Revista Iberoamericana de Automática e Informática Industrial

ISSN: 1697-7912. Vol. 7, Núm. 2, Abril 2010, pp. 68-80 http://www.revista-riai.org

Análisis de Detectores y Descriptores de Características Visuales en SLAM en Entornos Interiores y Exteriores

M. Ballesta, A. Gil, O. Reinoso, D. Úbeda

Departamento Ingenieria de Sistemas Industriales, Universidad Miguel Hernández, Avda. de la Universidad, s/n, 03202, Elche, Alicante, España
(e-mail: m.ballesta|arturo.gil|o.reinoso@umh.es|ubeda@umh.es)

Resumen: El objetivo de este artículo es encontrar un extractor de características visuales que pueda ser utilizado en un proceso de SLAM (*Simultaneous Localization and Mapping*). Este extractor de características consiste en la combinación de un detector que extrae puntos significativos del entorno, y un descriptor local que caracteriza dichos puntos. Este artículo presenta la comparación de un conjunto de detectores de puntos de interés y de descriptores locales que se utilizan como marcas visuales en un proceso de SLAM. El análisis comparativo se divide en dos fases diferenciadas: detección y descripción. Se evalúa la repetibilidad de los detectores, así como la invariabilidad de los descriptores ante cambios de vista, escala e iluminación. Los experimentos se han realizado a partir de un conjunto de secuencias de imágenes tanto interiores (entorno de oficinas) como exteriores, con diversas variaciones en la imagen (iluminación y posición), representando así de una forma bastante general los entornos típicos de un robot. Se considera que los resultados de este trabajo pueden ser útiles a la hora de seleccionar una marca adecuada en SLAM visual, tanto para entornos interiores como exteriores.

Palabras Clave: SLAM visual, marcas visuales, detectores de puntos de interés, descriptores locales.

1. INTRODUCCIÓN

La capacidad de construir un mapa del entorno y localizarse en él es esencial para que un robot sea considerado como autónomo. Por este motivo, el problema de SLAM (*Simultaneous Localization and Mapping*) ha sido estudiado con gran interés en los últimos años y llevado a la práctica en diversas aplicaciones tales como tareas de rescate (Ribas *et al.* (2008)), vehículos aéreos (Béjar and Ollero (2008)), asistencia a pesonas con minusvalías (Rodriguez-Losada *et al.* (2005)) o robots guía en museos (Burgard *et al.* (1998)). El problema de SLAM trata la situación en la que un robot construye un mapa mediante la información que obtiene del entorno a través de sus sensores, y, de forma simultánea, se localiza en dicho mapa. Se trata de un problema complejo, puesto que el ruido en la estimación de la pose del robot induce ruido en la estimación del mapa y viceversa. Muchos autores realizan tareas de SLAM utilizando sensores LASER o SONAR en dos y tres dimensiones (Grisetti *et al.* (2007); Hähnel *et al.* (2003); Biber *et al.* (2004); Eustice *et al.* (2005); Triebel and Burgard (2005)). Sin embargo, existe un creciente interés en el uso de cámaras como sensores. Esta propuesta se conoce como SLAM visual. Las principales ventajas del uso de cámaras es que estos dispositivos proporcionan mayor cantidad de información del entorno y son más económicos que los LASER. Además, es posible obtener información $3D$ si se utiliza visión estereoscópica. En la configuración más típica, la cámara está instalada a una altura y orientación fijas sobre el robot (Gil *et al.* (2006); Valls Miro *et al.* (2006)). En este caso, el movimiento de la cámara queda restringido a un plano paralelo al suelo.

La mayoría de soluciones de SLAM visual están basadas en características. Estas características visuales permiten tener conocimiento del entorno e interactuar en él (Soria *et al.* (2008)). En este artículo, las marcas (*landmarks*) que constituyen el mapa, son puntos distintivos del entorno. En la extracción de dichas marcas visuales intervienen principalmente dos pasos: detección y descripción. El primer caso consiste en la detección de puntos del entorno que puedan ser utilizados como marcas. Para ello, un requerimiento importante que deben cumplir estos puntos en SLAM visual es que se deben detectar desde distintas posiciones. En la práctica, cuando el robot navega, adquiere de forma sucesiva imágenes del entorno. En esta secuencia de imágenes, diversos puntos se observan con variaciones de punto de vista y/o ángulo, e incluso iluminación, debido al movimiento del robot. En el segundo paso, se realiza la descripción de estos puntos mediante un vector de características. El uso de un descriptor adecuado es decisivo para el problema de asociación de datos, en el que el robot tiene que decidir si la observación actual corresponde a una marca integrada previamente en el mapa o a una nueva. Cada vez que el robot observa una marca, calcula la distancia a dicha marca y le asocia un descriptor visual. Durante la navegación por el entorno, el robot suele revisitar una zona previamente explorada, en este caso, el robot se vuelve a encontrar con marcas previamente observadas. Si la correspondencia entre las observaciones actuales y las marcas del mapa se realiza correctamente, el robot puede reducir la incertidumbre en su pose. En caso contrario, se puede llegar a una situación en la que el mapa creado sea inconsistente. Por tanto, la correcta asociación de datos es una parte fundamental en el proceso de SLAM.

Publicado electrónicamente: 09/04/2010
DOI: 10.4995/RIAI.2010.02.07

Hasta ahora, se han utilizado diferentes combinaciones de detectores y descriptores en procesos de SLAM con visión monocular o estereoscópica. Como ejemplo, las características SIFT (Scale-Invariant Feature Transform) se han utilizado como marcas en el espacio $3D$ en (Se *et al.* (2001)). En (Little *et al.* (2002); Gil *et al.* (2006)) se realizó además un *tracking* de estas características para obtener las más robustas. En (Jensfelt *et al.* (2006)) se utilizó una versión invariante a rotación de SIFT en combinación con el detector Harris-Laplace en SLAM monocular. El detector de esquinas de Harris se utilizó en (Davison and Murray (2002); Hygounenc *et al.* (2004)). Otro ejemplo, son las características SURF (Speeded Up Robust Features) utilizadas en (Murillo *et al.* (2007)) en tareas de localización utilizando cámaras omnidireccionales. Existe, por tanto, una gran variedad de detectores y descriptores que se han utilizado en el contexto de SLAM visual. Sin embargo, en nuestra opinion, todavía no se ha llevado a cabo un estudio que determine el detector y el descriptor más adecuados para SLAM visual.

Se han presentado con anterioridad diversos estudios evaluando detectores y descriptores orientados al reconocimiento de objetos. Este es el caso de (Schmid *et al.* (2000)), donde se evalúa un conjunto de detectores. Las características detectadas se evalúan según su validez en la correspondencia de imágenes, reconocimiento de objetos y reconstrucción $3D$. Sin embargo, en este trabajo no se evalúan los detectores más comunes en SLAM visual.

Otras evaluaciones de detectores se pueden encontrar en (Mikolajczyk *et al.* (2005)) donde se estudian detectores de regiones afines y en (Fraundorfer and Bischof (2005)), similar al anterior pero incorporando un método de seguimiento para superficies no planares. Por otro lado, (Mikolajczyk and Schmid (2005)) se centran en la comparación de descriptores. En este caso el criterio de evaluación se basa en el número de correspondencias correctas e incorrectas entre pares de imágenes.

A diferencia de estos trabajos, en este artículo se realiza un estudio novedoso en el que se evalúa la estabilidad e invariabilidad de los puntos de interés detectados y la capacidad que tienen los descriptores de caracterizar dichos puntos a lo largo de secuencias con diferentes variaciones en la imagen (cambios de punto de vista o de ángulo e iluminación). Esta situación recrea las diferentes posiciones desde las que el robot observa un punto cuando realiza tareas de SLAM. En nuestro caso, una marca visual está representada por un conjunto, (*cluster*), de descriptores extraídos desde diferentes posiciones. De modo que, en lugar de tener pares de puntos correspondientes entre las imágenes como en Mikolajczyk and Schmid (2005), tenemos conjuntos de puntos. Un conjunto se compone de un punto que ha sido observado desde diferentes posiciones en una secuencia de imágenes y tiene asociado un descriptor en cada imagen. El objetivo de este artículo es determinar qué detector es capaz de extraer puntos de interés en secuencias con variaciones de ángulo, punto de vista e iluminación. En segundo lugar, se pretende encontrar un descriptor que sea invariante a estos cambios en el contexto de SLAM visual. Los primeros trabajos en este sentido aparecen en (Ballesta *et al.* (2007), Martínez Mozos *et al.* (2007), Gil *et al.* (2009)) . En este caso, además, consideramos la invarianza a iluminación del extractor de características y su aplicación en entornos exteriores.

2. EVALUACIÓN DE DETECTORES DE PUNTOS DE INTERÉS

En esta sección se detalla la metodología utilizada para llevar a cabo la evaluación de un conjunto de detectores de puntos de interés.

Los experimentos se han realizado con un conjunto de secuencias de imágenes capturadas para este fin. Se trata de secuencias con cambios de punto de vista, escala e iluminación. Estas últimas se han obtenido aprovechando cambios de iluminación naturales. La totalidad de secuencias se ha obtenido capturando entornos interiores, con escenarios en 2 dimensiones, tales como pósters y en 3 dimensiones, en las que existen objetos en diferentes planos. Además, se han capturado también secuencias de imágenes exteriores. En la figura 1 se observan algunas de las secuencias más representativas de estos experimentos.

El objetivo que se persigue en este apartado es seleccionar el detector que encuentre los mismos puntos en el espacio a pesar de que la escena sea observada desde diferentes posiciones y ángulos y que experimente cambios de iluminación. Para ello realizamos un seguimiento (*tracking*) de los puntos a lo largo de las secuencias de imágenes. El método de seguimiento utilizado se muestra en los apartados siguientes. No obstante, la valoración de este método de seguimiento de puntos no es uno de los objetivos primordiales de este trabajo.

Tras realizar el seguimiento de puntos, los seguidos a lo largo de las secuencias son caracterizados por medio de descriptores locales. De esta forma, los descriptores asociados a un mismo punto que se ha seguido a lo largo de una secuencia, forman un conjunto en el subespacio de descriptores. Esto se verá en el apartado 3.

En lo sucesivo se presenta el conjunto de métodos de detección evaluados en este trabajo, el método de seguimiento de puntos y el criterio de evaluación establecido para obtener el detector que mejor funcione en SLAM visual.

2.1 Métodos de detección

A continuación se presenta el conjunto de detectores que se evalúan en este trabajo. Muchos de ellos han sido aplicados anteriormente a tareas de SLAM.

Harris Corner Detector[1]*:* El detector de esquinas de Harris (Harris and Stephens (1998)) es uno de los detectores de puntos de interés más comunes. El punto característico viene dado por aquel píxel que tenga los valores propios elevados en el cálculo de la matriz de momentos de segundo orden. En (Davison and Murray (2002)), los puntos de Harris se utilizan para extraer marcas visuales en SLAM monocular.

Harris-Laplace[2]*:* Los puntos extraídos con Harris-Laplace son detectados mediante una función de Harris adaptada a escala. Se seleccionan en el espacio de escalas por un operador Laplaciano. Este detector se ha aplicado en (Mikolajczyk and Schmid (2001); Jensfelt *et al.* (2006)).

SUSAN [1]*:* El detector SUSAN (Smallest Univalue Segment Assimilating Nucleus) funciona aplicando una máscara circular sobre el píxel . Un píxel es punto característico en función del número de píxeles similares al píxel central situados dentro

[1] Downloaded from http://www.mathworks.com/matlabcentral/fileexchange

[2] Downloaded from http://lear.inrialpes.fr/people/dorko/downloads.html

(a)

(b)

(c)

(d)

(e)

Figura 1. Ejemplo con algunas de las imágenes que componen cada secuencia. Se muestran los tipos de secuencias representativos. (a) Imágenes exteriores (3D) con cambios de escala. (b) Imágenes exteriores (3D) con cambios de punto de vista. (c) Imágenes interiores (2D) con cambios de punto de vista. (d) Imágenes interiores (3D) con cambios de escala. (e) Imagenes con cambios de iluminación.

de la máscara. SUSAN ha sido utilizado tradicionalmente en aplicaciones de reconocimiento de objetos (Smith (1992)).

SIFT [2]*:* El algoritmo SIFT (Scale-Invariant Feature Transform) es un detector y descriptor de puntos característicos. En este apartado nos centramos en SIFT como detector. Los puntos característicos son detectados en las imágenes mediante la función DoG (Difference of Gaussian) aplicada en el espacio de escalas (Lowe (2004)). Los puntos se seleccionan como los extremos locales de la función DoG. El algoritmo fue presentado inicialmente en (Lowe (1999)) y utilizado en tareas de reconocimiento de objetos. Recientemente, SIFT se ha utilizado en aplicaciones de SLAM visual (Gil *et al.* (2006),Sim *et al.* (2005),Valls Miro *et al.* (2006)). En este articulo se separa la etapa de detección de la de descripción, de modo que cuando se evalúa como detector, los puntos se extraen utilizando la función DoG.

SURF [3] *:* Presentadas en (Bay *et al.* (2006)), las características *SURF* (Speeded Up Robust Features) superan, según sus autores, a otros métodos existentes en términos de robustez,

[3] Downloaded from http://www.vision.ee.ethz.ch/∼surf/

repetitibilidad y distinción de los descriptores. El método de detección está basado en la matriz Hessiana y depende de imágenes integrales para reducir el coste computacional. Al igual que con las características SIFT, nos centramos sólo en los puntos detectados cuando se evalúa como detector.

2.2 Seguimiento de puntos detectados

El seguimiento de los puntos detectados se ha realizado estableciendo únicamente restricciones geométricas. Ésto nos permite evaluar los detectores independientemente del método de descripción utilizado.

Para cada imagen en una secuencia, primero se extraen los puntos de interés con los métodos de detección presentados en el apartado 2.1. A continuación, se han implementado dos algoritmos diferentes para realizar el *tracking* de los puntos en imágenes sucesivas, según se trate de secuencias 2D ó 3D. Se considera que una escena es bidimensional cuando se constituye mayoritariamente por objetos planares (ej. un póster) visto desde diferentes distancias y ángulos. Una escena 3D contiene objetos en diferentes planos vistos también desde diferentes posiciones. Las secuencias con cambios de iluminación son estáticas en cuanto a la posición del punto de observación, por ello no es necesario aplicar estos métodos de seguimiento.

En el caso 2D se ha utilizado un método basado en la matriz de homografía como en (Schmid *et al.* (2000)). En este caso, dado un punto X en el espacio 3D, se asume que este punto proyecta en la posición $x_1 = K_1 X$ en la imagen I_1 y en la posición $x_i = K_i X$ en la imagen I_i, donde K_1 y K_i son matrices de proyección (Fig. 2). Si suponemos que el punto X se detecta en ambas imágenes, entonces

$$x_i = H_{1i} x_1, \; con \; H_{1i} = K_i K_1^{-1} \tag{1}$$

La matriz de homografía H_{1i} se calcula seleccionando, de forma manual, cuatro correspondencias de puntos coplanares entre las imágenes 1 y i. La matriz de homografía permite encontrar las correspondencias entre los puntos de interés para cada par de imágenes consecutivas. Dado un punto detectado en una imagen, se puede predecir su posición en la imagen consecutiva mediante la matriz de homografía. Si la posición predicha está situada a una distancia menor de ε píxeles respecto de un punto de interés detectado en la segunda imagen, enconces se considera que ambos puntos son correspondientes y, por tanto, el *tracking* del punto de interés se ha realizado satisfactoriamente. Por el contrario, si no hay ningún punto de interés detectado en el entorno de la posición predicha, entonces se considera que el punto se ha perdido. En este trabajo, se ha seguido el criterio de que si un punto se pierde durante el seguimiento, ya no se vuelve a buscar en posteriores imágenes de la secuencia. El método de la matriz de homografía se ha aplicado a secuencias de imágenes que contienen objetos coplanares (pósters).

En el caso de las imágenes con escenarios 3D, se ha implementado un método de seguimiento basado en la matriz fundamental. La matriz fundamental, de dimensión 3×3 y rango 2, relaciona los puntos correspondientes de dos imágenes estéreo. Dado un punto x en la imagen I_i, la matriz fundamental F calcula la línea epipolar sobre la que debe localizarse el punto correspondiente x' en la segunda imagen I_{i+1}. Esta línea epipolar se calcula como $E_p = F x_1$ (véase Figura 3). Como consecuencia, dos puntos correspondientes cumplirán la siguiente ecuación:

$$x_i'^T F x_i = 0 \tag{2}$$

Para cada punto x_i el punto correspondiente en la siguiente imagen x_i' se selecciona como el que tenga la menor distancia a la línea epipolar. Esta estrategia, en principio, puede dar lugar a numerosas falsas correspondencias, ya que puede haber muchos puntos cerca de la línea. Por ello, se establece una restricción, por la cual el punto x_i' debe caer dentro de una ventana de 10×10 píxeles centrada en el punto x_i. Esto es válido para las secuencias de los experimentos, ya que el movimiento de la cámara entre imágenes consecutivas es pequeño.

El cálculo de la matriz F se divide en dos pasos. En el primero, se seleccionan siete correspondencias entre cada par de imágenes consecutivas. Esto nos permite calcular una matriz fundamental F. En el segundo paso, a partir de esta matriz F se obtiene un conjunto de correspondencias entre ambas imágenes que se utiliza, a su vez, para el cálculo de una segunda matriz fundamental F'. En este segundo paso, la matriz fundamental se calcula utilizando RANSAC (RANdom Sample Consensus) como en (Zhang *et al.* (1995)). De este modo, el resultado es una matriz fundamental F' mejor estimada, que permite obtener el conjunto final de correspondencias con mayor precisión. La figura 4 muestra ejemplos de puntos seguidos a lo largo de diferentes secuencias. Concretamente, muestra puntos de interes obtenidos con Harris (cruces blancas). Primero, estos puntos se extraen en cada imagen de la secuencia. A continuación, se realiza el *tracking* a lo largo de dicha secuencia. En la figura 4 se muestran los puntos que se siguieron a lo largo de la secuencia completa. Los puntos que se pierden, incluso en una sola imagen, son rechazados por nuestro algoritmo de *tracking* ya que se considera que estos puntos no son suficientemente estables para nuestro propósito.

2.3 Criterio de evaluación

La evaluación de los detectores de puntos de interés se realiza estudiando la repetibilidad de dichos puntos ante cambios de punto de vista, escala e iluminación. Una vez realizado el seguimiento de los puntos en una secuencia, se puede definir una ratio de supervivencia rs_i en la imagen i de la secuencia como:

$$rs_i = \frac{np_i}{np_1} \cdot 100 \tag{3}$$

donde np_i es el número de puntos de interés seguidos hasta la imagen i en la secuencia y np_1 es el número de puntos de interés detectados en la primera imagen de la secuencia. El detector ideal sería el que detectase los mismos puntos a lo largo de la secuencia, es decir, $rs_i = 100\,\%$ para cada imagen de la secuencia. Sin embargo, como se comprobará en los resultados, se observa una tendencia decreciente de rs_i, lo cual significa que algunos de los puntos observados en la primera imagen se pierden en las sucesivas imágenes.

3. EVALUACIÓN DE DESCRIPTORES LOCALES

La evaluación de los descriptores locales presentados en el apartado 3.1 se realiza en tres pasos. Primero, se elige uno de los detectores del apartado 2.1 y lo aplicamos a cada imagen de las secuencias. En segundo lugar, se utiliza el método de seguimiento explicado en la sección 2.2 con el fin de obtener un conjunto de puntos que se hayan seguido a lo largo de

Figura 2. El punto X se proyecta sobre las imágenes de la secuencia en los puntos $x_1, x_2, ..., x_i$. Las matrices de homografia H_{12} y H_{2i} establece las correspondencias entre las imágenes $I_1 - I_2$ e $I_2 - I_i$ respectivamente.

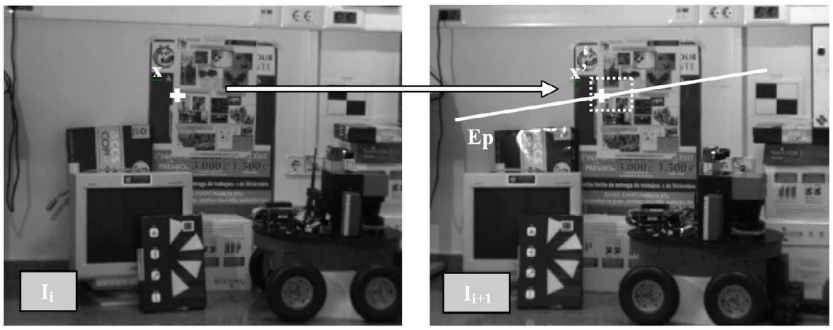


Figura 3. El punto p' es la correspondencia de p en la imagen I_{i+1}. Este punto está situado en la linea epipolar E_p calculada con la matriz fundamental F.

la secuencia completa. Finalmente, se extrae cada uno de los descriptores locales del apartado 3.1 en un entorno local de cada punto de interés seguido en toda la secuencia.

Como resultado de estos pasos, un punto de interés x, que se haya seguido en una secuencia completa $\{I_1, ..., I_N\}$, será representado por M conjuntos diferentes $D_1^x, ..., D_M^x$, donde el conjunto $D_M^x = \{d_m^x(1), ..., d_m^x(N)\}$ representa el punto a lo largo de la trayectoria utilizando el método de descripción m, y cada elemento $d_m^x(i)$ indica el descriptor representando al punto x en la imagen I_i de la secuencia. En este caso $m \in \{$Ventana de niveles de gris (*Patch*),SIFT,SURF,E-SURF, U-SURF, Zernike, Histograma$\}$.

En la figura 5(a) se observan dos puntos $\{x_1, x_2\}$ que se siguen a lo largo de una secuencia compuesta de tres imágenes $\{I_i, I_{i+1}, I_{i+2}\}$. En cada imagen, cada punto de interés se describe mediante ocho descriptores $\{d_1, ..., d_8\} = \{$Ventana de niveles de gris (*Patch*), SIFT, SURF, E-SURF, U-SURF, Zernike, Histograma$\}$. A partir del seguimiento d estos puntos, obtenemos dos conjuntos de vectores del descriptor d_1 que representa los puntos seguidos en esas imágenes:

$$D_1^{x_1} = \{d_1^{x_1}(i),\ d_1^{x_1}(i+1),\ d_1^{x_1}(i+2)\},$$
$$D_1^{x_2} = \{d_1^{x_2}(i),\ d_1^{x_2}(i+1),\ d_1^{x_1}(i+2)\},$$
(4)

Del mismo modo, podemos obtener dos conjuntos para el segundo descriptor d_2. En el caso general, tendremos V conjuntos de vectores para cada descriptor m, donde V es el número de puntos seguidos a lo largo de la secuencia completa. En este trabajo, se considera que cada punto seguido completamente es una marca visual, por tanto, se tendrán V marcas visuales.

Supongamos que nos centramos únicamente en uno de los descriptores, por ejemplo en d_1. Cada uno de los V conjuntos $D_1^{x_1}, ..., D_1^{x_V}$, pertenecientes al descriptor seleccionado, forman un conjunto en el subespacio de descriptores. Cada conjunto representa el punto x a lo largo de la secuencia de imágenes completa. En la figura 5(b) se muestra un ejemplo. En este caso, los dos puntos $\{x_1, x_2\}$ de la figura 5(a) se han seguido a lo largo de las tres imágenes en la secuencia $\{I_i, I_{i+1}, I_{i+2}\}$.Para una mayor claridad en la representación gráfica, se asume que el descriptor tiene dos componentes $d_1 = \{a, b\}$. La figura 5(b)

Figura 4. Las imágenes de la parte superior representan una escena típica del laboratorio con cambios de punto de vista. Las imágenes de la parte inferior corresponden a una escena exterior con cambios en escala. Estas imágenes son la primera, central y última de una secuencia de 21 imágenes (laboratorio) o 11 imágenes (exterior). Los puntos detectados están indicados con cruces blancas y con cruces rojas se resaltan aquellos que se siguen a lo largo de la secuencia.

muestra el caso ideal de comportamiento que se desea para un descriptor. Como se aprecia en la figura, los descriptores que representan un mismo punto están muy agrupados formando un conjunto y, a su vez, se diferencian claramente del conjunto formado por los descriptores de un punto diferente. Este descriptor sería bastante distintivo, y, en consecuencia, idóneo para integrarlo en un proceso de SLAM visual.

3.1 Descriptores Locales

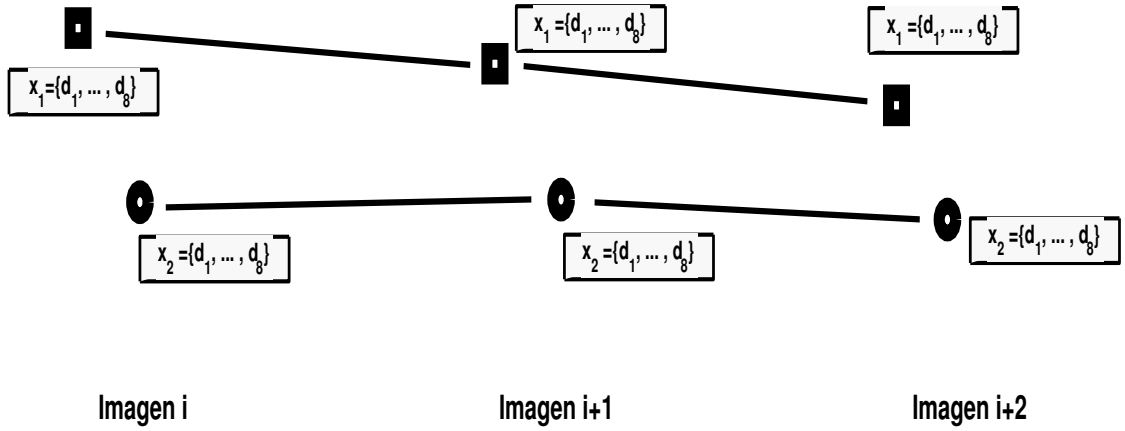
En este artículo se evalúan descriptores que han sido aplicados previamente en SLAM visual. A continuación, se describen brevemente los descriptores evaluados.

SIFT: En el algoritmo SIFT se asigna una orientación global a cada punto basada en las direcciones del gradiente en un entorno local. A continuación, se calcula un descriptor basado en los histogramas de orientación en una subregión 4×4 entorno al punto de interés, resultando en un vector de dimensión 128 Lowe (2004). Para obtener invarianza a iluminación, el descriptor se normaliza con la raíz cuadrada de la suma de los componentes al cuadrado.

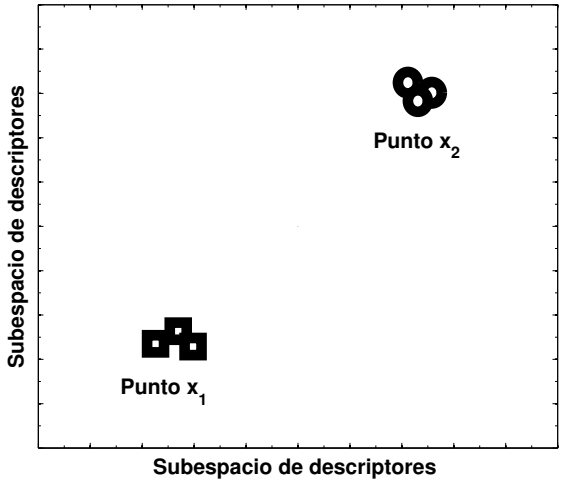
SURF: El descriptor SURF representa una distribución de respuestas Haar-wavelet en el entorno del punto de interés y utiliza imágenes integrales eficientemente. En este artículo se han estudiado las tres versiones de este descriptor: El descriptor SURF estándar, que tiene una dimensión de 64, la versión extendida (E-SURF) con 128 elementos y la versión *upright* (U-SURF). La versión U-SURF no es invariante a rotación y el vector tiene 64 elementos (Bay *et al.* (2006)).

Ventana de niveles de gris: Este método describe cada marca utilizando los valores de nivel de gris de una subregión entorno al punto de interés. Este descriptor ha sido utilizado en un contexto de SLAM monocular en (Davison and Murray (2002)).

Histograma de orientación: Los histogramas de orientación se calculan a partir de la imagen gradiente. Para cada píxel, se calcula un módulo y una orientación. Los valores de orientación se dividen en intervalos y el histograma se forma con los valores



(a)



(b)

Figura 5. (a) Dos puntos de interés se siguen a lo largo de tres imágenes consecutivas. En cada imágen, cada punto de interés x_1 y x_2 está representado por los ocho descriptores estudiados en este artículo: $\{d_1, ..., d_8\}$ = {Ventana de niveles de gris (*Patch*),SIFT,SURF,E-SURF, U-SURF, Zernike, Histograma}. (b) Ejemplo de representación en el subespacio de descriptores de los puntos de la figura 5(a) considerando un descriptor, ej., d_1.

del módulo. En (Kosecka *et al.* (2003)) los histogramas de orientación se aplican en navegación de robots móviles.

Momentos de Zernike: La formulación de los polinomios de Zernike (Zernike (1934)) parece ser una de las más populares en términos de redundancia y capacidad de reconstrucción. Los momentos complejos de Zernike se construyen mediante un conjunto de polinomios que forma un conjunto base ortogonal definido en el círculo unidad.

3.2 Criterio de evaluación

El comportamiento de los descriptores se ha evaluado en un contexto de asociación de características mediante las curvas de *recall* y *precision*. De este modo, se tienen en cuenta los falsos positivos. No obstante, a diferencia de (Mikolajczyk and Schmid (2005)) la evaluación se realiza considerando que cada marca está representada por un conjunto. Este conjunto está formado por los descriptores a las diferentes vistas de un punto en la escena. Esta situación se da con frecuencia en SLAM visual, ya que en muchas soluciones se realiza el seguimiento de puntos en la escena antes de ser integrados como marcas en el mapa (Gil *et al.* (2006); Se *et al.* (2001); Valls Miro *et al.* (2006)).

Al comenzar este apartado 3, se explicó que un punto de interés x, que se sigue completamente a lo largo de una secuencia, se representa por medio de M conjuntos diferentes $D_1^x, ..., D_M^x$. El conjunto $D_m^x = \{d_m^x(1), ..., d_m^x(N)\}$ se considera un conjunto y representa al punto, caracterizado por el descriptor m, cuando es visto desde diferentes posiciones. En este caso consideramos que existe un total de V conjuntos, que corresponden al número total de puntos que se han seguido en las secuencias completas. Dado un descriptor que representa una vista específica de una marca visual, nuestro propósito es asociar ese descriptor a su conjunto correspondiente utilizando una

medida de distancia. Para ello, se ha hecho uso de la distancia de Mahalanobis, que se define como:

$$\sqrt{(d_m^{x_i} - \mu_j)^T S_j^{-1} (d_m^{x_i} - \mu_j)} \quad (5)$$

donde $d_m^{x_i}$ es un descriptor que pertenece a la clase i,S_j es la matriz de covarianza asociada al conjunto w_j y μ_j es el vector media del conjunto.

En los experimentos, se ha dividido de forma aleatoria cada conjunto w_j en dos grupos de tamaño similar: uno de ellos se utiliza para calcular la matriz S_j asociada al conjunto, mientras que los A descriptores restantes se utilizan para evaluar el proceso de clasificación. En consecuencia, el número de posibles correspondencias será $A \cdot V$, donde V es el número de conjuntos y A el número de ejemplos que queremos clasificar. Por cada descriptor $d_m^{x_i}$ que queremos clasificar, se calcula la distancia de Mahalanobis a todos los conjuntos de la base de datos y se busca el conjunto que minimiza esta distancia. Además, el número de correspondencias correctas es conocido: una correspondencia es verdadera cuando el descriptor $d_m^{x_i}$ se asigna al conjunto w_i y falsa cuando el descriptor se asigna a un conjunto diferente. Como resultado, tenemos una lista de correspondencias realizadas, y a cada una se le asocia con una distancia de Mahalanobis al conjunto mínima y la indicación $verdadera/falsa$, según se haya hecho correctamente la asociación o no. A continuación, se ordena la lista de correspondencias en orden ascendente atendiendo a la distancia mínima de Mahalanobis. Esta lista se utliza para calcular los parámetros *recall* y $precision$, que se definen como:

$$recall = \frac{n^{\text{o}}\ correspondencias\ correctas\ seleccionadas}{n^{\text{o}}\ total\ correspondencias\ correctas}$$

$$precision = \frac{n^{\text{o}}\ correspondencias\ correctas\ seleccionadas}{n^{\text{o}}\ correspondencias\ seleccionadas}$$

En estas expresiones, $recall$ representa la habilidad de encontrar todas las correpondencias correctas, mientras que $precision$ representa la capacidad de obtener correspondencias correctas cuando el número de correspondencias seleccionadas varía. Tomando diferentes umbrales para la distancia de Mahalanobis en la lista ordenada, da lugar a diferentes conjuntos de correspondencias seleccionadas y, por tanto, a diferentes valores de $recall$ y $precision$. El factor $n^{\text{o}}\ correspondencias\ seleccionadas$ es el número de correspondencias en la lista cuya distancia es menor que un umbral establecido, y varía desde 1 hasta el número total de correspondencias que componen la lista ($A \cdot V$). La variable $correspondencias\ correctas\ seleccionadas$ es el número de correspondencias correctas obtenidas de la lista ordenada dado un umbral. El factor $n^{\text{o}}\ total\ correspondencias\ correctas$ es un valor constante, que expresa el número total de posibles correspondencias correctas. En nuestro caso, si la clasificación fuese perfecta se asignaría correctamente el total de descriptores ($A \cdot V$) a sus conjuntos correspondientes. Por tanto, el número total de posibles correspondencias correctas es $A \cdot V$. El descriptor ideal alcanzaría un valor de $recall$ igual a 1, lo que significa que todas las asociaciones realizadas son correctas.

En una curva $recall\ vs.\ precision$, un valor alto de $precision$ con un valor pequeño de $recall$ indica que se han obtenido sólo una pequeña parte del total de posibles correspondencias correctas. Por otro lado, un valor alto de $recall$ acompañado de un valor pequeño de $precision$ supone que se han encontrado bastantes correspondencias correctas, pero que también hay un número considerable de correspondencias incorrectas. Por este motivo, la situación ideal será encontrar el descriptor que obtenga valores elevados tanto de $recall$ como de $precision$, de modo que la curva $recall\ vs.\ precision$ se sitúe en la esquina superior derecha del gráfico.

4. RESULTADOS EXPERIMENTALES

Para llevar a cabo los experimentos de evaluación de los diferentes detectores y descriptores, se ha capturado un conjunto de imágenes que se describe en la Tabla 1.

Tabla1. Conjunto de secuencias de imágenes.

Tipo	Cambio en la imagen	nºsecuencias	nºimagenes en la secuencia
Interiores	Punto de vista	12	21
	Escala	14	12
	Iluminación	3	11
Exteriores	Punto de vista	5	11
	Escala	4	12
	Iluminación	10	14-17

Todas las secuencias se han capturado usando una cámara Videre Design MDCS3. En el caso de las secuencias con cambios en el punto de vista, la cámara describió una trayectoria semicircular tomando como centro un punto de la escena. La variación entre imágenes consecutivas es de 2,5 grados en imágenes interiores y de unos 20 grados en el caso de imágenes exteriores. En el caso de imágenes con cambios de escala, la cámara describió una trayectoria recta, moviéndose 0,1 metros entre imágenes consecutivas. Las secuencias de imágenes presentan escenas en 2D y 3D. Diremos que son escenas 2D, aquellas en las que los objetos se encuentran en un mismo plano (p.ej. póster). Por otro lado, las imágenes 3D, presentan objetos en diferentes planos (p.ej. imágenes del laboratorio e imágenes exteriores). Las imágenes con cambios de iluminación se han capturado aprovechando variaciones naturales de luminosidad en la escena. En la figura 1 se pueden ver ejemplos de todos los tipos de imágenes utilizadas. Finalmente, las imágenes se capturaron en diferentes resoluciones (320×240, 640×480 and 1280×960), de modo que el conjunto de imágenes sea lo más representativo posible.

4.1 Detectores de puntos

Como se ha descrito previamente, el proceso seguido para realizar la evaluación de detectores ha sido el siguiente. Primeramente, se extrajeron los puntos de interés con los detectores presentados en el apartado 2.1. A continuación, se realizó el seguimiento o *tracking* de estos puntos a lo largo de las secuencias tal y como se explicó en el apartado 2.2. Para cada secuencia se calculó la ratio de supervivencia mediante la ecuación 3. Los resultados obtenidos se muestran en las Figuras 6, 7, 8 y 9. Las figuras muestran el valor medio y desviación 2σ obtenidos para secuencias con el mismo tipo de variaciones en la imagen. En todos los casos, se observa que Harris es el detector con los mejores resultados. Por ejemplo, en la Fig. 6(b), este detector alcanza una ratio de supervivencia de $55\,\%$ al final de la secuencia. En lo que respecta al resto de detectores, se observa que Harris-Laplace y SURF tienen un comportamiento similar, con peores resultados que el detector de Harris. El detector SIFT obtiene resultados similares a Harris-Laplace y SURF en el caso de imágenes interiores (Figs. 6 y 7). En este caso, los peores resultados se obtienen con el detector SUSAN. Sin embargo, en los experimientos realizados con imágenes exteriores

y con cambios de iluminación (Figs. 8 y 9 respectivamente) se observa que SIFT muestra, junto con SUSAN, los peores resultados.

Si se compara el caso 2D y el 3D, en imágenes interiores, el detector Harris obtiene mejores resultados en el último caso, con mayor diferencia respecto del resto de detectores. Como ejemplo, en la Fig. 7(a) se observa que el detector Harris consigue que un $30\,\%$ de puntos se siga a lo largo de la secuencia completa, mientras que el segundo mejor detector (SURF) consigue entorno a un $20\,\%$. En el caso tridimensional esta diferencia es mayor, mientras que Harris alcanza una ratio de supervivencia de $50\,\%$, SURF tiene un $25\,\%$. Este resultado se podría explicar porque en las imágenes 3D la cantidad de esquinas o estructuras similares existentes es mayor, con lo que se favorece la detección de este tipo de puntos

En los experimentos con imágenes exteriores (Fig. 8), se observa, en general, un mayor descenso en el número de puntos seguidos. Este fenómeno puede ser debido a que los cambios entre imágenes consecutivas son más bruscos y a que la diversidad de características en la escena es mayor que en el caso de entornos interiores. No obstante, en estos casos Harris muestra también los mejores resultados.

A la vista de los resultados, el detector Harris, ha demostrado gran estabilidad ante cambios en el punto de vista, escala e iluminación.

4.2 Descriptores de puntos

En el siguiente experimento, se calcularon los descriptores correspondientes a los puntos seguidos en las secuencias de imágenes completas. Los descriptores se evalúan con los puntos detectados por Harris, ya que, según los resultados obtenidos, es el detector que mejor satisface el comportamiento deseado en SLAM visual. Los descriptores que se han evaluado son los descritos en el apartado 3.1.

Se ha implementado el criterio de *matching* de características explicado en el apartado 2.2. Las figuras 10 y 11 muestran los resultados obtenidos en imágenes con cambios en el punto de vista y escala respectivamente, para escenas $2D$ y $3D$ en entornos interiores. La figura 12 muestra los resultados obtenidos en entornos exteriores, con cambios de punto de vista (Fig. 12(a)) y cambios de escala (Fig.12(b)). Finalmente la figura 13 muestra una secuencia ejemplo con cambios de iluminación. Las figuras muestran las curvas de *recall* y *precision* para cada descriptor. Como se explicó en el apartado 3.2, se busca que la curva resultante se encuentre en la zona superior derecha del gráfico, indicando que se han encontrado bastantes correspondencias correctas y el número de incorrectas es muy bajo. Bajo este criterio, el descriptor U-SURF muestra los mejores resultados, puesto que obtiene los valores de *recall* y *precision* más cercanos a 1. SURF y E-SURF presentan resultados similares, que superan el resultado de SIFT.

El descriptor U-SURF destaca respecto del resto de descriptores. Es importante destacar que las secuencias utilizadas en nuestros experimentos no presentan cambios de rotación. Esto es debido a que en la mayoría de aplicaciones de SLAM visual (Gil *et al.* (2006); Jensfelt *et al.* (2006); Valls Miro *et al.* (2006)) la cámara sólo se mueve en un plano paralelo al suelo, y rota únicamente entorno a un eje vertical. Este hecho puede explicar los buenos resultados que se han obtenido con el descriptor U-SURF.

Si comparamos solamente descriptores invariantes a rotación (SURF, SIFT, E-SURF), se observa que SURF y E-SURF presentan resultados similares. Por este motivo, se podría afirmar que el coste computacional que supone calcular la versión E-SURF no merece la pena, ya que los resultados son similares. Si comparamos SURF y SIFT, SURF obtiene siempre mejores resultados que SIFT. Además, SIFT es superado en varios casos por el descriptor *Patch*.

Respecto al resto de descriptores (Histogramas y Momentos de Zernike), no presentan resultados relevantes.

5. CONCLUSIONES

En este artículo, se ha realizado una evaluación de detectores de puntos de interés y descriptores locales que caracterizan estos puntos para integrarlos como marcas visuales del entorno en un proceso de SLAM visual. Los experimentos se han realizado con una amplia batería de imágenes, que comprenden escenarios interiores, exteriores, 2D y 3D. De modo que los resultados obtenidos se puedan aplicar en multitud de situaciones.

La evaluación de los detectores y de los descriptores se ha llevado a cabo de forma independiente. En primer lugar, se realizó una evaluación de detectores estableciendo los requerimientos deseables en SLAM visual, como son la repetibilidad y la estabilidad de los puntos detectados. Para este estudio se capturaron secuencias de imágenes en las que el movimiento total de la cámara es significativo, al igual que ocurre en las aplicaciones de SLAM visual. Además, se definió una medida de evaluación denominada $ratio\ de\ supervivencia$, que mide el porcentaje de puntos detectados que persiste a lo largo de la secuencia completa. Tras este estudio, el detector que mejor se ajusta a los requerimientos establecidos es el detector de esquinas de Harris.

Una vez seleccionado el detector de puntos de interés, el siguiente paso es buscar el descriptor, que en combinación con este detector (Harris) obtenga los mejores resultados para SLAM visual. En este caso, los descriptores han sido evaluados según la capacidad de establecer correspondencias (*matching*). Para ello, se utilizaron las curvas de *recall* y *precision* que evalúan la capacidad de clasificación de los descriptores utilizando la distancia de Mahalanobis. Ambos criterios coinciden en que el descriptor U-SURF es el más adecuado en el contexto estudiado. Sin embargo, el descriptor U-SURF no es invariante a rotación. Por este motivo, estaría limitado a aplicaciones en las que la cámara solo rota en torno a un eje vertical, que es el caso que se estudia en este artículo. Si nos centramos únicamente en los descriptores invariantes a rotación (SURF,E-SURF y SURF), se observa, en los resultados, que SURF es el mejor descriptor. Éste sería el descriptor idóneo en las aplicaciones en las que la camara pueda rotar.

Es también importante destacar que el descriptor SURF obtiene mejores resultados que SIFT en la mayoria de los resultados obtenidos. Además, SURF tiene un menor coste computacional, lo cual facilita la extracción online de características visuales.

Los resultados muestran que el descriptor U-SURF en combinación con el detector Harris proporcionan marcas visuales que son estables y distintivas. Esto las hace idóneas para realizar tareas de SLAM visual.

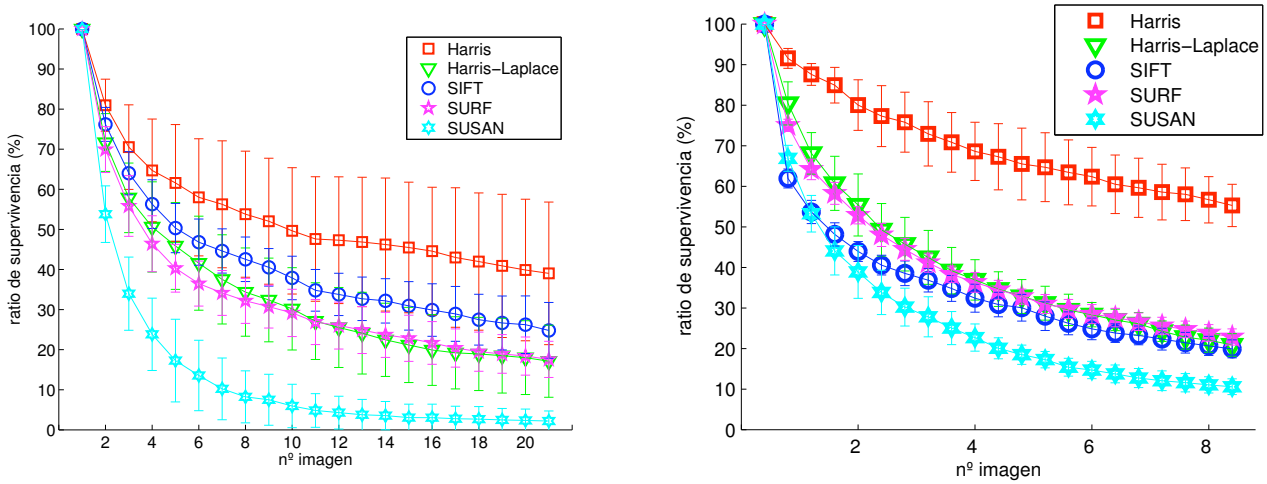


Figura 6. La imagen de la izquierda muestra la ratio de supervivencia medio de los puntos de interés en todas las secuencias 2D de imágenes interiores con cambios de punto de vista. La imagen de la derecha muestra los mismos resultados pero con imágenes 3D. En ambas figuras se muestra los límites de error 2σ calculados para secuencias con los mismos cambios en la imagen.

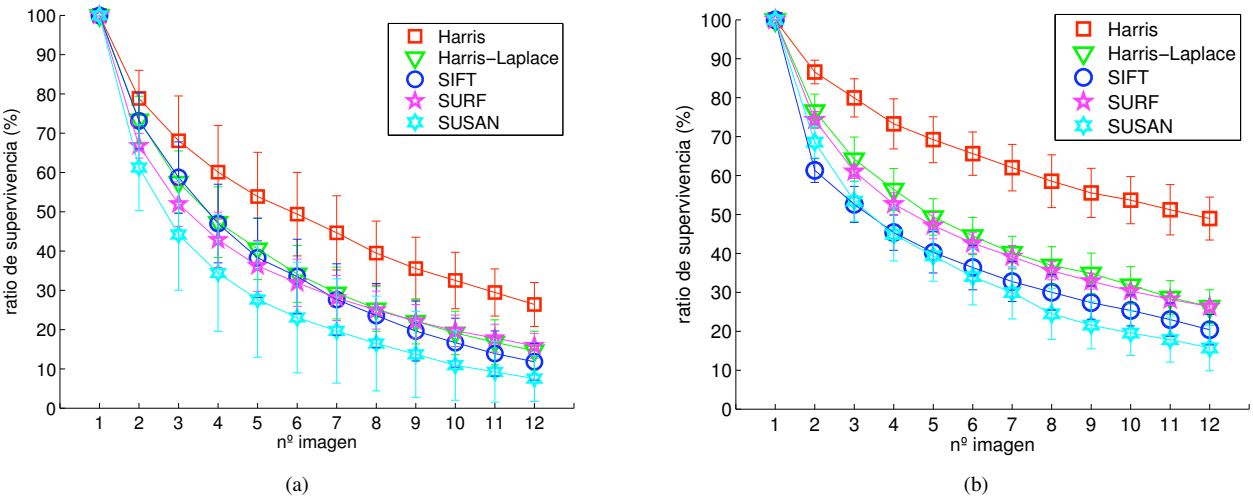


Figura 7. La imagen de la izquierda muestra la ratio de supervivencia medio de los puntos de interés en todas las secuencias 2D de imágenes interiores con cambios de escala. La imagen de la derecha muestra los mismos resultados pero con imágenes 3D. En ambas figuras se muestra los límites de error 2σ calculados para secuencias con los mismos cambios en la imagen.

AGRADECIMIENTOS

Este trabajo ha sido realizado parcialmente gracias al apoyo del Ministerio de Ciencia e Innovación con el proyecto CICYT DPI2007-61197 y de la Generalitat Valenciana con la beca BFPI/2007/096.

REFERENCIAS

Ballesta, M., A. Gil, O. Martínez Mozos and O. Reinoso (2007). Local descriptors for visual SLAM. In: *Workshop on Robotics and Mathematics (ROBOMAT07), Portugal*. pp. 209–215.

Bay, Herbert, Tinne Tuytelaars and Luc Van Gool (2006). SURF: Speeded up robust features. In: *European Conference on Computer Vision*.

Biber, P., H. Andreasson, T. Duckett and A. Schilling (2004). 3D modelling of indoor environments by a mobile robot with a laser scanner and panoramic camera. *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems* **4**, 3430– 3435.

Burgard, W., A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner and S. Thrun (1998). The interactive museum tour-guide robot. In: *Proc. of the National Conference on Artificial Intelligence*.

Béjar, M. and A. Ollero (2008). Modelado y control de helicópteros autónomos. revisión del estado de la técnica. *RIAI* **5**(4), 5–16.

Davison, Andrew J. and David W. Murray (2002). Simultaneous localisation and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 735–758.

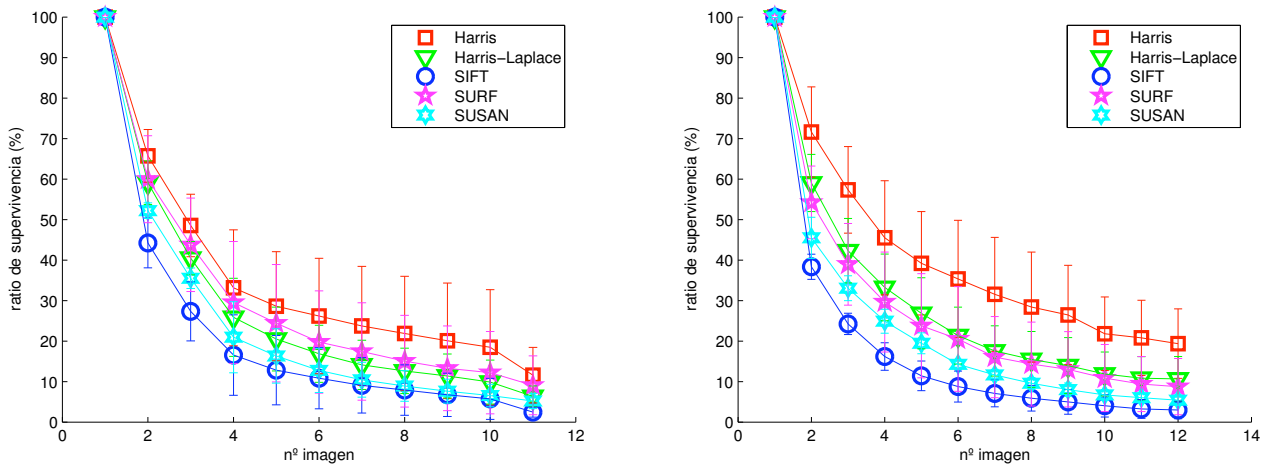


Figura 8. La imagen de la izquierda muestra la ratio de supervivencia medio de los puntos de interés en todas las secuencias de imágenes exteriores con cambios de punto de vista. La imagen de la derecha muestra la ratio de supervivencia medio de los puntos de interés en todas las secuencias de imágenes exteriores con cambios de escala. En ambas figuras se muestra los límites de error 2σ calculados para secuencias con los mismos cambios en la imagen.

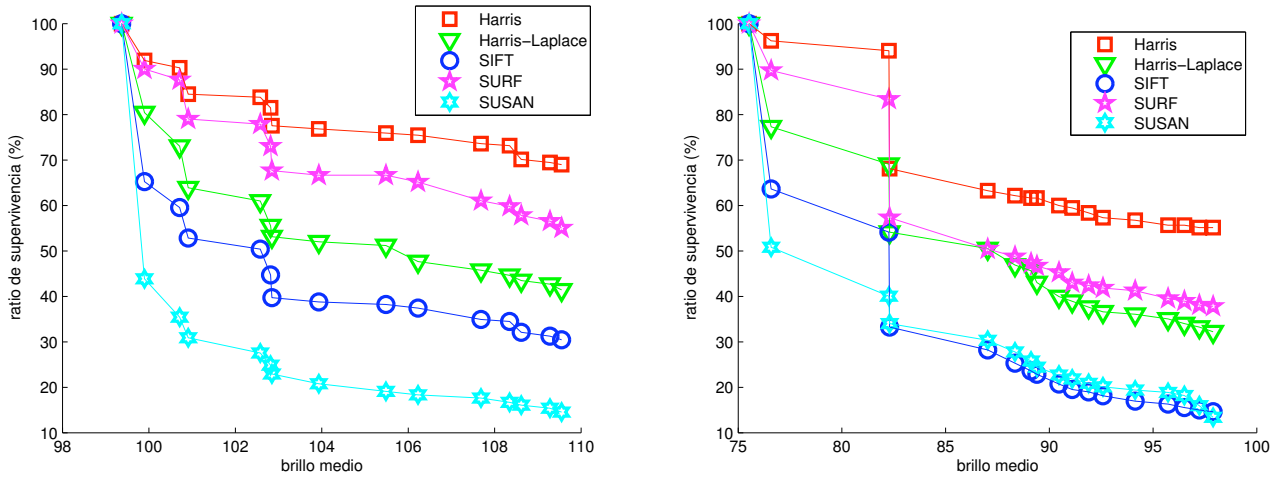


Figura 9. La imagen de la izquierda corresponde al ratio de supervivencia medio de los puntos de interés en una secuencia ejemplo con cambios de iluminación. La imagen de la derecha muestra otra secuencia ejemplo de las mismas características.

Eustice, R., H. Singh and J.J. Leonard (2005). Exactly sparse delayed-state filters. In: *IEEE Int. Conf. on Robotics & Automation*. pp. 2417– 2424.

Fraundorfer, F. and H. Bischof (2005). A novel performance evaluation method of local detectors on non-planar scenes. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

Gil, A., O. Martinez Mozos, M. Ballesta and O.Reinoso (2009). A comparative evaluation of interest point detectors and local descriptors for visual slam. In: *Machine Vision and Applications Journal*.

Gil, A., O. Reinoso, W. Burgard, C. Stachniss and O. Martínez Mozos (2006). Improving data association in rao-blackwellized visual SLAM. In: *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*. pp. 2076–2081.

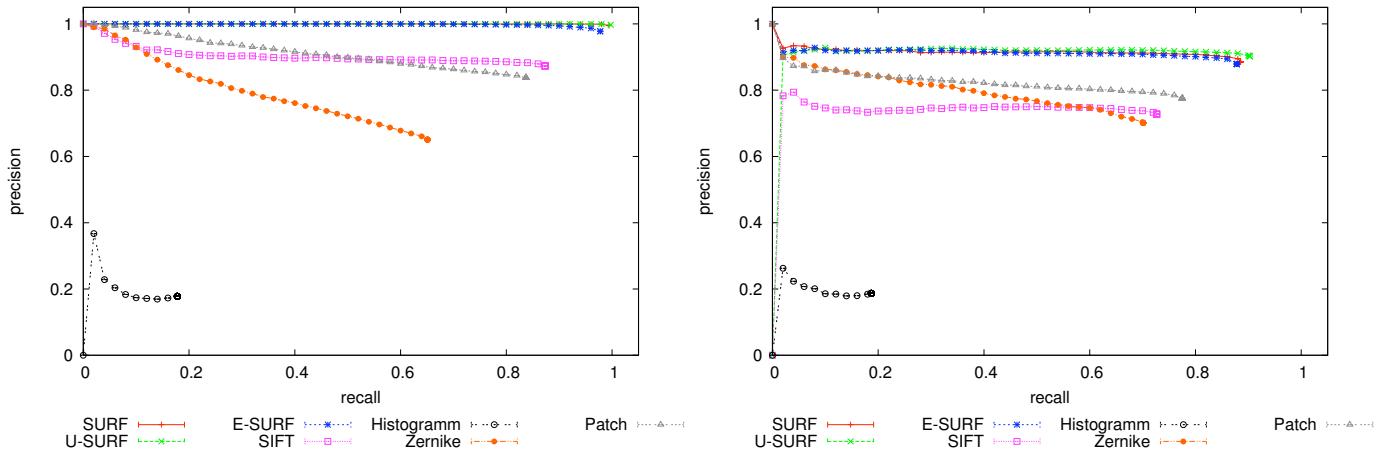
Grisetti, G., C. Stachniss and W. Burgard (2007). Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Transactions on Robotics* **23**(1), 34–46.

Hähnel, D., W. Burgard, D. Fox and S. Thrun (2003). An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements. In: *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*. Las Vegas, NV, USA. pp. 206–211.

Harris, C. G. and M. Stephens (1998). A combined corner and edge detector. In: *Alvey Vision Conference*. pp. 147–151.

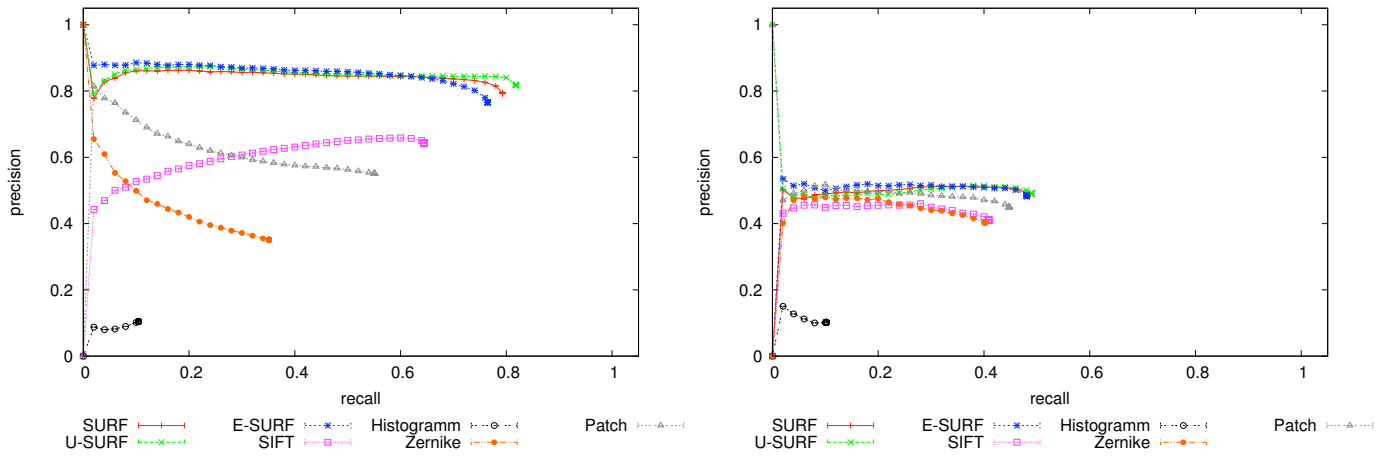
Hygounenc, Emmanuel, Il-Kyun Jung, Philippe Souères and Simon Lacroix (2004). The autonomous blimp project of laas-cnrs: Achievements in flight control and terrain mapping. *International Journal of Robotics Research* **23**(4–5), 473–511.

Jensfelt, Patric, Danica Kragic, John Folkesson and Mårten Björkman (2006). A framework for vision based bearing only 3D SLAM. In: *IEEE Int. Conf. on Robotics & Automation*. pp. 1944–1950.



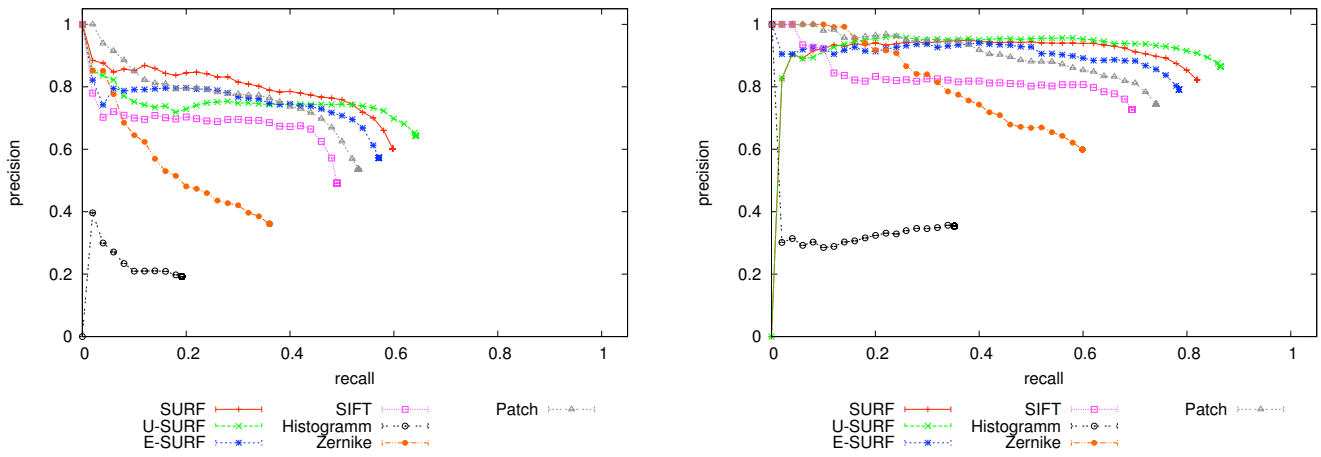
(a) (b)

Figura 10. La imagen de la izquierda muestra las curvas de *recall vs. precision* para las secuencias 2D interiores con cambios de punto de vista. La imagen de la derecha muestra los mismos resultados pero con imágenes 3D interiores.



(a) (b)

Figura 11. La imagen de la izquierda muestra las curvas de *recall vs. precision* para las secuencias 2D interiores con cambios de escala. La imagen de la derecha muestra los mismos resultados pero con imágenes 3D interiores.



(a) (b)

Figura 12. La imagen de la izquierda muestra las curvas de *recall vs. precision* para las secuencias exteriores con cambios de punto de vista. La imagen de la derecha muestra los mismos resultados pero imagenes exteriores con cambios de escala.

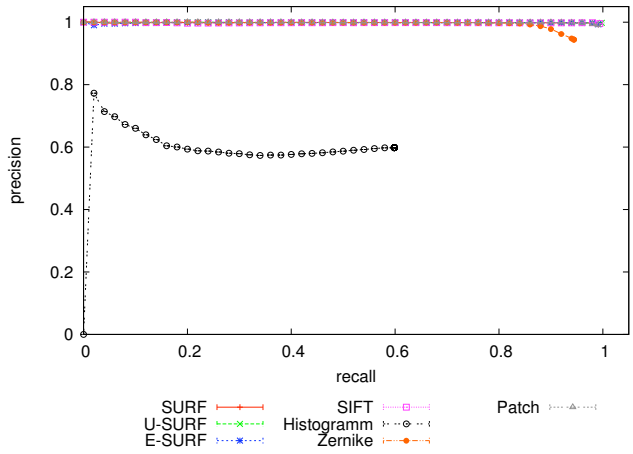


Figura 13. La imagen muestra las curvas de *recall vs. precision* para una secuencia ejemplo con cambios de iluminación.

Kosecka, J., L. Zhou, P. Barber and Z. Duric (2003). Qualitative image based localization in indoor environments. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. pp. 3–8.

Little, J., S. Se and D.G. Lowe (2002). Global localization using distinctive visual features. In: *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*. pp. 226–231.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **2**(60), 91–110.

Lowe, D.G. (1999). Object recognition from local scale-invariant features. In: *Int. Conf. on Computer Vision*. pp. 1150–1157.

Martínez Mozos, O., A. Gil, M. Ballesta and O. Reinoso (2007). Interest point detectors for visual slam. In: *Proc. of the XII Conference of the Spanish Association for Artificial Intelligence (CAEPIA), Salamanca, Spain*. pp. 217–226.

Mikolajczyk, K. and C. Schmid (2001). Indexing based on scale invariant interest points. In: *Int. Conf. on Computer Vision*. p. 525.

Mikolajczyk, K. and C. Schmid (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(10), 1615–1630.

Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool (2005). A comparison of affine region detectors. *International Journal of computer Vision* **65(1/2)**, 43–72.

Murillo, A. C., J. J. Guerrero and C. Sagüés (2007). Surf features for efficient robot localization with omnidirectional images. In: *IEEE Int. Conf. on Robotics & Automation*.

Ribas, D., P. Ridao, J.D. Tardós and J.Ñeira (2008). Underwater slam in man-made structured environments. In: *Journal of Field Robotics*. pp. 1–24.

Rodriguez-Losada, D., F. Matia, A. Jimenez, R. Galan and G. Lacey (2005). Guido, the robotic smartwalker for the frail visually impaired. In: *First International Congress on Domotics, Robotics and Remote Assistance for All. DRT4ALL'05.*. Madrid, Spain. pp. 155–169.

Schmid, C., R. Mohr and C. Bauckhage (2000). Evaluaton of interest point detectors. *International Journal of computer Vision* **37**(2), 151–172.

Se, Stephen, David G. Lowe and Jim Little (2001). Vision-based mobile robot localization and mapping using scale-invariant features. In: *IEEE Int. Conf. on Robotics & Automation*. pp. 2051–2058.

Sim, R., P. Elinas, M. Griffin and J. Little (2005). Vision-based slam using the rao-blackwellised particle filter. In: *IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR)*.

Smith, S.M. (1992). A new class of corner finder. In: *British Machine Vision Conference*. pp. 139–148.

Soria, C., F. Roberti, R. Carelli and J.M. Sebastian (2008). Control servo-visual de un robot manipulador planar basado en pasividad. *RIAI* **5**, 54–61.

Triebel, R. and W. Burgard (2005). Improving simultaneous mapping and localization in 3D using global constraints. In: *National Conference on Artificial Intelligence (AAAI)*. Vol. 3. pp. 1330–1335.

Valls Miro, J., W. Zhou and G. Dissanayake (2006). Towards vision based navigation in large indoor environments. In: *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*. pp. 2096–2102.

Zernike, F. (1934). Diffraction theory of the cut procedure and its improved form, the phase contrast method. *Physica* **1**, 689–704.

Zhang, Z., R. Deriche, O. Faugeras and Q. Luong (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence* **78**, 87–119.