Proc. 6th IEEE-RAS Int. Conf. on Humanoid Robots, 2006.

(Bandera et al, 2007) J.P. Bandera, R. Marfil, L. Molina-Tanco, J.A. Rodríguez, A. Bandera y F. Sandoval, Robot learning by active imitation. In M. Hackel, ed., Humanoid robots: human-like machine, chapter 12, ARS, 2007.

(Asfour et al, 2006) T. Asfour, F. Gyarfas, P. Azad y R. Dillmann, Imitation learning of dual-arm manipulation tasks in humanoid robots. In 6th IEEE-RAS Int. Conf. on Humanoid Robots, 2006.

(McNamara, 1986) T.P. McNamara, Mental representations of spatial relations. Cognitive Psychology 18, 87–121, 1986.

(Zivkovic et al, 2007) Z. Zivkovic, O. Booij y B. Kröse, From images to rooms. Robotic and Autonomous Systems 55 (5), 411-418, 2007

(Cottret y Devy, 2006) M.Cottret y M.Devy, Active learning of local structures from attentive and multi-resolution vision. In 9th Int.Symp. on Intelligent Autonomous Systems (IAS'2006), 2006.

(Vázquez-Martín et al, 2006) R. Vázquez-Martín, J. Martínez, J. C. del Toro, P. Núñez y F. Sandoval, A Software Control Architecture based on Active Perception for Mobile Robotics. WSEAS Trans. on Circuits and Systems 5(6), 797-804, 2006.

(Vázquez-Martín et al, 2006b) R. Vázquez-Martín, P. Núnez, J.C. del Toro, A. Bandera y F. Sandoval, Adaptive observation covariance for EKF-SLAM in indoor environments using laser data. In 13th IEEE Mediterranean Electrotechnical Conference (MELECON 2006), 2006.

(Núñez et al, 2008) P. Núnez, R. Vázquez-Martín, J.C. del Toro, A. Bandera y F. Sandoval, Natural landmark extraction for mobile robot navigation based on an adaptive curvature estimation. accepted to Robotics and Autonomous Systems, 2008.

(Topp et al, 2006) E.A.Topp, H.Huettenrauch, H.I.Christensen y K.Severinson-Eklundh, Acquiring a shared environment representation. In Proc. of the 2006 Human Robot Interaction (HRI2006), 2006.

(Núñez et al, 2006) P. Núñez, J. P. Bandera, J. M. Pérez-Lorenzo y Francisco Sandoval, A human-robot interaction system for navigation supervision based on augmented reality. In 13th IEEE Mediterranean Electrotechnical Conference (MELECON 2006), 2006.