CENTRO DE INVESTIGACIONES
EN ÓPTICA, A.C.

LOCALIZACIÓN DE ROBOTS HUMANOIDES BASADA EN APARIENCIA A PARTIR DE UNA MEMORIA VISUAL

Como requisito para obtener el grado de:

Maestro en Optomecatrónica

Asesor:

Dr. J. Ascención Guerrero Viramontes

Co-asesor:

Dr. Héctor Manuel Becerra Fermín

Presenta:

Ing. Noé Guadalupe Aldana Murillo

"Versión Definitiva. Incluye cambios sugeridos por revisores."

Octubre de 2014
León, Guanajuato, México

Agradecimientos

Agradezco en primer lugar al Dr. Jean Bernard Hayet (CIMAT) por su disponibilidad, paciencia, dedicación y confianza para transmitirme sus conocimientos y poder realizar el presente trabajo de tesis.

Agradezco también a mis asesores Dr. J. Ascención Guerrero Viramontes y Dr. Héctor Manuel Becerra Fermín (CIMAT), por sus consejos y quienes sin su ayuda no podría haber terminado este trabajo.

A mis padres, hermanos y hermana.

Le doy gracias al Dr. David Moreno Hernández por tomarse el tiempo de revisar el presente trabajo y por sus sabios consejos en cuanto el desarrollo del mismo.

Y por último apremio al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado durante mis estudios.

Dedicatoria

A MIS PADRES

Mil gracias por ser los mejores ejemplos que he podido tener. Les doy gracias por todos sus cuidados y porque siempre creyeron en mi. Son las mejores personas que conozco y les dedico este trabajo. Gracias por todo su amor y apoyo. Este trabajo es suyo.

Resumen

El objetivo planteado en esta tesis es la localización de un robot humanoide a partir de una memoria visual. Una memoria visual es un conjunto de imágenes organizadas como un mapa topológico que modela un ambiente. El problema de localización en una memoria visual consiste en encontrar la imagen de la memoria visual más similar a la observada por la cámara a bordo del robot en su ubicación actual. Para la solución de este problema de localización basada en apariencia, se requiere comparar la imagen de entrada con todas las imágenes de la base de datos que forman la memoria visual. Para la localización del robot, se utilizó como herramienta el algoritmo de bolsa de palabras (BoW, por sus siglas en inglés). Mediante esta herramienta, se construyeron siete vocabularios visuales, todos basados en el uso de características locales: como BRIEF, ORB, SURF, y características en el espacio de color, en particular en el espacio HLS. Se construyó un vocabulario para uso especifico en robots humanoides: el vocabulario BRIEFROT. Después de haber creado los vocabularios visuales, se realizaron pruebas para sintonizar los parámetros de los vocabularios a partir de la mayor cantidad de resultados correctos obtenidos. Una vez ajustados los parámetros, se ejecutaron pruebas con cinco bases de datos distintas: tres en interiores y dos en exteriores. Las pruebas se realizaron conociendo de antemano la localización del robot y sin el conocimiento previo de la posición actual del robot. Los algoritmos se implementaron en C++ usando las librerías OpenCV 2.4.3 y DBow2. Los resultados demuestran que el vocabulario BRIEFROT, creado en esta tesis para el uso en robots humanoides, logró un buen desempeño durante las pruebas realizadas con las bases de datos. Además, dicho vocabulario también mostró una adecuada robustez a oclusiones y cambios de iluminación.

Índice general

Agradecimientos **I**

Dedicatoria **III**

Resumen **V**

Índice de figuras **XIII**

Índice de tablas **XV**

1. Introducción **1**

1.1. Antecedentes 1

1.2. Planteamiento del problema 2

1.3. Objetivos 3

1.3.1. Objetivo general 3

1.3.2. Objetivos particulares 3

1.4. Contribuciones del trabajo de tesis 3

1.5. Estado del arte 5

1.6. Estructura del documento 7

2. Fundamentos teóricos **9**

2.1. Técnicas de comparación de imágenes 10
2.2. Algoritmo de bolsa de palabras 11
2.2.1. Extracción de características 12
2.2.2. Creación del vocabulario visual 16
2.2.3. Representación de una imagen en bolsa de palabras. 17
2.3. Localización basada en apariencia mediante BoW 20
2.4. Evaluación de la salida de la localización 24
2.4.1. Cuando el ground truth no está disponible 24
2.4.2. Ground truth disponible 25

3. Desarrollo **27**
3.1. Creación de vocabularios visuales 29
3.2. Ajuste de parámetros para la creación de vocabularios visuales 31
3.3. Evaluación de características visuales con ground truth 32
3.4. Evaluación de características visuales sin ground truth 33

4. Análisis de resultados **35**
4.1. Vocabularios visuales 35
4.2. Ajuste de parámetros 37
4.2.1. Parámetros del árbol y medida de similitud 40
4.2.2. Mejora en el desempeño de los vocabularios de malla de color . . . 42
4.3. Evaluación de las características visuales con ground truth 47
4.3.1. CIMAT-NAO-A 47
4.3.2. CIMAT-NAO-B 49
4.3.3. Bicocca 2009-02-25b 54
4.3.4. New College 56

4.3.5. Málaga 56

4.4. Evaluación de las características visuales sin ground truth 60

5. Conclusiones y Trabajo Futuro **67**

Referencias **75**

Índice de figuras

1.1. Problema de localización basado en apariencia en el contexto de navegación usando una memoria visual. 4

2.1. Representación de una imagen en bolsa de palabras. 11

2.2. Punto candidato FAST. p es el punto candidato y los cuadrados resaltados son los píxeles del círculo de radio 3. Crédito: Rosten and Drummond (2006). 14

2.3. Proceso de construcción de un árbol de vocabulario. El grupo inicial de descriptores se va dividiendo en subgrupos cada vez más refinados hasta llegar a las palabras visuales. En este caso se tiene $k_w = 3$ (número de sub-grupos en cada nodo) y $L_w = 2$ (profundidad del arbol). 18

2.4. Conversión de descriptores en palabras visuales. El descriptor recorre cada nivel L_w del árbol hasta llegar a la palabra visual más cercana. En este caso se tiene $k_w = 3$ y $L_w = 2$. 19

2.5. Metodología para la comparación de la imagen de entrada con las imágenes de la memoria visual basada en un algoritmo de bolsa de palabras. 21

3.1. Metodología de desarrollo de un esquema de localización usando el algoritmo de bolsa de palabras. 28

4.1. Clases del método propuesto para la localización del robot. 36

4.2. Ejemplo de imágenes de prueba de la base de datos para ajustar los parámetros. 39

4.3. Diagramas de caja con $k_w = 8$ grupos para 4 y 5 niveles de árbol del vocabulario BRIEFROT. 41

4.4. Diagramas de caja con $k_w = 8$ grupos para 6, 7 y 8 niveles de árbol del vocabulario BRIEFROT. 43

4.5. Diagramas de caja con $k_w = 9$ grupos para 6, 7 y 8 niveles de árbol del vocabulario BRIEFROT. 44

4.6. Diagramas de caja con $k_w = 10$ grupos para 6 y 7 niveles de árbol del vocabulario BRIEFROT. 45

4.7. Combinaciones de agrupamientos con niveles del árbol para el vocabulario BRIEFROT. 46

4.8. Comparación de vocabulario de malla de color con diferentes tamaño de empalme. 46

4.9. Niveles de confianza obtenidos por los vocabularios visuales para la base de datos CIMAT-NAO-A. 48

4.10. Imágenes de prueba con su correspondiente imagen de la base de datos de acuerdo al vocabulario BRIEFROT. 50

4.11. Niveles de confianza obtenidos por los vocabularios visuales para la base de datos CIMAT-NAO-B. 51

4.12. Imágenes de prueba con su correspondiente imagen de la base de datos de acuerdo con el vocabulario ORB. 53

4.13. Niveles de confianza obtenidos por los vocabularios visuales para la base de datos Bicocca25b. 55

4.14. Niveles de confianza obtenidos por los vocabularios visuales para la base de datos New College. 57

4.15. Niveles de confianza obtenidos por los vocabularios visuales para la base de datos Málaga. 59

4.16. κ_1 para la base de datos CIMAT-NAO-A. La línea roja muestra los resultados correctos. 61

4.17. Niveles de confianza κ_1 y κ_2 para la base de datos CIMAT-NAO-B. La línea roja muestra los resultados correctos. 62

4.18. Niveles de confianza κ_1 y κ_2 para la base de datos Bicocca25b. La línea roja muestra los resultados correctos. 63

4.19. Niveles de confianza κ_1 y κ_2 para la base de datos New College. La línea roja muestra los resultados correctos. 64

4.20. Niveles de confianza κ_1 y κ_2 para la base de datos Málaga. La línea roja muestra los resultados correctos. 65

Índice de tablas

2.1. Detectores/descriptores de características de imagen empleados en la tesis. 13

4.1. Bases de datos usadas para la evaluación de características. 47

4.2. Porcentajes de resultados correctos para la base de datos CIMAT-NAO-A. 49

4.3. Porcentajes de resultados correctos para la base de datos CIMAT-NAO-B. 52

4.4. Porcentajes de resultados correctos para la base de datos CIMAT-NAO-B con cambio de condiciones. 54

4.5. Porcentajes de resultados correctos para la base de datos Bicocca25b. . . . 56

4.6. Porcentajes de resultados correctos para la base de datos New College. . . 58

4.7. Porcentajes de resultados correctos para la base de datos Málaga. 60

Capítulo 1

Introducción

1.1. Antecedentes

Una estrategia de navegación de robots que utiliza un sistema de visión como único sensor es la navegación basada en una memoria visual. La navegación basada en una memoria visual imita el comportamiento de un ser humano de recordar imágenes de donde ha estado previamente, para después usarlas para ubicarse y moverse más eficientemente cuando se encuentra en el mismo ambiente. Adicionalmente, se ha mostrado que la gente confía más en un mapa topológico que en un métrico al momento de desplazarse en su entorno [van Veen et al., 1998]. Un mapa topológico es un conjunto de imágenes de referencia que permiten recorrer un entorno sin necesidad de tener información métrica.

Una memoria visual es un conjunto de imágenes capturadas por una cámara fija a bordo de un robot y ordenadas como un mapa topológico que modela el ambiente por el cual, el robot tiene permitido navegar. Al conjunto de imágenes que forman una memoria visual también se les denomina imágenes clave. Estas imágenes se capturan previamente en una etapa de aprendizaje del ambiente, típicamente supervisada. Desde un punto de vista computacional, una memoria visual es una estructura de datos formada como un multigrafo, cuyos vértices son las imágenes clave, ligados por aristas que son los caminos visuales (grafos dirigidos). La composición de caminos visuales forman una ruta visual que permitirá ir desde un punto inicial a un punto deseado del multigrafo [Courbon et al., 2009].

La navegación con visión como único sensor también puede usar el enfoque de la

construcción densa de modelos 3D o de los mapas dispersos 3D obtenidos a partir de los datos visuales de entrada, como en [Royer et al., 2007] y [Davison et al., 2007]. Sin embargo, este enfoque es más demandante en tiempo de cómputo y en memoria. También, para modelos grandes el error métrico se acumula, por lo que las posiciones estimadas tienen deriva a lo largo del tiempo, como sucede en el planteamiento de la construcción de mapas y localización simultanea (SLAM) [Dissanayake et al., 2001]. Los mapas topológicos superan estas desventajas y son útiles para navegación, lo cual significa que, en la mayoría de los casos, un mapa geométrico contiene más información que la necesaria para que un robot se mueva en un ambiente.

1.2. Planteamiento del problema

En este trabajo de tesis, se aborda el problema de la localización de un robot humanoide en el contexto de navegación basada en una memoria visual, donde el robot tiene la tarea de llegar a una ubicación deseada que es especificada mediante una imagen objetivo capturada previamente. Específicamente, se afronta el problema de la localización basada en apariencia —es decir, localizarse usando únicamente características visuales contenidas en las imágenes—. En este problema, se requiere una comparación de la imagen observada actualmente por el robot con todas las imágenes de la memoria visual para encontrar la "más similar".

Este es un problema que se debe resolver de forma eficiente y que presenta retos interesantes debido a la baja calidad de las imágenes capturadas por plataformas humanoides típicas y, más aún, debido a problemas de imágenes borrosas provocadas por la locomoción de estos sistemas robóticos.

En un esquema de navegación de robots basado en una memoria visual, se especifican como entradas del sistema de navegación un número de identificación de una de las imágenes de la memoria visual asociada a la ubicación de destino (imagen objetivo final), así como la imagen inicial que corresponde a la observada por la cámara a bordo en la ubicación actual (imagen inicial). Una vez especificada la imagen inicial, el robot debe localizarse en el mapa topológico de imágenes del que dispone; enseguida, debe planificar una ruta visual formada por una secuencia de imágenes clave que deberá seguir para alcanzar la ubicación especificada por la imagen objetivo. La generación de la ruta visual no se aborda en el presente trabajo de tesis. En la Figura 1.1, se puede observar un esquema

del problema concreto abordado en esta tesis: la localización en el contexto de navegación basada en una memoria visual. Cabe resaltar que en esta tesis se supone que la memoria visual ha sido dada previamente.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un sistema de localización de un robot humanoide mediante un método basado en apariencia en el contexto de navegación utilizando una memoria visual.

1.3.2. Objetivos particulares

- Explorar la combinación de métodos locales y globales en el proceso de obtención de similitud de las imágenes capturadas por una cámara a bordo de un robot humanoide.
- Definir los mejores descriptores locales que permitan una eficiente comparación de las imágenes capturadas por una cámara a bordo de un robot humanoide.
- Resolver el problema de localización de forma eficiente computacionalmente considerando robustez a cambios de iluminación y oclusiones.

1.4. Contribuciones del trabajo de tesis

- Creación de un vocabulario visual binario para uso específico en robots humanoides.
- Desarrollo de tres vocabularios visuales de color usando el espacio de color HLS (Matiz, Iluminación y Saturación).
- Implementación de métricas para la evaluación de la salida del algoritmo de localización.
- Integración de los métodos desarrollados en una librería existente de C++ para localización basada en apariencia.

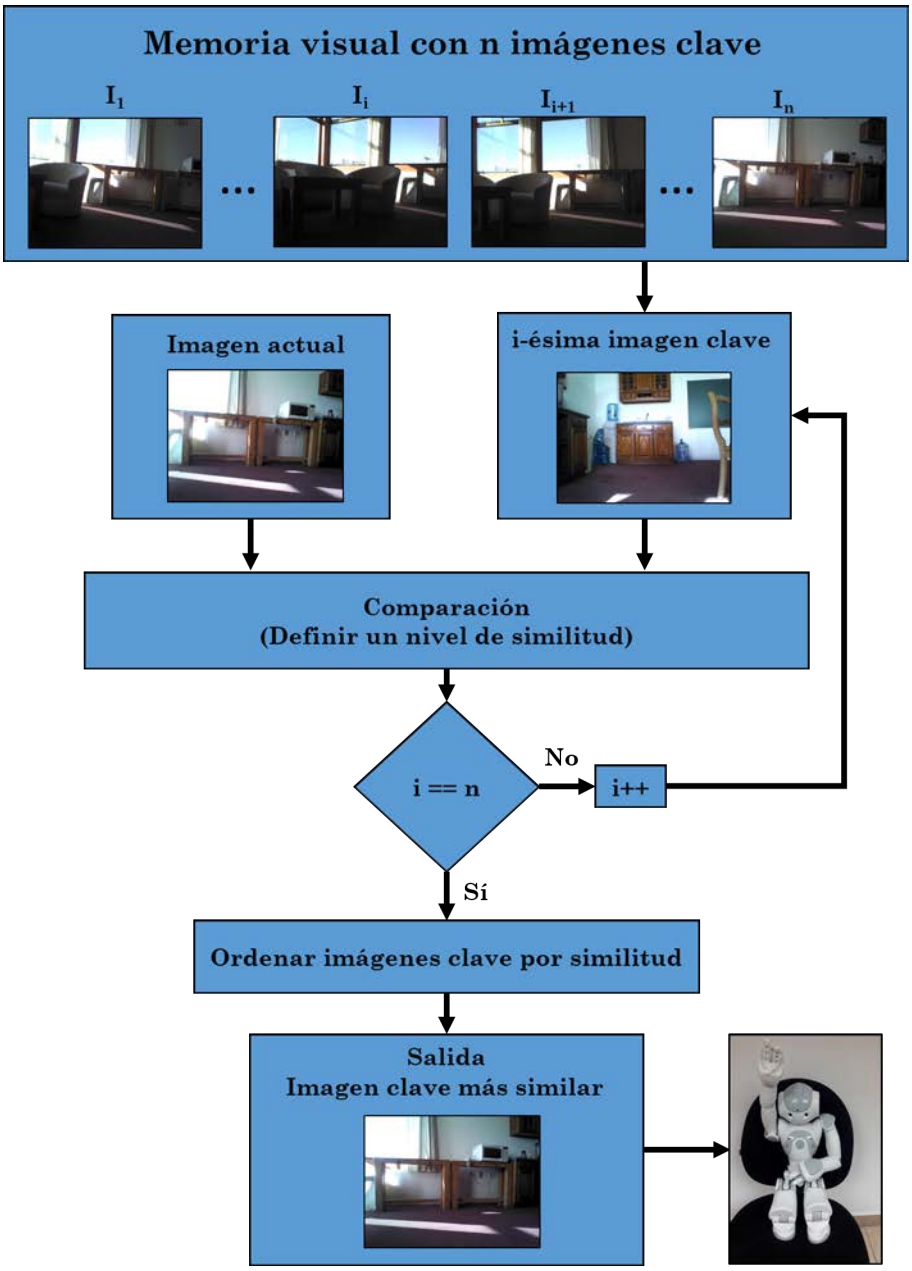


Figura 1.1: Problema de localización basado en apariencia en el contexto de navegación usando una memoria visual.

1.5. Estado del arte

La localización global exige un tiempo considerable de cómputo, ya que en este punto no se tiene idea de la localización del robot y hay que comparar la imagen actual con cada imagen del mapa topológico para encontrar la mejor correspondencia. Sin embargo, esto se realiza fuera de línea y sólo se requiere una vez antes de iniciar la navegación autónoma.

Una vez localizado el robot en su mapa geométrico, topológico o combinación de ambos, se realiza la planificación de la ruta a seguir en el espacio correspondiente, siendo una planificación geométrica clásica en el caso de contar con un mapa de reconstrucción 3D, como en [Royer et al., 2007]. En este caso, la localización se realiza emparejando puntos de interés entre la imagen actual y cada una de las imágenes clave y se calcula la postura de la cámara usando un enfoque robusto como Random Sample Consensus (RANSAC). La postura obtenida con el mayor número de correspondencias correctas (inliers) es una buena estimación de la postura actual de la cámara. Esta postura inicial se usa entonces para actualización en línea, considerando incertidumbre en la estimación con un enfoque Bayesiano. Aunque el esquema de [Goedemé et al., 2007] no realiza la reconstrucción de los puntos de interés en un ambiente 3D, también estima la localización inicial y la actualiza en la navegación mediante el uso de hipótesis refinadas mediante un enfoque probabilístico, es decir, un algoritmo de Simultaneous Localization And Mapping (SLAM) basado en un Extended Kalman Filter (EKF). A partir de la postura actual calculada por el algoritmo de localización, se puede encontrar una ruta usando, por ejemplo, el algoritmo de Dijkstra. Esta ruta se expresa como una serie de lugares topológicos, los cuales deben ser recorridos. Ambos esquemas basados en localización geométrica requieren una cámara calibrada.

Por otro lado, se encuentran los esquemas de localización que utilizan información puramente extraída del espacio de la imagen. Tanto en [Matsumoto et al., 1996] como [Matsumoto et al., 1999] se utiliza el emparejamiento de una plantilla, es decir, una porción de la imagen de entrada. El proceso de emparejamiento se realiza dos veces (para dos vistas memorizadas consecutivas), y la imagen que tiene la mayor correlación con la plantilla proporciona la localización. El mismo grupo de trabajo de Matsumoto ha usado el mismo enfoque de localización basado en correlación de una plantilla para imágenes capturadas por un robot humanoide [Ido et al., 2009]. En este caso, el cálculo de la correlación puede ser seriamente afectado por el motion blur generado por la locomoción humanoide.

El método de [Menegatti et al., 2004] proporciona una forma directa de localización

jerárquica, al comparar el espectro de frecuencia de la imagen actual con el espectro de frecuencia del conjunto de imágenes clave. Este método de localización es jerárquico debido a que se inicia comparando el espectro de baja frecuencia y se va refinando al aumentar componentes de mayor frecuencia. En consecuencia, la precisión de la localización aumenta al tomar más componentes de frecuencia. Una vez localizado, el robot no realiza un plan de navegación, sino que lo hace de manera reactiva, de acuerdo con el movimiento actual relativo al conjunto de imágenes de la memoria visual.

Otro método de localización jerárquica basada en imagen para un conjunto de vistas omnidireccionales es el propuesto en [Courbon et al., 2009]. La primera fase del método se basa en una representación geométrica de la imagen que se deriva a partir de interpolación de superficies, esto es, las imágenes se representan como superficies 3D, y la altitud de algunos puntos de control de la superficie aproximada se usa como descriptores. Esto define una distancia global, y por medio de un umbral sobre esta distancia se propone un conjunto de imágenes candidatas similares a la imagen actual. Enseguida, se usa un enfoque local clásico basado en Zero Normalized Cross Correlation (ZNCC) entre parches alrededor de puntos de Harris, para finalmente encontrar la imagen más similar y concluir la localización. La etapa de planificación en este esquema consiste en encontrar un camino de longitud mínima en la memoria visual que conecta los vértices asociados a la imagen actual y a la imagen clave final. Los caminos visuales están pesados por el número de imágenes clave que contienen. De esta manera, la ruta de mínima longitud se puede encontrar con el algoritmo de Dijkstra.

Otros esquemas de navegación basados en memoria visual se despreocupan del problema de localización al ser proporcionada o conocida previamente [Šegvić et al., 2009], [Diosi et al., 2011]. Mediante una interfaz, un usuario selecciona la imagen clave más cercana a la ubicación actual del robot. Además, dado que los mapas topológicos generados en estos esquemas son relativamente simples, no se tiene una etapa explícita de planificación de la ruta visual.

El problema de localización usando información puramente de imagen ha sido ampliamente tratado fuera del contexto de navegación basada en memoria visual mediante diferentes técnicas. Por ejemplo, se ha usado la imagen completa como descriptor [Ulrich and Nourbakhsh, 2000] buscando similitudes por histogramas de color y técnicas de aprendizaje; se han usado descriptores de frecuencias sobre representaciones cilíndricas de imágenes panorámicas [Pajdla and Hlaváč, 1999]; se ha explotado el uso de un conjunto

de descriptores locales [Kosecka and Yang, 2004] y una restricción geométrica que los relacione [Murillo et al., 2007]. Además de estas técnicas relacionadas con visión, se integran técnicas para hacer robusta la búsqueda de similitud entre imágenes, como filtros de partículas [Tamimi et al., 2006] o una bolsa de palabras visuales (visual bag of words) [Filliat, 2007]. [Gálvez-López and Tardós, 2012] usa el algoritmo de bolsa de palabras en el reconocimiento de lugares ya visitados para detectar cierre de trayectorias mediante el uso de robots móviles tipo carro. Para verificar que realmente se obtuvo una buena comparación, se realiza una verificación geométrica entre las características de la imagen de entrada y la imagen candidata.

1.6. Estructura del documento

La presentación del trabajo de tesis, se organizó en los siguientes apartados:

En el capítulo 1, se muestran los antecedentes, el planteamiento del problema, los objetivos, las contribuciones y el estado del arte, los cuales proporcional al lector un panorama general del trabajo realizado en la tesis.

En el segundo capítulo, se presentan las herramientas usadas para la localización del robot humanoide: técnicas para la comparación de imágenes de forma general, el algoritmo de bolsa de palabras, medidas de similitud utilizadas para la localización del robot y la evaluación de la salida del algoritmo de localización.

El capítulo 3 describe la metodología empleada para la localización del robot humanoide, creación de vocabularios visuales, ajuste de parámetros del árbol para la creación del vocabulario visual, elección de una medida de similitud, realización de pruebas con ground truth y sin ground truth.

Dentro del capítulo 4, se exponen los resultados logrados al implementar el algoritmo de bolsa de palabras: los vocabularios visuales creados, los parámetros ajustados, las pruebas conociendo de antemano la localización del robot (ground truth) y sin el conocimiento de la misma (sin ground truth).

Finalmente, en el capítulo 5, se presentan las conclusiones del proyecto de tesis y la propuesta de trabajo a futuro.

Capítulo 2

Fundamentos teóricos

Para que el robot se pueda localizar dentro un ambiente previamente modelado mediante una memoria visual, se debe de realizar la comparación de la imagen de entrada, que es la observación actual del robot, con las imágenes del mapa topológico para encontrar la más similar. Para realizar la comparación, se puede usar la información de la imagen completa, pero esto conllevaría a utilizar una gran cantidad de memoria para almacenar todas las imágenes y, además, se requeriría una cantidad considerable de tiempo de cómputo. Utilizar representaciones de la imagen completa, por ejemplo histogramas, puede reducir el tiempo de cómputo, pero no podría hacer frente oclusiones y a severos cambios de puntos de vista. Una manera de solventar esto, es usar características locales que describen las áreas más sobresalientes de una imagen. Aunque se pueden representar a las imágenes mediante sus características locales, aún se puede necesitar una gran cantidad de memoria y un considerable tiempo de cómputo.

Una solución a estos problemas es transformar las características locales de las imágenes en *bolsas de palabras*. Mediante este esquema, una imagen puede ser representada por un simple vector numérico usando un vocabulario visual. Esto permite realizar comparaciones con grandes conjuntos de imágenes empleando menos tiempo de cómputo y también se requiere una menor cantidad de memoria para almacenar dicha información. Por lo tanto, se ha elegido usarlo para el esquema de localización propuesto.

A lo largo de este capítulo se revisan las herramientas usadas para la localización del robot humanoide. La primera parte, la sección 2.1, se centra en las técnicas empleadas para la comparación de imágenes de forma general. En el apartado 2.2, se describe con detalle

la técnica particular utilizada en esta tesis: el algoritmo de bolsa de palabras. Mientras que en 2.3, se explican las medidas de similitud utilizadas para la localización del robot. Finalmente, en la parte 2.4, se comenta el proceso de decisión de la salida del algoritmo de localización.

2.1. Técnicas de comparación de imágenes

Un buen método de comparación de imágenes es de suma importancia en un enfoque de navegación basado en visión. Para realizar la comparación, existen varias técnicas que se pueden emplear. Por un lado, hay métodos globales que calculan una medida de similitud usando todos los píxeles de la imagen. Descriptores globales tales como emparejamiento basado en histogramas de color y técnicas de aprendizaje [Ulrich and Nourbakhsh, 2000], transformada de Fourier-Mellin [Stricker et al., 2001], eigen-imágenes panorámicas [Jogan and Leonardis, 1999], proporcionan información de toda la imagen en un simple descriptor. Aunque estos métodos son rápidos, no pueden hacer frente a, por ejemplo, oclusiones y severos cambios de puntos de vista. Por otro lado, técnicas que trabajan a una escala local, extrayendo y reconociendo características de manera individual, pueden ser robustas a estos efectos. Por ejemplo, esquinas [Harris and Stephens, 1988],[Rosten and Drummond, 2006]; puntos de interés sobresalientes [Lowe, 2004a], [Bay et al., 2008]; segmentos de líneas [Canny, 1986], etc. La principal desventaja de las técnicas locales es el tiempo de cómputo empleado para realizar la comparación.

Las características locales describen las zonas sobresalientes de las imágenes y pueden soportar algunas transformaciones de imagen como escalamiento y rotación, pero necesitan una gran cantidad de memoria para su almacenamiento. Para reducir este inconveniente, se han propuesto varios esquemas de comparación. Uno de ellos es el algoritmo de bolsa de palabras, el cual reduce las características de la imagen mediante un *vocabulario visual* (también conocido como *diccionario*) [Sivic and Zisserman, 2003] a un vector numérico denominado vector de bolsa de palabras, con lo cual se obtiene una representación compacta de la imagen. Debido a esto, el tiempo de cómputo empleado para la comparación de imágenes se reduce.

Por lo anterior, para identificar con eficacia pares de imágenes con abundantes descriptores en común, un enfoque eficiente es utilizar el algoritmo de bolsa de palabras,

el cual se explica en la siguiente sección.

2.2. Algoritmo de bolsa de palabras

El algoritmo de bolsa de palabras visuales es semejante al algoritmo usado para recuperar información en archivos de texto [Baeza-Yates and Ribeiro-Neto, 1999]. En la recuperación de textos, los documentos se representan como histogramas de frecuencias que presentan la aparición de cada palabra. Para determinar la semejanza entre dos documentos, se cuenta el número de palabras similares que contienen y lo representativas que son. En el caso de las imágenes, éstas son representadas como documentos formados por *palabras visuales* que representan sus características, como se puede apreciar en la Figura 2.1.

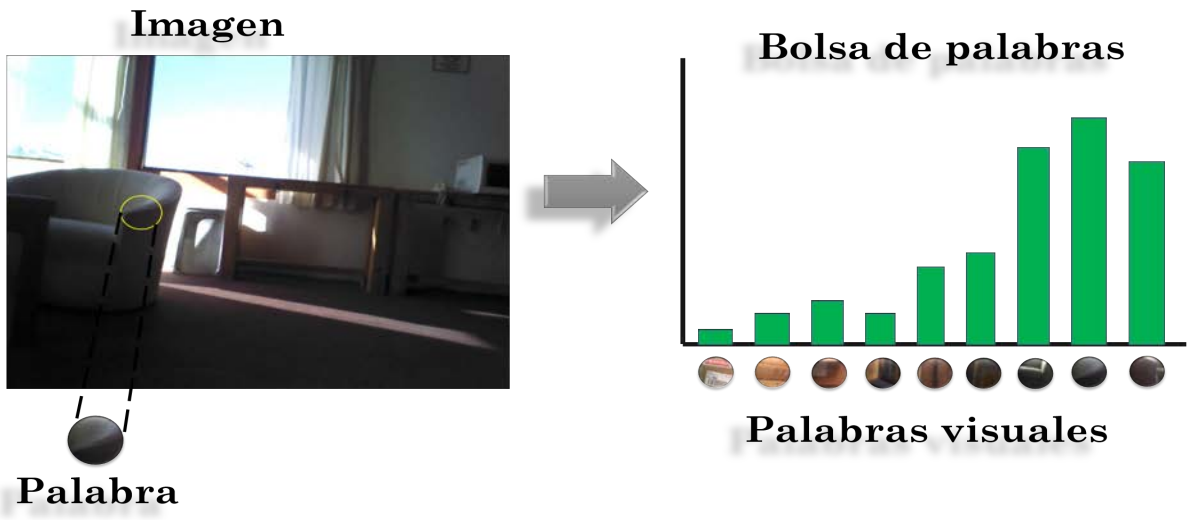


Figura 2.1: Representación de una imagen en bolsa de palabras.

La conversión de un conjunto de características de imagen en un vector de palabras visuales, a partir de un vocabulario visual, recibe el nombre de *bolsa de palabras*. A cada palabra se le asigna un peso de acuerdo con la relevancia que ésta tiene en el conjunto de características de entrenamiento; la ponderación disminuye en las palabras que aparecen de forma frecuente y que son consideradas menos discriminatorias.

Un vocabulario visual, como en el caso de la recuperación de información en

documentos de texto, es una distribución para representar un conjunto de características de una imagen en un histograma, lo que permite una representación compacta con bajos requerimientos de almacenamiento y una rápida comparación de imágenes.

A modo de resumen, el algoritmo de bolsa de palabras, para el entrenamiento, se puede dividir en los siguientes pasos:

1. Extraer un conjunto de características provenientes de algunas imágenes de entrenamiento (subsección 2.2.1).
2. Creación del vocabulario visual al agrupar características "similares"(subsección 2.2.2).
3. Representación de las imágenes de entrenamiento en bolsas de palabras (subsección 2.2.3).

A continuación se explican cada uno de estos pasos.

2.2.1. Extracción de características

La primera etapa en el algoritmo de bolsa de palabras es extraer, a partir de una secuencia de imágenes de entrenamiento, un conjunto de características locales distribuidas sobre todas esas imágenes, para, posteriormente, crear un vocabulario visual. Las características locales pueden ser definidas en puntos de interés o regiones de interés. Son utilizadas de manera amplia en visión por computadora para el reconocimiento de patrones, análisis de imagen, rastreo en vídeo, entre otros.

Los puntos de interés, también conocidos como *keypoints*, son puntos en una imagen que tienen una posición bien definida y pueden ser detectados de forma robusta en imágenes de una misma escena capturadas desde diferentes ubicaciones. El contenido de las vecindades de los puntos de interés puede ser descrito mediante vectores, mientras que las regiones de interés son porciones de imagen que pueden ser descritas ya sea por histogramas de color, textura, etc.

Para obtener las características locales de una imagen, se requieren dos etapas: detectar y describir dichas características. La etapa de detección se refiere a localizar las particularidades de la imagen que mejor la describen mediante algoritmos de detección;

a tales algoritmos también se les conoce sólo como detectores. En cambio, en la etapa de descripción se obtienen los parámetros (e.g., textura, color...) que definan las características encontradas en la etapa de detección. Esta descripción tiene forma de vector, por lo que se le conoce como vector descriptor o simplemente descriptor.

Para la creación del vocabulario, son necesarios los vectores descriptores de las características locales. Para encontrar descriptores de características locales, existen algoritmos que describen puntos de interés tales como BRIEF, BRISK, FREAK, ORB, SIFT y SURF. En el presente trabajo de tesis, se emplean descriptores tales como BRIEF, ORB y SURF. El descriptor SURF se empleó debido a que es robusto a cambios de iluminación, escala y rotación, aunque calcularlo requiere un costo computacional relativamente alto. Los descriptores ORB y BRIEF (aunque BRIEF no es invariante a cambios de rotación) son vectores binarios y requieren menos costo computacional que SURF. Las regiones de interés, por su parte, fueron descritas por histogramas de color, es decir, los histogramas de color son los descriptores de las regiones de interés. La tabla 2.1 resume los algoritmos usados en esta tesis para la detección/descripción de características de imagen. Los fundamentos teóricos de estos algoritmos se explican de forma breve a continuación:

Tabla 2.1: Detectores/descriptores de características de imagen empleados en la tesis.

Descriptor	Detector usado para ubicar puntos de interés
ORB	FAST
BRIEF	ORB
SURF	SURF

Puntos FAST

El algoritmo ORB usa el detector FAST para determinar un punto de interés, por lo que se hablará un poco de este detector. Las características FAST (Features from Accelerated Segment Tests) [Rosten and Drummond, 2006] son puntos que se detectan mediante la comparación de algunos píxeles —en escala de grises— en un círculo de radio 3. Si la intensidad $I_p - t$ de un píxel p de prueba es menor que la intensidad de n píxeles

consecutivos o la intensidad $I_p + t$ es mayor que la intensidad de dichos píxeles, entonces el píxel p es seleccionado como una esquina, donde t es un umbral (ver Figura 2.2).

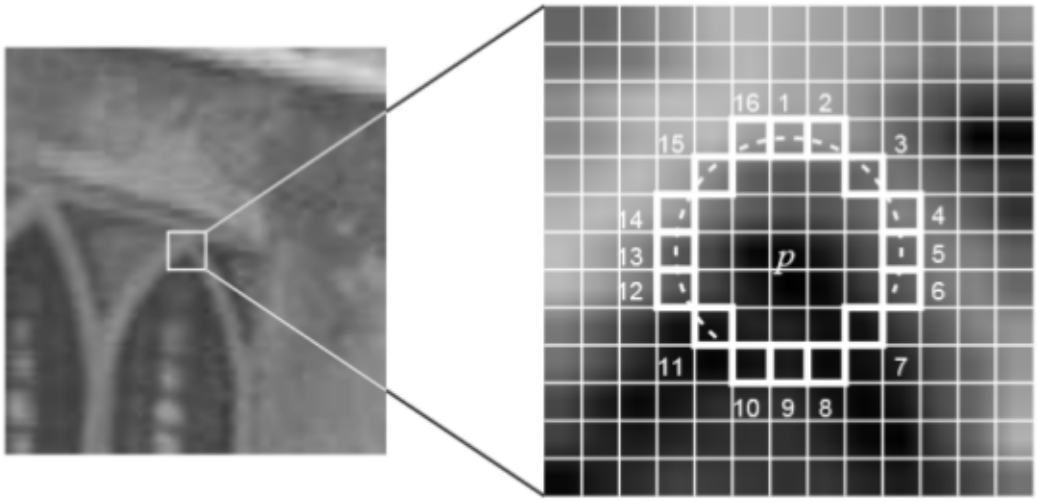


Figura 2.2: Punto candidato FAST. p es el punto candidato y los cuadrados resaltados son los píxeles del círculo de radio 3. Crédito: Rosten and Drummond (2006).

Descriptor BRIEF

BRIEF (Binary Robust Independent Elementary Features) [Calonder et al., 2010]. Para definir este descriptor, se selecciona un parche alrededor de un punto de interés. En esta tesis se usaron puntos ORB. El parche, cuyo tamaño es dado de antemano, es previamente suavizado por un kernel Gaussiano para reducir el ruido. El descriptor BRIEF es un vector binario donde cada bit es el resultado de una comparación entre un par de píxeles dentro del parche. El conjunto de pares de píxeles se seleccionan de manera aleatoria en una etapa previa a la creación de los descriptores. La longitud del descriptor es el número de pruebas a realizar (L_b). Cada bin del descriptor se define como:

$$\tau(i; x, y) = \begin{cases} 1, & \text{si I(x) < I(y)} \\ 0, & \text{de otra manera} \end{cases}$$

donde x y y son los píxeles seleccionados de manera aleatoria, I(.) la intensidad del píxel en la imagen suavizada e i es el índice del descriptor. La principal ventaja de los

descriptores BRIEF es que son muy rápidos de calcular y, por consiguiente, la comparación de descriptores es rápida.

Descriptor ORB

El descriptor ORB (Oriented FAST and Rotated BRIEF) [Rublee et al., 2011] es un vector binario que es invariante a los cambios de rotación. ORB emplea puntos FAST para detectar esquinas y construir descriptores binarios al igual que BRIEF.

El descriptor BRIEF se calcula realizando comparaciones de píxeles que no toman en cuenta la rotación de la imagen y, por lo tanto, no es robusto a esas transformaciones. ORB supera esta deficiencia mediante el cálculo de una orientación dominante entre el centro del punto de interés y el centroide de la intensidad de su parche [Rosin, 1999]. El patrón de comparaciones se rota con la finalidad de que el descriptor calculado sea poco propenso a variar cuando la imagen gira en el plano bidimensional.

La longitud del descriptor es de 256 bits. El valor de cada bit depende del resultado de la comparación de dos píxeles alrededor del punto clave. ORB difiere de BRIEF en la forma en que se realiza la comparación. BRIEF selecciona pares de píxeles de forma aleatoria en cada parche en una etapa previa al cálculo de los descriptores, por lo que para poder realizar las comparaciones, dichos píxeles deben ser recuperados. ORB utiliza un patrón de comparación fijo, obtenido por los pares de píxeles que proporcionan la variación más alta de bits en el conjunto de parches de entrenamiento. Debido a la variación máxima, se producen bits menos correlacionados, lo que mejora la naturaleza descriptiva de la característica. La variación a la escala no se obtiene directamente, pero en [Rublee et al., 2011] sugieren la extracción de características ORB después de escalar la imagen varios niveles.

Estas propiedades hacen de ORB un descriptor binario rápido para calcular y con un poder discriminatorio equiparable, en algunos casos, con SURF [Rublee et al., 2011].

Descriptor SURF

SURF (Speeded-Up Robust Features) fue desarrollado por [Bay et al., 2008] como un detector de puntos de interés y descriptor robustos. El descriptor SURF guarda cierta similitud con SIFT [Lowe, 2004b], pero es más rápido para calcular. Se realiza en cuatro

etapas: 1. Cálculo del determinante del Hessiano: En esta etapa, se realiza un calculo de una cantidad que permitirá detectar puntos de interés que sean invariantes a los cambios de rotación y de escala. SURF utiliza el valor del determinante de la matriz Hessiana para el cálculo de la posición y la escala de los puntos de interés. 2. Localización de puntos clave: La localización de los puntos de interés se obtiene mediante la eliminación de los puntos que no cumplan con una condición de máximo en un vecindario del punto. 3. Asignación de orientación: Se asigna una orientación a cada punto de interés de acuerdo a la respuesta obtenida, por los puntos de interés, a filtros de Haar. De esta forma se obtiene invarianza a la rotación. 4. Descriptor del punto de interés: En esta etapa, los puntos de interés son representados mediante vectores descriptores como una medida de las respuestas a filtros de Haar en las direcciones horizontal y vertical, respectivamente, relativas a la orientación del punto de interés. El tamaño del vector descriptor es de 64 (256 bytes).

Descriptor de regiones de color

Para el caso de las regiones locales, en esta tesis, se usaron parches o ventanas rectangulares de imagen, en las cuales, se utilizaron histogramas de color como descriptores. El modelo HSL (matiz o tonalidad, luminosidad, saturación) facilita la caracterización de los colores ya que en este modelo las características de saturación, iluminación y tonalidad se encuentran de forma explícita como partes del modelo. La tonalidad corresponde a la percepción del color; la saturación describe la pureza del color, es decir, la intensidad o palidez del mismo. Se trabajó sólo con los canales de tonalidad y saturación con la finalidad de proporcionar robustez ante cambios de iluminación. El vector descriptor es un histograma 2D de matiz y saturación.

2.2.2. Creación del vocabulario visual

Un vocabulario visual se crea cuando los descriptores de entrenamiento se asocian a un conjunto finito de grupos enumerados, de manera que cualquier descriptor puede ser representado por el grupo al cual pertenece. Gracias a este agrupamiento, se puede representar —por el índice de su grupo— a un vector descriptor multidimensional como un número entero (una “palabra”); también, facilita la correspondencia entre las características similares.

El vocabulario visual es creado fuera de línea mediante la división de los vectores descriptores en W grupos: las denominadas palabras visuales. [Sivic and Zisserman, 2003] presentaron una simple distribución de los descriptores en K agrupamientos; en cambio, [Nister and Stewenius, 2006] propusieron un agrupamiento jerárquico para mejorar el rendimiento. Nuestro enfoque usa también este agrupamiento jerárquico. En este tipo de vocabulario, el espacio descriptor se agrupa de forma recursiva en k_w regiones, las cuales se van haciendo cada vez más finas, hasta L_w niveles, lo que produce $W = k_w^{L_w}$ grupos finales o palabras visuales.

Con el enfoque propuesto por [Sivic and Zisserman, 2003], un punto en el espacio de los descriptores debe compararse con K grupos para seleccionar uno al cual pertenece, es decir, debe compararse con todos los descriptores; en cambio, el método propuesto por [Nister and Stewenius, 2006] disminuye el número de comparaciones a $k_w \cdot L_w$, porque sólo un subconjunto de los grupos se considera en cada paso, lo que resulta más eficiente que el método de [Sivic and Zisserman, 2003].

El árbol de vocabulario jerárquico se construye a partir de un conjunto de características de algunas imágenes de entrenamiento. El método se inicia con un proceso de agrupamiento en los datos de entrenamiento para generar un primer nivel usando un algoritmo k-mean, el cual es un algoritmo de agrupamiento que tiene como objetivo la partición de un conjunto de n datos en k grupos y que realiza la elección de datos de forma aleatoria. En particular, se usa el algoritmo k-means++ [Arthur and Vassilvitskii, 2007], que además de elegir los datos de forma aleatoria, los toma de forma bien distribuida. De esta manera, en el primer proceso de agrupamiento el algoritmo k-means++ elige k_w datos que serán los centros de los grupos. Los descriptores son distribuidos entre los k_w grupos, donde cada grupo consiste de los descriptores más cercanos a su centro. Este agrupamiento forma el primer nivel del árbol de vocabulario. Niveles subsecuentes se construyen repitiendo la operación de agrupamiento k-means++ con los descriptores asociados a cada nodo hasta L_w niveles. La Figura 2.3 muestra el proceso para la construcción del árbol de vocabulario.

2.2.3. Representación de una imagen en bolsa de palabras.

Para convertir una imagen en bolsa de palabras (BoW) a partir de un vocabulario visual, se requiere representar dicho vocabulario en un vector numérico con la finalidad de

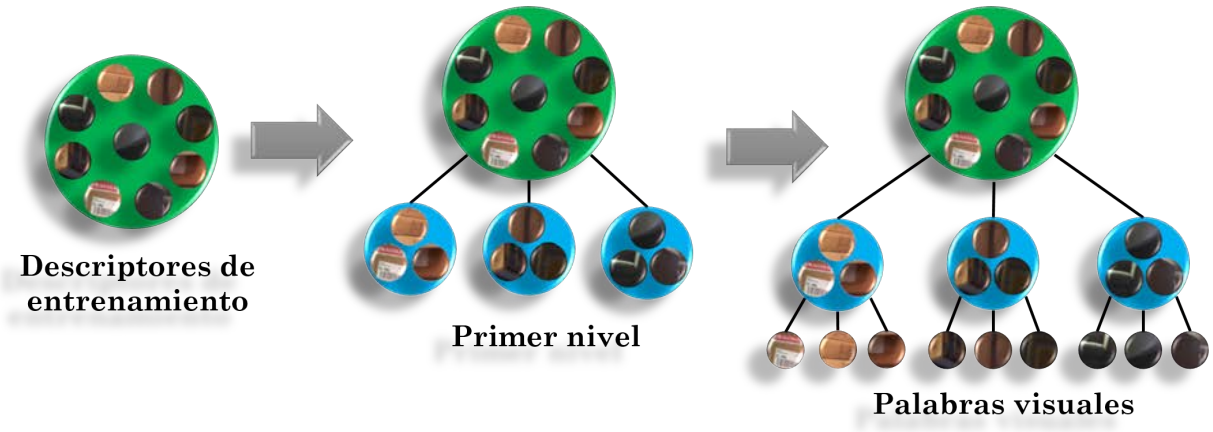


Figura 2.3: Proceso de construcción de un árbol de vocabulario. El grupo inicial de descriptores se va dividiendo en subgrupos cada vez más refinados hasta llegar a las palabras visuales. En este caso se tiene $k_w = 3$ (número de sub-grupos en cada nodo) y $L_w = 2$ (profundidad del arbol).

realizar la comparación de las imágenes. En el presente trabajo de tesis, se usa un peso para cada palabra visual, como lo ha propuesto [Sivic and Zisserman, 2003]. En dicho método, a cada palabra $\mathbf{w}_i$ del vocabulario se le asigna un peso de acuerdo a su aparición durante la etapa de entrenamiento: si la palabra aparece de manera frecuente, el valor de este peso disminuye, debido a que es menos discriminatoria. La ponderación de cada palabra $\mathbf{w}_i$ involucra el término denominado *idf* (inverse document frequency), el cual está dado por:

$$\mathrm{idf}(i) = \log\left(\frac{N}{n_i}\right), \tag{2.1}$$

donde N es el número total de las imágenes de entrenamiento; n_i, el número de imágenes en donde aparece la palabra $\mathbf{w}_i$. $\mathrm{idf}(i)$ es, entonces, más grande cuando aparece con menos frecuencia en la base de datos.

Para convertir una "nueva"imagen $\mathbf{I}$ en un vector de bolsa de palabras $\mathbf{v} \in \mathbb{R}^W$, donde W es el vocabulario visual, cada descriptor encontrado en $\mathbf{I}$ cruza el árbol, comparándose en cada nivel con los descriptores asociados a los nodos y eligiendo el más cercano. Una vez que el descriptor haya recorrido el árbol hasta una hoja, habrá encontrado la palabra de imagen, es decir el descriptor en el árbol que mejor lo representa (ver Figura 2.4). Esto permite

calcular el llamado término de frecuencia (tf) de cada palabra en esa "nuevaïmagen:

$$\mathrm{tf}(i, \mathbf{I}) = \frac{n_{iI}}{n_I}, \tag{2.2}$$

donde n_{iI} es el número de veces que aparece la palabra $\mathbf{w}_i$ en la imagen $\mathbf{I}$; n_I, el número total de palabras en $\mathbf{I}$. Luego, a la i-ésima entrada del vector $\mathbf{v}$ se le asigna el valor:

$$v^i = \mathrm{tf}(i, \mathbf{I}) \cdot \mathrm{idf}(i), \tag{2.3}$$

que es la ponderación final de la palabra i denominada como ($tf-idf$), la cual fue propuesta por [Sivic and Zisserman, 2003], y donde idf es el inverse document frecuency descrito arriba.

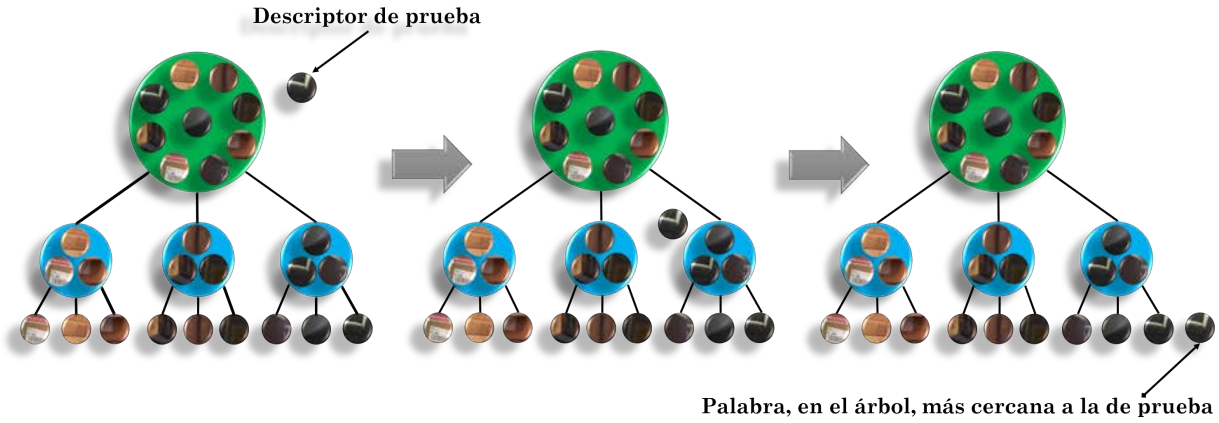


Figura 2.4: Conversión de descriptores en palabras visuales. El descriptor recorre cada nivel L_w del árbol hasta llegar a la palabra visual más cercana. En este caso se tiene $k_w = 3$ y $L_w = 2$.

Una vez hecho esto, la imagen queda representada como un vector numérico —bolsa de palabras— de frecuencias de aparición, ponderadas por la "rareza"de la palabra. Con esta representación, la comparación de imágenes se hace más rápida, y además, la comparación es precisa debido a que se emplean características locales. Los vectores de bolsas de palabras se normalizan con la finalidad de mitigar el efecto del número de palabras diferentes entre vectores cuando son comparados [Nister and Stewenius, 2006].

Con la representación de imágenes en vectores de bolsa de palabras, la comparación de imágenes es más rápida, pero también más eficiente. Eso es debido a que —mediante esta comparación— se obtienen las ventajas que los métodos locales poseen: una comparación precisa y robusta debido a que se emplean características locales y, además, rápida, ya

que la comparación se realiza entre vectores numéricos, los cuales son pocos debido a la estructura del árbol.

2.3. Localización basada en apariencia mediante BoW

La localización del robot se realiza comparando la imagen observada por el robot desde su ubicación actual (la cual es la entrada al algoritmo) con las imágenes del mapa topológico que modela el entorno, con la finalidad de encontrar la más similar. La comparación empleando el algoritmo de bolsa de palabras se representa en la Figura 2.5 y sigue estos pasos:

1. Creación de un vocabulario visual (W) empleando las imágenes de la memoria visual (offline).
2. Representación de las imágenes en bolsas de palabras a partir del vocabulario visual, generado en el paso 1, para crear una base de datos BoW_M (offline).
3. Representación de la imagen de entrada en bolsa de palabras (BoW_I). Se obtiene el vector $\mathbf{v}$ descrito arriba (online).
4. Comparación del vector de bolsa de palabras de la imagen de entrada BoW_I con los vectores de la base de datos BoW_M que tienen, al menos, cinco palabras en común con la imagen de prueba (online).

La base de datos obtenida mediante la representación en bolsa de palabras permite manejar una gran cantidad de imágenes, gracias a su representación ligera. Dicha base de datos, debe permitir una rápida comparación de imágenes. Para realizar la comparación de la base de datos de forma rápida, se usa un "archivo invertido".

Un archivo invertido es una estructura que almacena una lista de imágenes $\mathbf{I}_k$, en el vocabulario visual, donde aparece cada palabra $\mathbf{w}_i$. Esto es muy útil, ya que permite comparar las imágenes de la base de datos que tengan algunas palabras en común con la imagen de prueba. Además de las imágenes, en el archivo inverso, se almacena el peso de cada palabra en esas imágenes para permitir un rápido acceso. Por lo tanto, un archivo

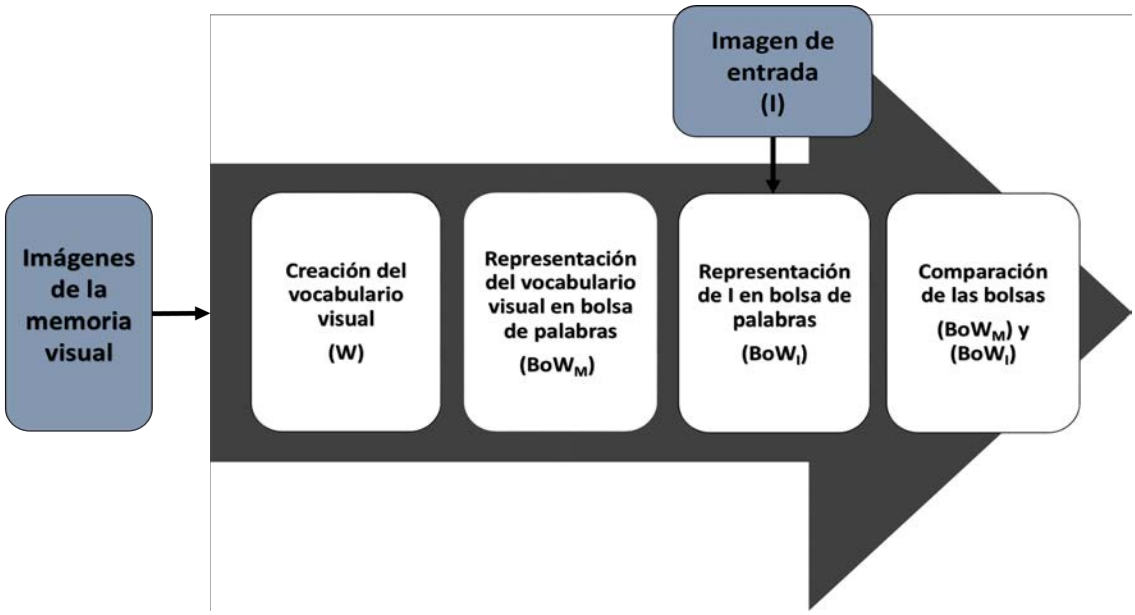


Figura 2.5: Metodología para la comparación de la imagen de entrada con las imágenes de la memoria visual basada en un algoritmo de bolsa de palabras.

inverso contiene para cada palabra $\mathbf{w}_i$ una lista de pares de valores $(\mathbf{I}_k,\ v_k^i)$, donde v_k^i es el valor de la i-ésima palabra en la imagen $\mathbf{I}_k$.

Lo que se obtiene es un conjunto limitado de vectores $\mathbf{w}_k$ de imágenes de la base de datos que comparten características con la imagen actual. Lo siguiente consiste en comparar cada una de estas imágenes con la imagen actual, con base en su vector $\mathbf{v}$ de bolsa de palabras.

La comparación se realiza para los vectores $\mathbf{w}_k$ que tienen, al menos, cinco palabras en común con el vector $\mathbf{v}$ de la imagen $\mathbf{I}_k$ de prueba. Para esto, se verifica el archivo inverso de cada palabra en el vector $\mathbf{v}$, donde se pueden obtener las imágenes que contienen la palabra. Gracias a esta operación, se encuentran las imágenes que tienen algunas palabras en común, y además, el peso de cada palabra. Es decir, en una sola operación, se encuentran las imágenes con, al menos, cinco palabras en común y el peso de cada una de ellas (el vector de bolsa de palabras). Después de obtener el peso de las palabras y las imágenes en común, se realiza la comparación sólo con estas imágenes, lo que reduce el tiempo de confrontación.

La comparación de los vectores de bolsa de palabras se realiza empleando una medida de similitud. Una métrica de similitud es una función matemática que asocia un valor

numérico entre un par de descriptores. Tal valor numérico es conocido con el nombre de distancia. Hay varias métricas estadísticas que se pueden aplicar para comparar dos vectores. En la presente tesis se emplearon las métricas: $L_1 - norm$, $L_2 - norm$, coeficiente de Bhattacharyya, distancia χ^2 y el producto punto. De cada métrica depende el valor numérico del resultado de la comparación. Los resultados obtenidos se escalan al intervalo [0...1], donde el valor 0 indica que no existe correspondencia entre vectores y 1, perfecta correspondencia. Las métricas empleadas para calcular la similitud s entre dos vectores $\mathbf{u}$ y $\mathbf{v}$ se explican a continuación.

L_1-*norm*

Para esta métrica, los vectores formados por frecuencias de palabras son normalizados con la norma L_1, donde $\| \mathbf{u} \|_1 = \sum_{i=1}^{n} |u_i|$ y $\| \mathbf{v} \|_1 = \sum_{i=1}^{n} |v_i|$. La puntuación de similitud usando la distancia L_1 se observa en la ecuación 2.4.

$$s(\mathbf{u}, \mathbf{v}) = \| \frac{\mathbf{u}}{\|\mathbf{u}\|_1} - \frac{\mathbf{v}}{\|\mathbf{v}\|_1} \|_1. \tag{2.4}$$

L_2-*norm*

La métrica L_2 es similar a la métrica L_1. En este caso, $\| \mathbf{u} \|_2 = \left(\sum_{i=1}^{n} |u_i^2| \right)^{1/2}$ y $\| \mathbf{v} \|_2 = \left(\sum_{i=1}^{n} |v_i^2| \right)^{1/2}$. La métrica L_2 se aprecia en la ecuación 2.5.

$$s(\mathbf{u}, \mathbf{v}) = \| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \|_2. \tag{2.5}$$

Producto punto

En ésta métrica se calcula el producto punto entre dos vectores $\mathbf{u}$ y $\mathbf{v}$ mediante la ecuación 2.6.

$$s(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v}. \tag{2.6}$$

Notar que en este caso, mientras más grande sea el valor de s, más cercanos serán $\mathbf{u}$ y $\mathbf{v}$.

Distancia χ^2

La distancia Chi-Cuadrada entre dos vectores proporciona una media ponderada de la diferencia entre todas las posiciones de estos vectores. Se puede calcular esta distancia χ^2, entre dos vectores, $\mathbf{u}$ y $\mathbf{v}$ usando la ecuación 2.7.

$$\chi^2(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{n} \frac{(\hat{u}_i - \hat{v}_i)^2}{(\hat{u}_i + \hat{v}_i)} \Rightarrow s(\mathbf{u},\mathbf{v}) = \sum_{i|\hat{u}_i+\hat{v}_i\neq 0}^{n} \frac{(\hat{u}_i - \hat{v}_i)^2}{(\hat{u}_i + \hat{v}_i)}, \tag{2.7}$$

donde los vectores unitarios $\hat{\mathbf{u}}$ y $\hat{\mathbf{v}}$ son normalizados como $\frac{\mathbf{u}}{\|\mathbf{u}\|_1}$ y $\frac{\mathbf{v}}{\|\mathbf{v}\|_1}$ respectivamente.

Dado que algunas entradas de los vectores tienen valores de cero, la distancia χ^2 no se puede aplicar a cada una de las componentes de los vectores, debido a que provoca problemas de división por cero. Como se muestra en la ecuación, esto se evita no considerando el índice para el cual la suma de componentes es nula.

Distancia Bhattacharyya

El coeficiente Bhattacharyya se utiliza para evaluar la similitud entre 2 vectores. La ecuación 2.8 describe el cálculo de la distancia:

$$s(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{n} \sqrt{\hat{u}_i \cdot \hat{v}_i}, \tag{2.8}$$

donde los vectores unitarios correspondientes son $\hat{\mathbf{u}}$ y $\hat{\mathbf{v}}$ normalizados como $\frac{\mathbf{u}}{\|\mathbf{u}\|_1}$ y $\frac{\mathbf{v}}{\|\mathbf{v}\|_1}$ respectivamente.

Cuando se compara un vector $\mathbf{v}$ —que representa el vector de bolsa de palabras de la imagen de entrada— con los vectores $\mathbf{w}_k$ de la base de datos, las distancias s se obtienen para aquellos vectores $\mathbf{w}_k$ que tienen al menos cinco palabras en común con $\mathbf{v}$. La serie de distancias obtenidas durante la comparación se ordenan de mayor a menor, de tal manera que el resultado de la imagen con el mayor parecido aparece en primer lugar.

2.4. Evaluación de la salida de la localización

Cuando se localiza una imagen $\mathbf{I}$ mediante el algoritmo de bolsa de palabras, se obtiene un conjunto de resultados o "scores" $s_l(\mathbf{I})$, donde l es un índice que se refiere a una imagen $\mathbf{I}$ de la memoria visual. En el presente trabajo de tesis, se evalúan dichos resultados con la finalidad de verificar si la localización se realizó correctamente. El problema es determinar un nivel de confianza para los resultados obtenidos en la etapa de comparación. En particular, interesa determinar la imagen o las imágenes de la memoria visual que maximicen un criterio de confianza en los resultados de comparación. Dependiendo de si se supone conocida o no la información real de localización, la salida del método propuesto se puede evaluar considerando dos casos: cuando se conoce el ground truth y cuando no se cuenta con ground truth. En seguida, se explica la evaluación propuesta para la salida del algoritmo para los dos casos.

2.4.1. Cuando el ground truth no está disponible

Este es el caso de interés general que se presenta en la realidad. Para encontrar un nivel de confianza, cuando no se tiene idea de la verdadera localización del robot, una opción es tomar en cuenta la distribución de los resultados $z_l(k)$, entre la imagen k de la base de datos y los de la imagen l de la misma base. Sea k^* el índice de la imagen que el método de localización arroja como la más parecida a la de prueba $\mathbf{I}$ —de acuerdo al diccionario actual—. El nivel de confianza se define como:

$$\kappa_1 = \frac{\sum_l z_l(k^*) s_l}{\sqrt{\sum_l (z_l(k^*)^2)} \sqrt{\sum_l s_l^2}}. \tag{2.9}$$

Esta ecuación es una correlación entre histogramas, un valor alto indica que existen histogramas similares, es decir, muestra que dos imágenes son semejantes, en el caso de un valor bajo quiere decir que no existe similitud entre las imágenes. La idea es que si una imagen es parecida a una de la base de datos, el conjunto de las puntuaciones obtenidas con esa misma imagen de referencia deberían de ser similares.

2.4.2. Ground truth disponible

En este caso, se da la imagen $\bar{k}$ más parecida a la imagen de prueba, para lo cual se pueden definir dos métricas μ de evaluación. Es claro que estas métricas son aplicables únicamente para efectos de evaluación del método de localización.

Así pues, dada la imagen k de la base de datos más parecida a la imagen de prueba, se puede usar la posición en la que la imagen $\bar{k}$ es encontrada por el algoritmo de comparación: un valor de uno es el mejor resultado posible para el algoritmo y significa que el algoritmo ha determinado a la imagen k en la primera posición del conjunto de resultados, y un valor elevado significa que se obtuvo un resultado incorrecto y significa que el algoritmo ha colocado a la imagen k lejos de la primera posición en el conjunto de resultados. De esta manera, se define μ_1 como:

$$\mu_1 = rank(\bar{k}), \tag{2.10}$$

donde el operador $rank$ regresa la posición de la imagen k en el conjunto de resultados $s_l(\mathbf{I})$.

El inconveniente de la ecuación 2.10 es que puede arrojar imágenes cercanas a $\bar{k}$ dentro de la base de datos y el resultado puede ser evaluado como "malo". Además no toma en cuenta la posible similitud entre las imágenes de la base de datos, por lo que se propone emplear una métrica que relacione ésta semejanza. Esto se logra al definir una métrica que tome en cuenta las distribuciones $z_l(\bar{k})$ de los resultados obtenidos al comparar las imágenes $\bar{k}_l$ de la base de datos con cada imagen l de la misma base. De esta manera, otra medida del nivel de confianza se define como:

$$\mu_2 = \sum_l \frac{z_l(\bar{k})}{\sum_{l'} z_{l'}(\bar{k})} rank(l) \tag{2.11}$$

Donde un valor de uno es el mejor valor posible e indica que las imágenes son semejantes. Un valor elevado significa un mal resultado de la comparación.

Como comentario final de esta sección, para el caso donde no se obtienen resultados $s_l(\mathbf{I})$, debido a que tienen menos de cinco palabras en común, las confianzas μ_1 y μ_2 se establecieron a cero para evitar problemas numéricos de división por cero durante el cálculo de la confianza μ_2.

Capítulo 3

Desarrollo

A partir del conocimiento del algoritmo de bolsa de palabras, se analizó su funcionamiento de una forma sistemática, es decir, se estudió la interrelación de cada etapa del algoritmo con el método de comparación de imágenes. A partir de lo anterior, se desarrollaron siete vocabularios visuales, empleando diferentes descriptores. Posteriormente, se llevó a cabo el ajuste de los parámetros del árbol para la creación del vocabulario visual, con la finalidad de obtener los mejores resultados posibles. Hecho esto, se procedió a la realización de pruebas empleando distintas bases de datos, con el propósito de verificar el funcionamiento de los vocabularios en diferentes condiciones ambientales. Por un lado, se hicieron pruebas conociendo los resultados que deberían de obtenerse en la comparación (con ground truth). Por otro lado, se efectuaron pruebas sin conocer el resultado de la comparación (sin ground truth).

Para la realización de esta tesis, se consideraron cuatro etapas fundamentales, la cuales se muestran en la Figura 3.1, y se describen a continuación en forma breve.

Etapa 1. Creación de vocabularios visuales.

Se realizaron vocabularios visuales empleando diferentes descriptores con la finalidad de ver el comportamiento del algoritmo de bolsa de palabras con distintas características de las imágenes.

Etapa 2. Ajuste de parámetros en la creación de vocabularios visuales.

En esta etapa, se ajustaron los parámetros empleados durante la creación de los vocabularios visuales. Específicamente, los parámetros del número de centros de

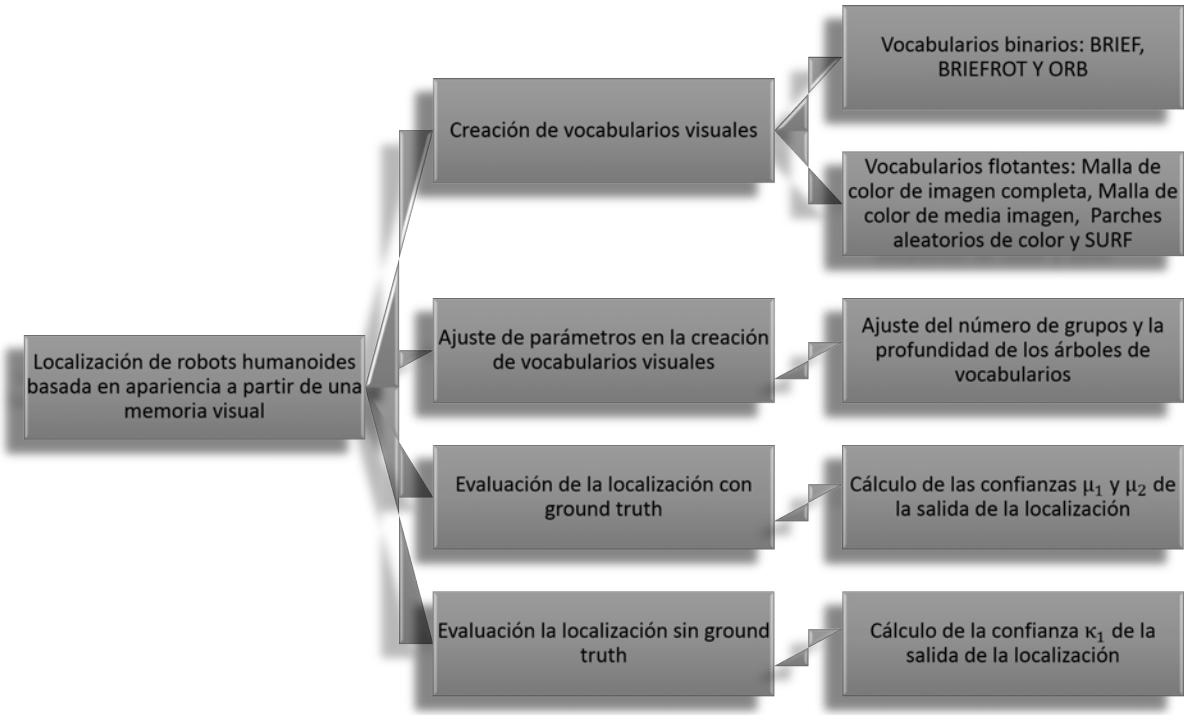


Figura 3.1: Metodología de desarrollo de un esquema de localización usando el algoritmo de bolsa de palabras.

agrupamiento k_w así como los niveles de profundidad L_w del árbol de vocabulario fueron modificados hasta obtener los mejores resultados en la comparación de imágenes.

Etapa 3. Evaluación de la localización con ground truth.

Se ejecutaron pruebas empleando cinco bases de datos distintas con la finalidad de verificar los resultados obtenidos bajo diferentes circunstancias ambientales. En estas pruebas, se conocían, de antemano, los resultados de las comparaciones, es decir, se sabía la localización del robot y se usó para verificar el desempeño del método propuesto.

Etapa 4. Evaluación de la localización sin ground truth.

Para finalizar, se efectuaron pruebas sin conocer previamente la ubicación del robot dentro del mapa topológico, situación real que se presenta al inicio de la navegación basada en una memoria visual.

3.1. Creación de vocabularios visuales

Los vocabularios desarrollados en esta tesis se basaron en los propuestos en [Gálvez-López and Tardós, 2012] donde se reporta el desarrollo de dos diccionarios, en los cuales se utilizaron descriptores SURF y BRIEF, para el reconocimiento de lugares ya visitados (cierre de trayectorias) en el problema de SLAM mediante el uso de robots móviles tipo carro. En la tesis doctoral [Gálvez-López, 2013], el autor presenta resultados usando vocabularios con descriptores ORB para el reconocimiento de objetos 3D. En la presente tesis, además de los vocabularios antes mencionados, se desarrollaron cuatro vocabularios adicionales: BRIEFROT, *Malla de color de imagen completa*, *Malla de color de media imagen* y *Parches aleatorios de color*.

Cada uno de los vocabularios contiene funciones tanto para implementar el algoritmo de bolsa de palabras como para realizar la localización del robot. Es decir, cada vocabulario extrae características de las imágenes, crea el vocabulario visual, representa las imágenes como bolsa de palabras y efectúa la localización del robot. Enseguida, sólo se explica la manera en que se obtuvieron los vectores descriptores para la creación de los vocabularios, ya que la etapa de localización es similar para todos los diccionarios, tal como se explicó en la sección 2.3.

Los descriptores BRIEF se han utilizado en el reconocimiento de lugares pa-

ra encontrar cierre de trayectorias empleando robots móviles tipo carro, como en [Gálvez-López and Tardós, 2012]. Con este tipo de robots, el movimiento de la cámara fija al robot es sólo en un plano y las imágenes capturadas no sufren cambios de rotación respecto al eje óptico. En el caso de los robots humanoides, existen variaciones de rotación respecto al eje óptico en las imágenes durante el desplazamiento, debido al efecto del balanceo generado por la locomoción bípeda. Para tratar de solventar dichas variaciones, el vocabulario BRIEF trabajado en esta tesis —a diferencia del creado en [Gálvez and Tardós, 2012]— emplea puntos ORB para la definición de los descriptores BRIEF con la finalidad de obtener puntos invariantes a la rotación.

El nuevo vocabulario, denominado BRIEFROT se desarrolló a partir de BRIEF y fue planteado específicamente para el uso en robots humanoides. También, usa puntos ORB para la definición de los descriptores BRIEF. Cuenta con tres vocabularios internos independientes, dos de los cuales están rotados un cierto ángulo: uno en sentido antihorario y el otro en sentido horario. Mediante experimentación se encontró que para la plataforma humanoide Nao un valor adecuado para la rotación de los vocabularios es 10 grados. Dichos vocabularios rotados, se implementaron con la idea de solventar las ligeras variaciones en la rotación provocadas por la locomoción de estos sistemas robóticos. Los vocabularios rotados representan a las imágenes del mapa topológico como imágenes rotadas. El tercer vocabulario es idéntico al BRIEF explicado con anterioridad. La idea de emplear tres vocabularios es la siguiente: si la imagen de entrada está rotada con respecto a alguna imagen del mapa topológico, entonces dicha imagen es detectada por alguno de los vocabularios rotados; si la imagen de entrada no está rotada con respecto a alguna imagen del mapa que modela el ambiente, entonces se detecta con el vocabulario sin rotación.

En el vocabulario SURF, se emplearon los detectores/descriptores SURF sin modificación alguna, es decir, se emplearon los descriptores originales. Se optó, además, por realizar un vocabulario a partir de los detectores/descriptores ORB sin modificar debido a que son rápidos de obtener e, invariantes a la rotación. Los descriptores ORB son vectores binarios y por lo tanto son rápidos de comparar, lo que los convierte en una opción viable para la localización de robots humanoides. Para crear el vocabulario ORB se utilizaron los descriptores obtenidos a partir del algoritmo original.

El *vocabulario Malla de color de imagen completa* emplea histogramas de color, a partir de parches o ventanas de la imagen, como descriptores. Los mejores resultados se obtuvieron empleando los canales de matiz (H) y saturación (S) del espacio de color

HLS. Para crear el vocabulario, la imagen se dividió en una malla de 10×10 y se calculó el histograma normalizado 2D de cada parche (o ventana). Debido a que no se obtuvieron buenos resultados a partir de la malla anterior, ésta se modificó de tal manera que las ventanas consecutivas se empalmaran —la mitad de su área— tanto en la dirección horizontal como en la dirección vertical, obteniendo así una rejilla de 19×19. El tamaño de cada descriptor se fijó en 64 bits , el cual se sintonizó de forma experimental hasta obtener resultados satisfactorios. El primer elemento del vector descriptor contiene, principalmente, información de la iluminación por lo que se estableció a cero con la finalidad de mitigar los cambios de iluminación.

Debido a que la información contenida en la parte inferior de las imágenes de algunas bases de datos de prueba era similar (suelo), se hizo el vocabulario *Malla de color de media imagen* con la finalidad de comparar sólo la parte superior de las imágenes. De esta forma se descarta la información repetida que resulta irrelevante para efectos de comparación. El vocabulario es creado de la misma forma que el vocabulario *Malla de color de imagen completa*, pero sólo se emplea la mitad superior de la imagen. La ventaja de este método es que la comparación entre imágenes se agiliza debido a que se necesita menos información y, por lo tanto, el tiempo requerido es menor. Cabe mencionar que estos vocabularios se realizaron después de haber ajustado los parámetros del árbol y, elegido la medida de similitud que arrojó mejores resultados.

El vocabulario *Parches aleatorios de color* se realizó al seleccionar 500 parches aleatorios dentro de la imagen. Los parches son rectángulos cuyo tamaño es de 48 x 64 píxeles. Se emplearon como vectores descriptores los histogramas normalizados 2D de los canales H y S. El primer elemento del vector descriptor se estableció a cero con la finalidad de mitigar los efectos de cambios en la iluminación.

3.2. Ajuste de parámetros para la creación de vocabularios visuales

En esta etapa, se procedió a realizar el ajuste de los parámetros en la creación de los vocabularios visuales, con la finalidad de obtener los mejores resultados. Los parámetros que se sintonizaron fueron el número de grupos k_w y los niveles L_w del árbol de vocabulario. Para realizar el ajuste, se seleccionó un conjunto de imágenes tomadas por un robot

humanoide y se usó como memoria visual. Otro conjunto de imágenes se utilizó como entrada para realizar pruebas de localización. Enseguida, se realizaron pruebas variando el número de grupos y los niveles del árbol. Una vez finalizadas las pruebas, se analizaron los resultados obtenidos para seleccionar los parámetros que obtuvieron los mejores resultados de acuerdo al conocimiento dado de la verdadera localización del robot (con ground truth). Además, se determinó la medida de similitud que dio mejores resultados. La base de datos utilizada contenía imágenes con cambios drásticos de iluminación, borrosas, con poca textura entre otros problemas, para probar el algoritmo bajo diferentes condiciones.

Después de ajustar los parámetros del árbol y definir una medida de similitud, se realizaron pruebas para mejorar el rendimiento en los vocabularios de color. En esta etapa, se realizaron modificaciones en los parámetros (definición de las rejillas, etc.) de los vocabularios de malla de color, las cuales se explicaron en la sección 3.1.

3.3. Evaluación de características visuales con ground truth

Ya que se tenían los parámetros que daban los mejores resultados, se ejecutaron pruebas con *ground truth*, es decir, la imagen en la memoria visual que más se parece a la imagen de entrada se conocía de antemano. Las pruebas realizadas se efectuaron con cinco bases de datos diferentes con la finalidad de analizar el comportamiento de los vocabularios. Se realizaron pruebas con dos bases de datos en ambientes exteriores y tres en ambientes interiores. Las bases de datos en exteriores no son para robots humanoides, pero se usaron debido a dos factores: son bases de datos públicas y porque tienen diferentes condiciones ambientales, para analizar el comportamiento de los vocabularios visuales bajo estas condiciones. Algunas imágenes de prueba seleccionadas contenían cambios de rotación o de iluminación o de puntos de vista; también había combinación de esas características y además, con ruido. Dichas imágenes se eligieron con la finalidad de probar los vocabularios bajo distintas condiciones. Para evaluar la salida de la localización, se emplean las fórmulas de confianza 2.10 y 2.11, explicadas en el capítulo 2.

3.4. Evaluación de características visuales sin ground truth

En esta etapa las pruebas se realizaron utilizando las mismas bases de datos que en la etapa anterior. La diferencia con la etapa de ground truth es que en esta ocasión no se conocía la localización por anticipado. Para la evaluación de la salida se utilizaron los niveles de confianza obtenidos por la fórmula 2.9, explicada en el capítulo 2. Los resultados obtenidos se mostraron en una gráfica de frecuencias para encontrar un umbral que permitiera detectar falsos positivos a partir del resultado arrojado por la fórmula de confianza. Mediante el análisis de los resultados experimentales, se observó que con los resultados arrojados por la fórmula 2.9 no era posible determinar un umbral para discriminar los falsos positivos, así que se decidió modificar la fórmula y, además, se propuso una nueva para tratar de encontrar un umbral de decisión.

Capítulo 4

Análisis de resultados

En el presente capítulo se muestran los resultados obtenidos al aplicar la metodología propuesta para la localización. En primer lugar, se muestran las funciones usadas en el código para la creación de los vocabularios visuales. Enseguida, se revisan los resultados obtenidos durante el ajuste de los parámetros del árbol de vocabulario visual. Después, se muestran los resultados arrojados por las pruebas cuando se conoce de antemano la localización del robot. Finalmente, se analizan los resultados de las pruebas efectuadas sin el previo conocimiento de la localización del robot.

4.1. Vocabularios visuales

Para la realización de cada vocabulario, se creó una clase en código C++ utilizando como base las librerías OpenCV 2.4.3 y DBow2 [Gálvez-López and Tardós, 2012]. Tales clases, implementan el algoritmo de bolsa de palabras, es decir, la conversión de imágenes a vectores numéricos; localización y evaluación de la salida del algoritmo con ground truth y sin éste. Las clases se heredaron a partir de una clase "CVoc"genérica para encapsular las funciones que los vocabularios tienen en común (ver Figura 4.1). Las funciones en común son cinco *GenerarVocabulario*, *Cotejar*, *CompareWithoutGroundTruth*, *InsertarParametros* e *IntroducirRuta*. Estas funciones trabajan de manera similar en cada vocabulario, por lo que sólo se explicarán una vez. Además de estas funciones, cada clase define las variables necesarias para su funcionamiento. Enseguida se explican de forma breve cada una de las funciones.

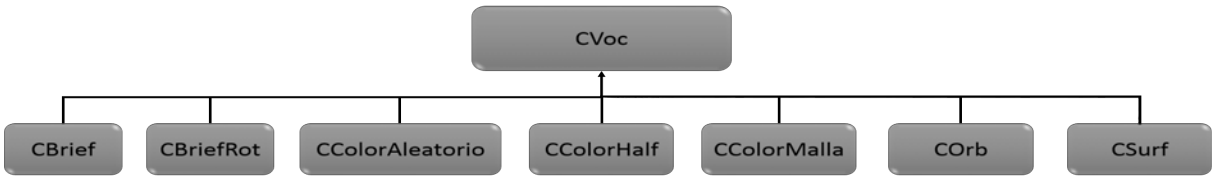


Figura 4.1: Clases del método propuesto para la localización del robot.

1. void **GenerarVocabulario** (string *nombre_archivo_db*). Esta función recibe como parámetro el nombre del archivo que contiene los nombres de las imágenes de la base de datos. Dicha función comienza por generar el vocabulario visual, de modo que se extraen características del conjunto de imágenes de la memoria visual. Cada clase contiene una función auxiliar que extrae los descriptores adecuados para cada vocabulario visual. Enseguida, crea el vocabulario visual a partir de las características extraídas y lo guarda en una base de datos. Finalmente, realiza una comparación de cada imagen de la base de datos con las imágenes de la misma base, con la finalidad de utilizar estos resultados en la evaluación de la salida de la localización (definición de κ_1, por ejemplo).

2. **Cotejar** (string *nombre_archivo_pruebas*). El parámetro de esta función es el nombre de un archivo con el nombre de las imágenes de prueba. Como su nombre lo indica, esta función coteja las imágenes de entrada con las imágenes de la base de datos para localizar el robot con ground truth. Para esto, primero extrae las características de la imagen de prueba, usando la función correspondiente a cada vocabulario. En seguida, realiza la comparación aplicando una medida de similitud. Para finalizar, se calculan los niveles de confianza con las formulas 2.10 y 2.11, a partir de los resultados obtenidos en la etapa de comparación.

3. void **CompareWithoutGroundTruth** (string *nombre_archivo_pruebas*). El parámetro de esta función es el nombre de un archivo que contiene el nombre de las imágenes de prueba. Esta función trabaja de manera similar a la función cotejar sólo que aquí no se emplea el ground truth y la salida del algoritmo se evalúa mediante la fórmula 2.9.

4. void **InsertarParametros** (int *k_in*, int *L_in*, WeightingType *weight_in*, ScoringType *score_in*). Esta función recibe como parámetros el número de grupos k_w para generar el árbol, el número de niveles L_w del árbol, el tipo de ponderación para cada palabra del vocabulario visual (la librería DBow2 implementa por separado las

ponderaciones para las palabras: tf , idf y $tf - idf$) y el tipo de métrica a utilizar en la comparación de imágenes. El objetivo de la función es insertar los parámetros para la creación de vocabularios y la métrica para la comparación de imágenes.

5. void **IntroducirRuta** (string *ruta*). El parámetro de la función es el nombre de un archivo que contiene las rutas para almacenar los resultados obtenidos durante las pruebas de la localización del robot.

Cada clase creada para la localización del robot, se dividió en dos archivos: CVocabulario.h y CVocabulario.cpp. En el primer archivo se declaran las funciones y los atributos de cada clase. Mientras que en el segundo archivo se definen las funciones. De esta forma, se puede trabajar por separado con cada vocabulario visual.

4.2. Ajuste de parámetros

Una vez que se tienen los vocabularios visuales, el siguiente paso es ajustar los parámetros para la creación de los vocabularios, tales como el número de grupos y la cantidad de los niveles del árbol. Dicho ajuste, se lleva a cabo mediante el empleo del ground truth. Específicamente, se usó el nivel de confianza μ_2, ya que toma en cuenta la posible similitud entre las imágenes de la base de datos. Además de los parámetros del árbol, se debe elegir una de las métricas propuestas en la sección 2.3. Tanto los parámetros del árbol como la medida de similitud, se seleccionaron de modo que la mayoría de los niveles de confianza μ_2, obtenidos durante las pruebas, se aproximaran a uno; es decir, los parámetros se seleccionaron con base en la mayor cantidad de resultados correctos obtenidos durante las pruebas.

Después de la obtención de los parámetros, se procedió a trabajar con los vocabularios de malla de color para mejorar su funcionamiento. En estos vocabularios, se hicieron pruebas usando los parches empalmados por la mitad y a tres cuartas partes de su área. Se utilizó el mismo criterio empleado en la selección de los parámetros del árbol: el tipo de empalme que arrojó la mayor cantidad de niveles de confianza μ_2 cercanos a la unidad obtenidos durante las pruebas.

Para realizar la selección de los parámetros del árbol y de la medida de similitud, los resultados se representaron por medio de gráficas de caja. La gráfica de caja más simple se

basa en los siguientes cinco números: valores más pequeños del conjunto de datos; cuarto inferior, es la mediana de la mitad del conjunto de datos más pequeña; la mediana del conjunto de datos; cuarto superior, es la mediana de la mitad de los datos más grande; los valores más grandes del conjunto de datos. Suelen ser muy útiles para comparar dos o más conjuntos de datos compuestos de resultados de la misma variable, en el caso de esta tesis la variable μ_2.

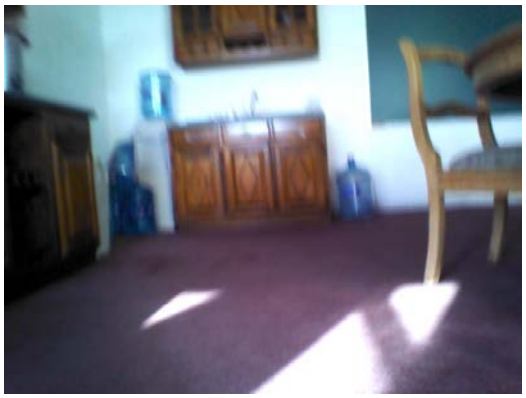
La construcción de la gráfica de caja se lleva a cabo mediante los siguientes pasos. Primero, se traza un rectángulo sobre una escala de medición vertical; el lado inferior del rectángulo está en el cuarto inferior y el lado superior en el cuarto superior del conjunto de datos. Se coloca un segmento de línea horizontal dentro del rectángulo en la ubicación de la mediana; la posición del símbolo de la mediana con respecto a los dos lados del rectángulo da información sobre asimetría en el 50 % medio de los datos. Finalmente, se trazan "bigotes" hacia afuera de ambos extremos del rectángulo dirigidos a los datos más pequeños y más grandes.

Para realizar el ajuste de los parámetros así como para mejorar el desempeño de los vocabularios de malla de color, se utilizó una base de datos que contiene algunas imágenes de buena calidad, otras son borrosas, tienen rotaciones introducidas por la locomoción humanoide y están afectadas por cambios de iluminación, con la intención de trabajar en distintas condiciones. Dicha base de datos, fue tomada por un robot humanoide Nao en movimiento, en el Centro de Investigación en Matemáticas (CIMAT). El robot fue guiado manualmente para visitar la mayor parte de un ambiente que es una sala de trabajo de aproximadamente 9 m $\times$ 9 m. La sala tiene en dos de sus lados grandes ventanas que dan al exterior, las cuales proporcionaron los cambios de iluminación natural con la que fue capturada la secuencia de imágenes. A partir de esta secuencia de imágenes, se eligieron manualmente 187 imágenes para formar una memoria visual. Además, para realizar las pruebas de localización, se usaron 258 imágenes de la misma base de datos empleada para la memoria visual pero diferentes. Las imágenes fueron tomadas a color y su resolución es de 640 $\times$ 480 píxeles. En la Figura 4.2, se muestran algunas imágenes utilizadas para el ajuste de los parámetros.

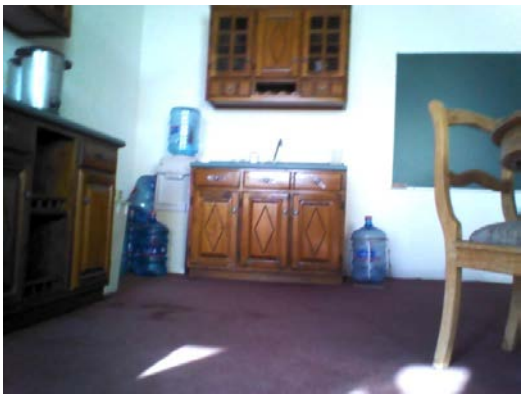
A continuación, se muestran los resultados obtenidos durante el ajuste de los parámetros, la elección de la medida de similitud y, por último, la mejora del desempeño de los vocabularios de malla de color.

(a) Imagen normal

(b) Imagen con cambio de iluminación



(c) Imagen borrosa



(d) Imagen rotada

Figura 4.2: Ejemplo de imágenes de prueba de la base de datos para ajustar los parámetros.

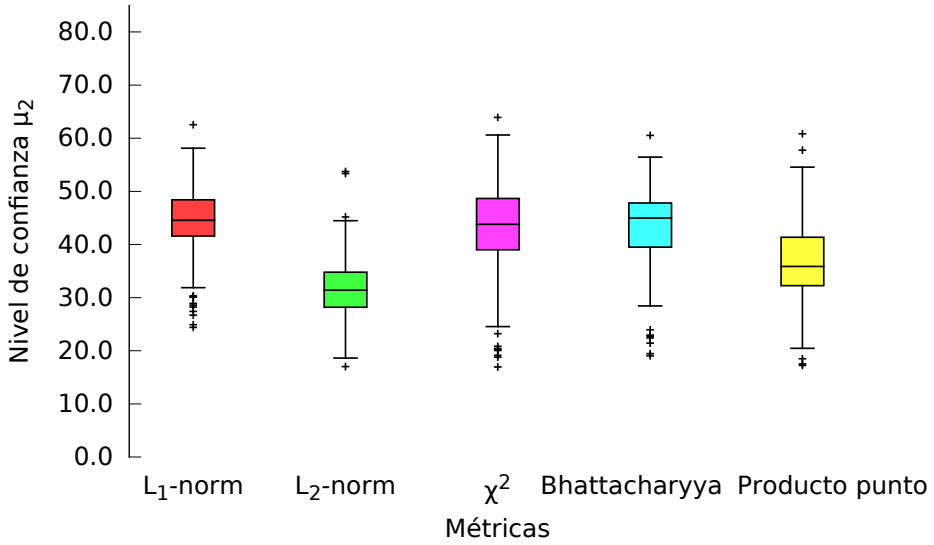
4.2.1. Parámetros del árbol y medida de similitud

En esta sección, se muestra el ajuste de los parámetros del árbol y la elección de la medida de similitud a partir de las 258 pruebas de localización con las 187 imágenes de la memoria visual. Para comenzar, el número de grupos se fijó en $k_w = 8$ y los niveles del árbol se variaron desde $L_w = 4$ hasta $L_w = 8$; se realizaron pruebas combinando el número de centros $k_w = 8$ con estos niveles de profundidad. Se hicieron, además, otro conjunto de pruebas con $k_w = 9$ y $k_w = 10$ de igual manera como se hizo para $k_w = 8$.

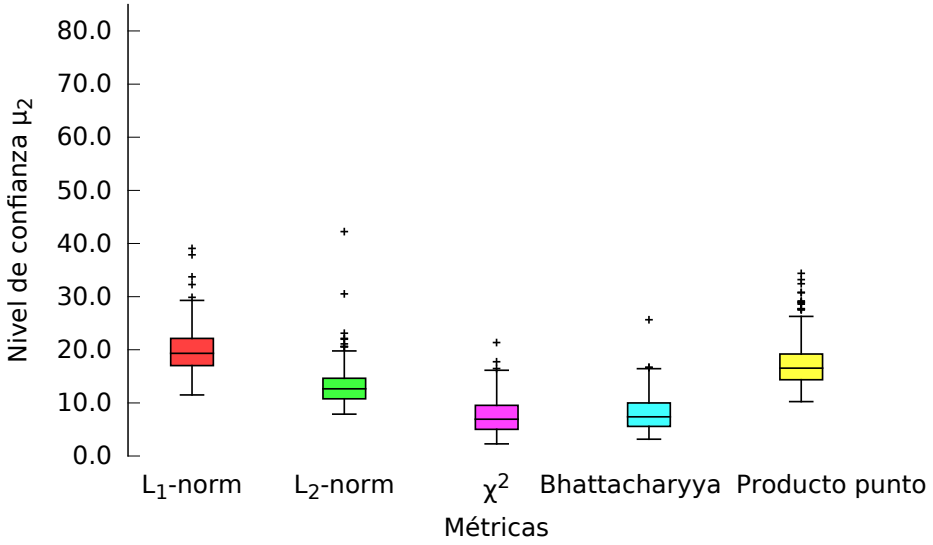
Debido a la cantidad de información obtenida por los diferentes vocabularios creados y, a que se obtuvieron resultados similares, sólo se muestran aquellos obtenidos con el vocabulario BRIEFROT. Además, este es el vocabulario de mayor interés, ya que fue creado en esta tesis específicamente para robots humanoides. Se comenzó por elegir la medida de similitud que ofreció la mayor cantidad de resultados en los cuales el valor de μ_2 se aproximó a uno. Gráficamente, se puede interpretar al elegir los rectángulos de las gráficas de caja con menor área y además, las más próximas al valor de uno. En el caso ideal se observaría una línea horizontal a la altura del valor de uno.

Los resultados obtenidos en las pruebas con los niveles $L_w = 4$ y $L_w = 5$ para $k_w = 8$, se muestran en la Figura 4.3. Como se puede apreciar, dichos niveles de profundidad arrojan resultados alejados del valor $\mu_2 = 1$, esto indica que los resultados obtenidos fueron incorrectos. Resultados similares se obtuvieron para estos niveles con $k_w = 9$ y $k_w = 10$, por lo cual, se decidió desechar los niveles $L_w = 4$ y $L_w = 5$ de la elección de los parámetros.

La elección de la métrica se realizó empleando los niveles $L_w = 6$, $L_w = 7$ y $L_w = 8$. Las gráficas obtenidas con estas combinaciones se muestran en las Figuras 4.4, 4.5 y 4.6. En dichas figuras, se observa que las métricas χ^2 y Bhattacharyya arrojaron los niveles de confianza más cercanos al valor ideal unitario para μ_2. Para las gráficas con $k_w = 8$, la métrica χ^2 obtuvo un mejor comportamiento; en las combinaciones obtenidas con $k_w = 9$, la medida de similitud Bhatacharrya fue la que se comportó de mejor forma; mientras que en las gráficas para $k_w = 10$, la métrica χ^2 obtuvo un comportamiento mejor. El caso $k_w = 10$ y $L_w = 8$ no se llevó a cabo debido a falta de memoria la computadora usada para las pruebas. Estos parámetros generan 10^9 palabras que no fue posible mantenerlas en memoria. Resultados similares se obtuvieron para los demás vocabularios visuales por lo cual, se decidió optar por la métrica χ^2 como la medida de similitud para la localización



(a) $L_w = 4$



(b) $L_w = 5$

Figura 4.3: Diagramas de caja con $k_w = 8$ grupos para 4 y 5 niveles de árbol del vocabulario BRIEFROT.

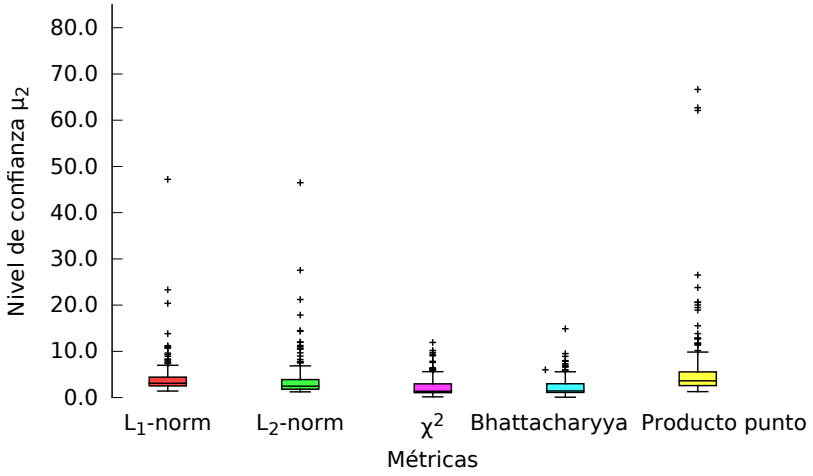
del robot humanoide.

En el caso de la elección de los parámetros del árbol de vocabulario, es decir, la elección de k_w y L_w, se compararon las diferentes combinaciones del número de agrupamientos con los niveles del árbol empleando la métrica χ^2. En la Figura 4.7, se muestran dichos resultados. Al igual que con el resto de los vocabularios, la combinación de $k_w = 8$ con $L_w = 8$ arrojó resultados más cercanos hacia el valor de nivel de confianza ideal, es decir hacia el valor de uno. Por lo tanto, se eligieron $k_w = 8$ y $L_w = 8$ como los parámetros para el árbol de vocabulario.

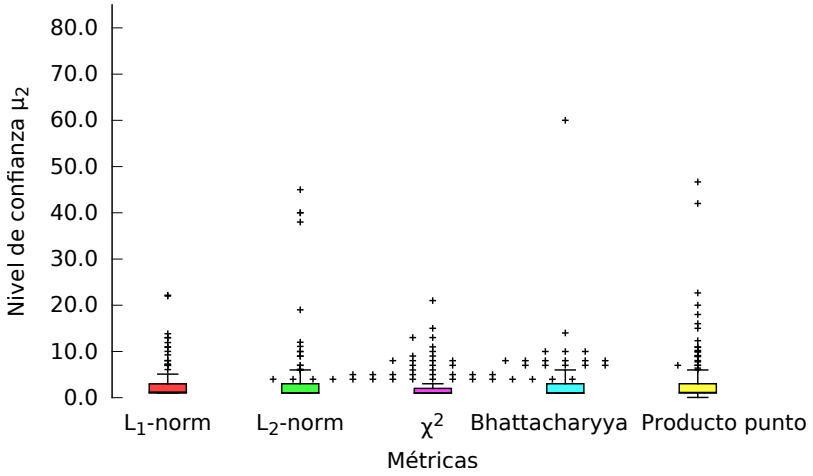
La combinación $k_w = 8$ y $L_w = 6$ arrojó resultados cercanos al nivel de confianza ideal, pero tiene varios resultados que se alejan demasiado de dicho valor (outliers) al igual que para $k_w = 8$ y $L_w = 7$. Los grupos $k_w = 9$ con sus combinaciones $L_w = 6$ y $L_w = 7$ mostraron comportamientos similares a los anteriores: tienen resultados cercanos al nivel de confianza ideal, pero tienen varios outliers muy alejados de la unidad. Para $k_w = 9$ y $L_w = 8$ además de mostrar varios valores atípicos, la caja del diagrama está más extendida lo que quiere decir que la mediana de la segunda mitad de los resultados se aleja del valor ideal. También es el caso para $k_w = 10$ y $L_w = 6$. Mientras que para $k_w = 10$ y $L_w = 7$ tiene varios outliers alejados del nivel de confianza ideal. La combinación $k_w = 8$ y $L_w = 8$, también tiene varios valores atípicos, pero éstos están menos alejados de la unidad, por lo que se decidió usarla como los parámetros del árbol de vocabulario.

4.2.2. Mejora en el desempeño de los vocabularios de malla de color

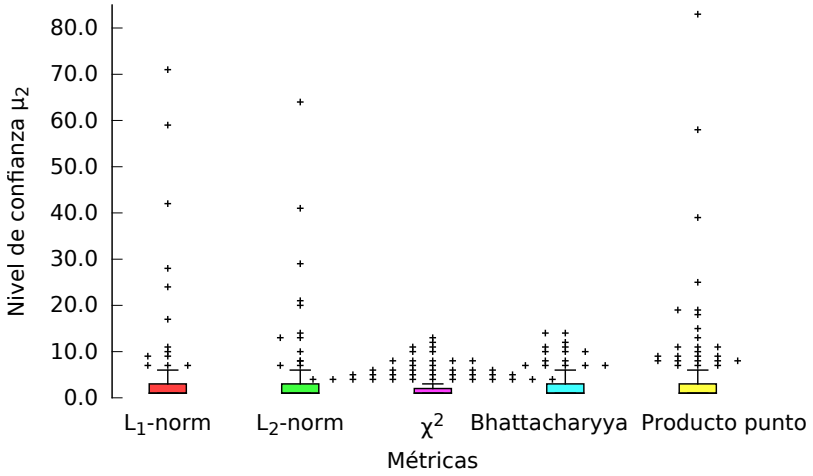
Debido a que inicialmente los vocabularios *Malla de color de imagen completa* y *Malla de color de media imagen* no arrojaban niveles de confianza μ_2 cercanos a uno, se decidió trabajar en mejorar estos vocabularios. Se obtuvieron mejores resultados haciendo el empalme de los parches por la mitad y a tres cuartas partes de su área. En la Figura 4.8 se muestran los resultados obtenidos con empalme y sin empalme. Se puede observar que empalmar la mitad del área de los parches obtuvo valores de μ_2 más próximos a uno; el empalme a tres cuartos del área obtuvo mejores resultados que la malla sin empalme, pero no mejoró el empalme por la mitad del área. Por lo tanto, para los vocabularios de malla de color, se eligió empalmar los parches por la mitad de su área.



(a) $L_w = 6$

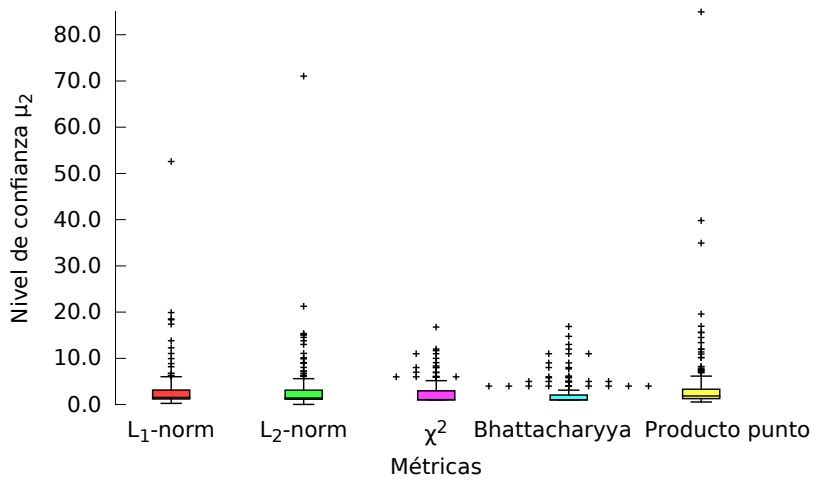


(b) $L_w = 7$

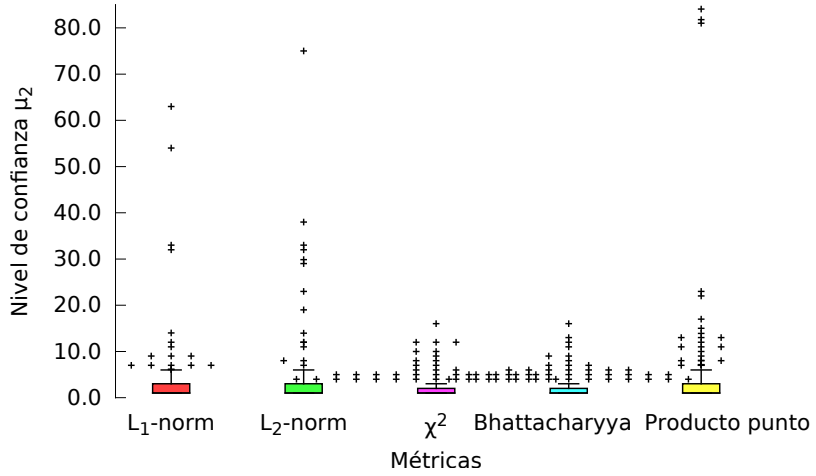


(c) $L_w = 8$

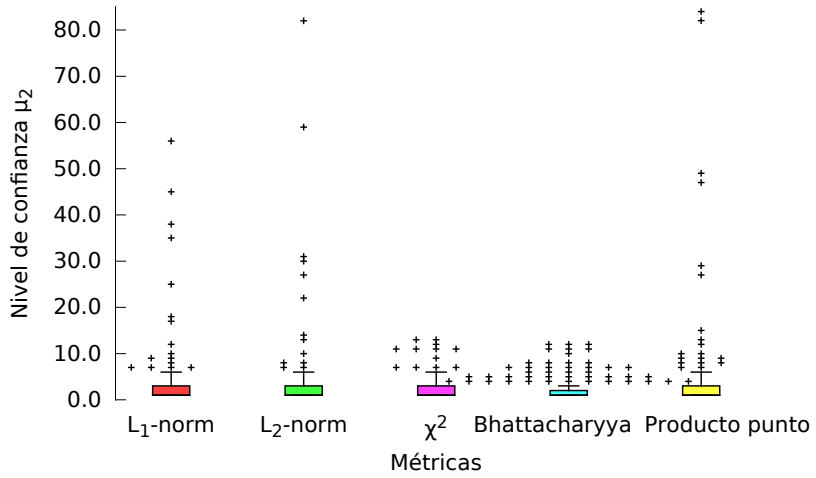
Figura 4.4: Diagramas de caja con $k_w = 8$ grupos para 6, 7 y 8 niveles de árbol del vocabulario BRIEFROT.



(a) $L_w = 6$

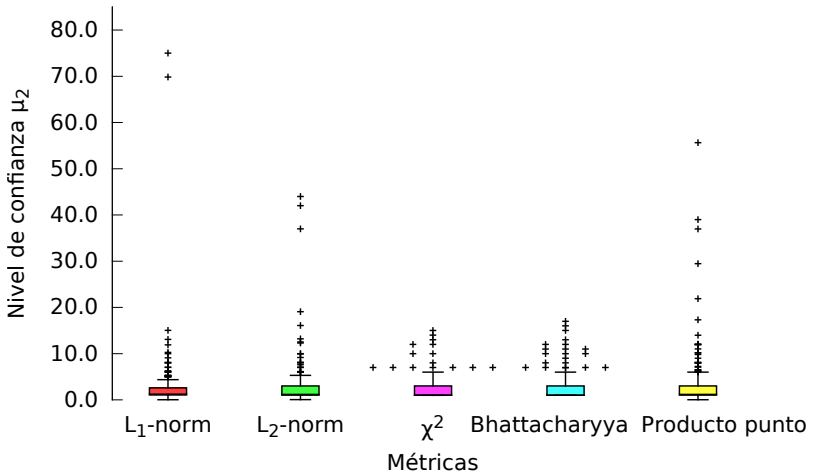


(b) $L_w = 7$

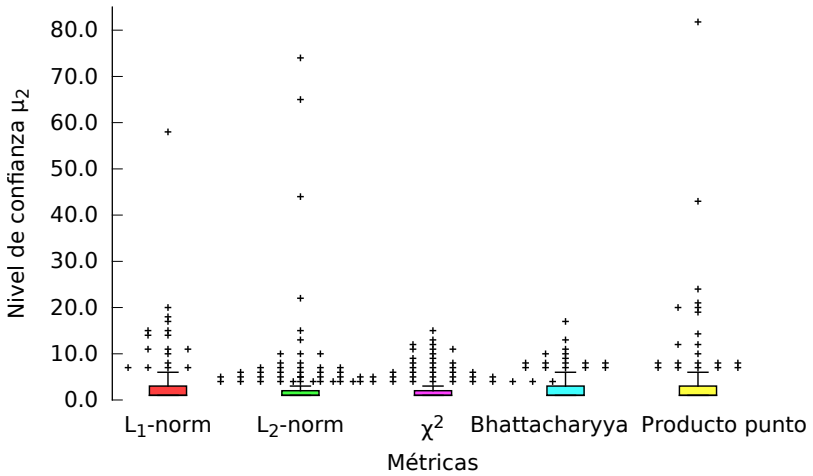


(c) $L_w = 8$

Figura 4.5: Diagramas de caja con $k_w = 9$ grupos para 6, 7 y 8 niveles de árbol del vocabulario BRIEFROT.



(a) $L_w = 6$



(b) $L_w = 7$

Figura 4.6: Diagramas de caja con $k_w = 10$ grupos para 6 y 7 niveles de árbol del vocabulario BRIEFROT.

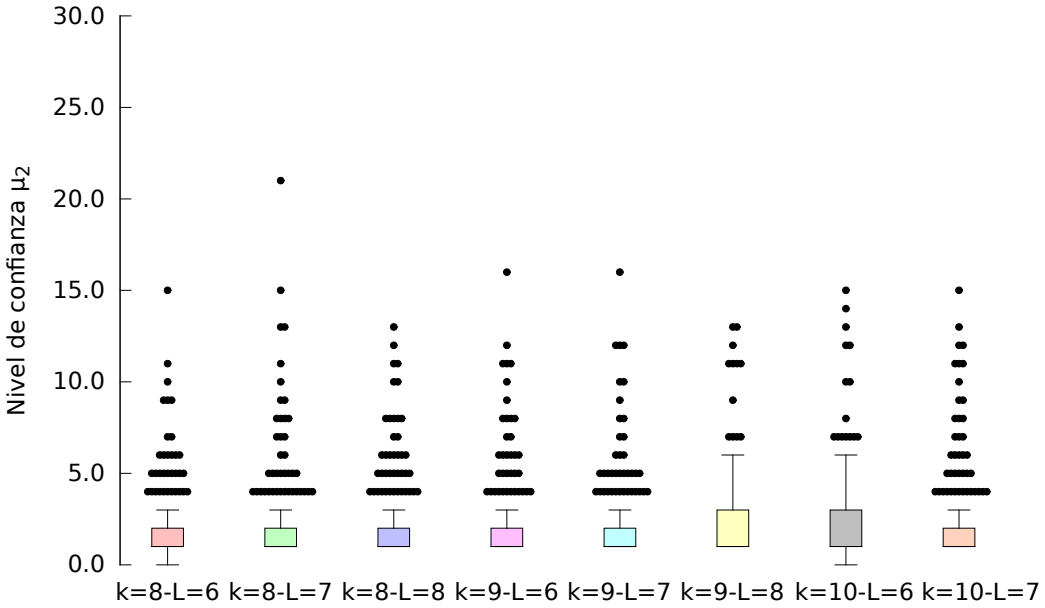


Figura 4.7: Combinaciones de agrupamientos con niveles del árbol para el vocabulario BRIEFROT.

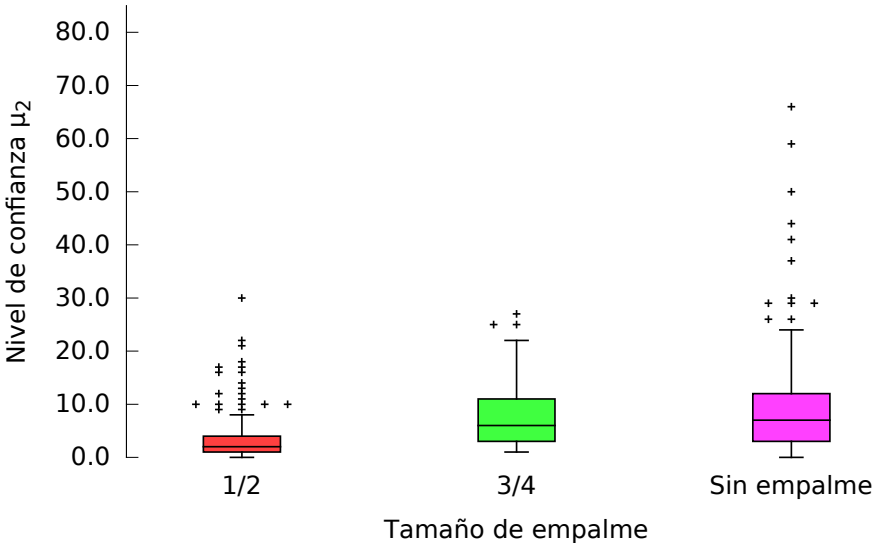


Figura 4.8: Comparación de vocabulario de malla de color con diferentes tamaño de empalme.

4.3. Evaluación de las características visuales con ground truth

Para evaluar las características visuales, es decir, los vocabularios visuales, se efectuaron pruebas en cinco bases de datos distintas, como se muestra en la Tabla 4.1. Se emplearon tres bases de datos en ambientes interiores (CIMAT-NAO-A, CIMAT-NAO-B y Bicocca) y dos en ambientes exteriores (New College y Málaga). A continuación se muestran los resultados obtenidos para cada base de datos. Para la realización de las pruebas, se empleó una laptop usando Ubuntu 12.04 con 4 Gb de memoria RAM y procesador AMD de 1.30 Ghz.

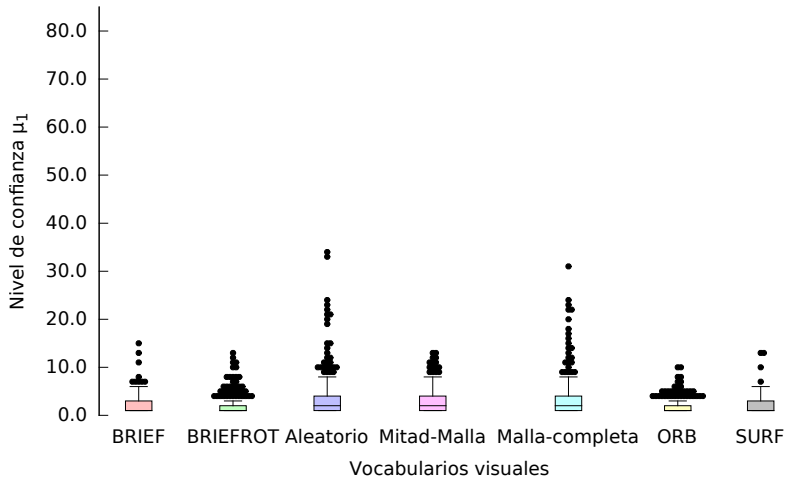
Tabla 4.1: Bases de datos usadas para la evaluación de características.

Base de datos	Descripción	Tamaño de la imagen (px × px)
CIMAT-NAO-A	Interior, dinámica	640 × 480
CIMAT-NAO-B	Interior, dinámica	640 × 480
Bicocca 2009-02-25b	Interiores, estática	320 × 240
New College	Exteriores, dinámica	384 × 512
Málaga	Exteriores, dinámica	1024 × 768

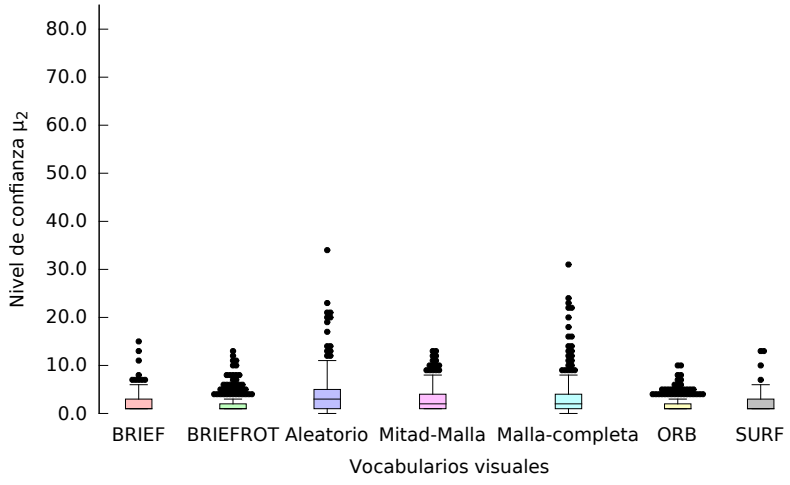
4.3.1. CIMAT-NAO-A

La base de datos *CIMAT-NAO-A* fue tomada en una sala de trabajo dentro del CIMAT mediante un robot humanoide Nao en movimiento. Dicha base de datos, se utilizó para ajustar los parámetros del árbol y para mejorar el desempeño de los vocabularios de malla de color. A partir de esta base de datos, se usaron 187 imágenes como la memoria visual y 258 como imágenes de prueba. Los tiempos reportados fueron medidos desde la etapa de la extracción de características de la imagen de prueba hasta la etapa de comparación de su vector de bolsa de palabras con los vectores de la base de datos.

En la Figura 4.9 se muestran los niveles de confianza μ_1 y μ_2 obtenidos por los vocabularios visuales. μ_1 muestra la posición que la imagen de prueba obtuvo en los resultados. Como se puede apreciar, el vocabulario BRIEFROT consiguió los mejores resultados, en este caso 60.85 % de resultados correctos; le sigue el vocabulario ORB



(a) μ_1



(b) μ_2

Figura 4.9: Niveles de confianza obtenidos por los vocabularios visuales para la base de datos CIMAT-NAO-A.

con 55.81 % resultados acertados. En la Tabla 4.2 se resumen los porcentajes de pruebas correctas logradas por todos los vocabularios y los tiempos promedio de la comparación. Además, se aprecia que el vocabulario *Malla de color de media imagen*, cosechó los peores resultados con un 40.31 % de resultados correctos. Los vocabularios *Malla de color de imagen completa* y *Parches aleatorios de color* obtuvieron la misma cantidad de resultados correctos: 42.64 %.

En la cuestión de los tiempos promedio de comparación, el vocabulario SURF consumió la mayor cantidad de tiempo con 296.5 ms. Seguido por BRIEFROT con 132.4 ms. El vocabulario *Malla de color de media imagen*, ocupó la menor cantidad de tiempo: 93.1 ms. En la Figura 4.10 se muestran algunas imágenes de prueba y su correspondiente imagen de la base de datos más cercana resultado del algoritmo de localización.

Tabla 4.2: Porcentajes de resultados correctos para la base de datos CIMAT-NAO-A.

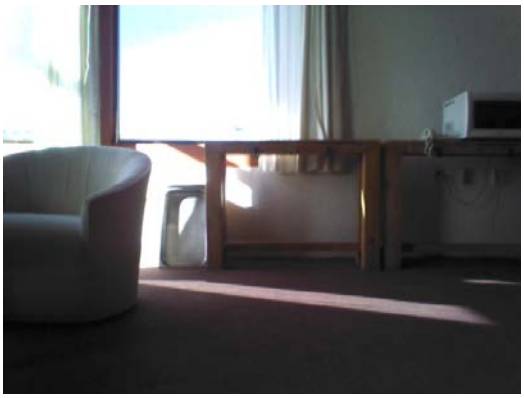
Vocabulario	Número de pruebas	Pruebas correctas	Efectividad (%)	Tiempos promedio de comparación (ms)
BRIEF	258	132	51.16	122.6
BRIEFROT	258	157	60.85	132.4
Aleatorio	258	110	42.64	129.4
Mitad Malla	258	104	40.31	93.1
Malla Completa	258	110	42.64	101.8
ORB	258	144	55.81	107.5
SURF	258	135	52.32	296.5

4.3.2. CIMAT-NAO-B

La base de datos *CIMAT-NAO-B* se obtuvo en el laboratorio de robótica del CIMAT por un robot humanoide Nao en movimiento, por lo que dicha base es de interiores. Se usaron 94 imágenes como memoria visual y 94 como pruebas. Los niveles de confianza se muestran en la Figura 4.11 mediante diagramas de caja y en la Tabla 4.3, se resumen los porcentajes de resultados correctos producidos por cada uno de los vocabularios visuales

(a) Imagen de prueba

(b) Resultado correcto

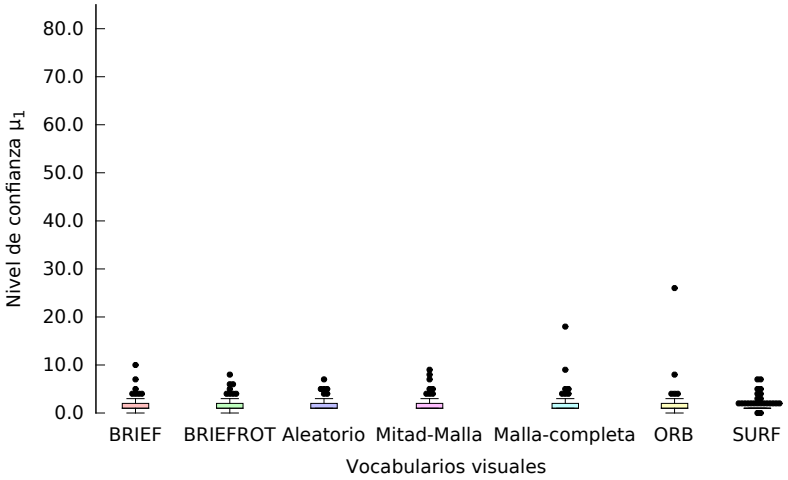


(c) Imagen de prueba

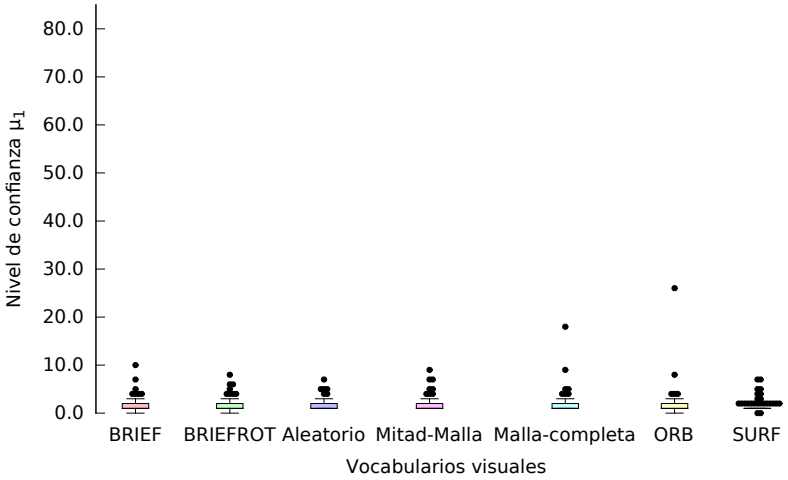
(d) Resultado incorrecto

Figura 4.10: Imágenes de prueba con su correspondiente imagen de la base de datos de acuerdo al vocabulario BRIEFROT.

además, de los tiempos de comparación.



(a) μ_1



(b) μ_2

Figura 4.11: Niveles de confianza obtenidos por los vocabularios visuales para la base de datos CIMAT-NAO-B.

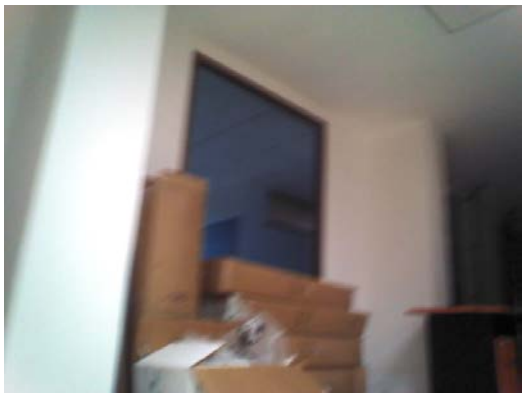
A partir de la figura se observa que, en este caso, el vocabulario SURF alcanzó los mejores resultados, con el 74.46 %; seguido por el vocabulario ORB con 73.40 % de resultados acertados y 267.9 ms empleados para la comparación. Además, el vocabulario de *Parches aleatorios de color* cosechó la menor cantidad de resultados correctos con 65.96 % y con un tiempo de 109.9 ms. El vocabulario BRIEFROT logró un 69.14 % de efectividad quedando en la cuarta posición y utilizó 120 ms de tiempo durante las pruebas. Con base

en los resultados anteriores, todos los vocabularios obtuvieron mejores resultados con esta base de datos debido a que los cambios de iluminación son menos drásticos que en CIMAT-NAO-A, ya que en el laboratorio de robótica no hay ventanas al exterior que permitan el paso de la luz natural.

Tabla 4.3: Porcentajes de resultados correctos para la base de datos CIMAT-NAO-B.

Vocabulario	Número de pruebas	Pruebas correctas	Efectividad (%)	Tiempos promedio de comparación (ms)
BRIEF	94	63	67.02	87.73
BRIEFROT	94	65	69.14	112.0
Aleatorio	94	62	65.96	109.9
Mitad Malla	94	64	68.09	63.2
Malla Completa	94	68	72.34	73.1
ORB	94	69	73.40	77.7
SURF	94	70	74.46	267.9

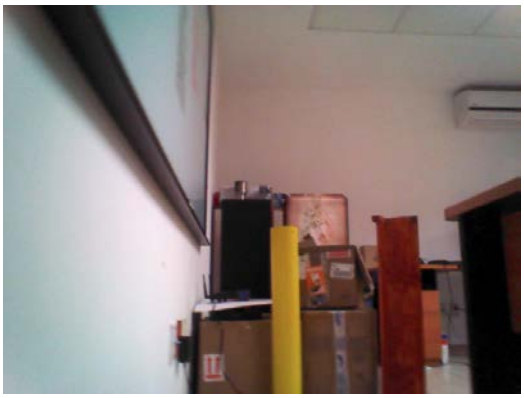
Se realizaron 21 pruebas adicionales con diferentes cambios de iluminación y se modificó la posición de algunos objetos para generar algunas obstrucciones, con la finalidad de probar los vocabularios bajo distintas circunstancias. La Tabla 4.4 muestra los resultados obtenidos. En este caso, el vocabulario SURF obtuvo los mejores resultados, con 11 resultados correctos de 21; le sigue el ORB con 10 de 21 resultados correctos. El vocabulario de *Parches aleatorios de color* fue el que peor comportamiento alcanzó con un total de 4 resultados correctos de 21 posibles. En la Figura 4.12, se muestran algunas imágenes de prueba con su correspondiente imagen de la base de datos que más se parece; los resultados corresponden al vocabulario ORB. En dicha figura, se observa que el vocabulario ORB es capaz de reconocer el lugar aunque haya cambios de rotación y oclusiones.



(a) Imagen de prueba

(b) Resultado correcto

(c) Imagen de prueba



(d) Resultado correcto

Figura 4.12: Imágenes de prueba con su correspondiente imagen de la base de datos de acuerdo con el vocabulario ORB.

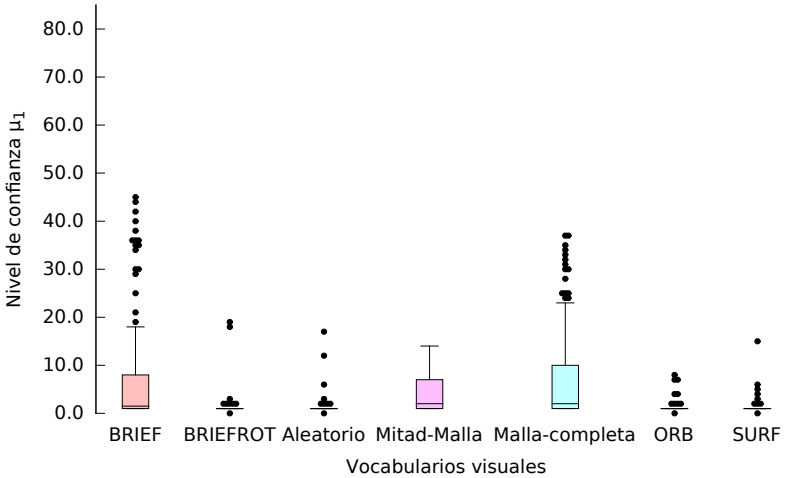
Tabla 4.4: Porcentajes de resultados correctos para la base de datos CIMAT-NAO-B con cambio de condiciones.

Vocabulario	Número de pruebas	Pruebas correctas	Efectividad (%)	Tiempos promedio de comparación (ms)
BRIEF	21	5	23.81	83.9
BRIEFROT	21	8	38.1	104.2
Aleatorio	21	4	19.05	108.5
Mitad Malla	21	5	23.81	60.1
Malla Completa	21	5	23.81	71.7
ORB	21	10	47.62	74.8
SURF	21	11	52.38	234.5

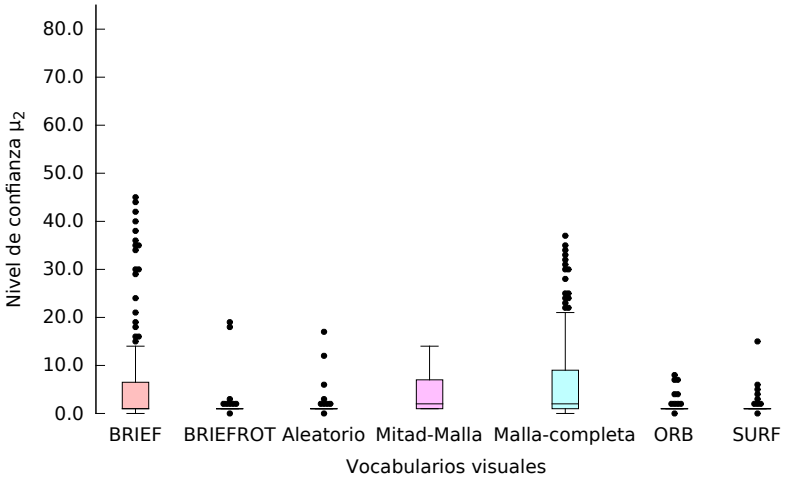
4.3.3. Bicocca 2009-02-25b

La base de datos interior, Bicocca25b [Bonarini et al., 2006], tiene 120 imágenes tanto para la memoria visual como para las pruebas. Fue tomada dentro del edificio de una universidad por una plataforma robótica sobre ruedas, por lo que no se tienen cambios drásticos de rotación. Esta base de datos, se eligió para probar los vocabularios bajo distintas condiciones.

En la Figura 4.13, se muestran los diagramas de caja para los niveles de confianza y en la Tabla 4.5, se resumen los porcentajes de pruebas correctas y los tiempos de ejecución de las pruebas. Como se aprecia, los vocabularios BRIEFROT, *Parches aleatorios de color* y SURF obtuvieron los mejores resultados con un total de 111 sobre 120, seguidos muy de cerca por el vocabulario ORB con 110 resultados correctos de 120 posibles. El vocabulario *Malla de color de media imagen* alcanzó los peores resultados con un total de 57 aciertos de 120 posibles. El vocabulario SURF empleó la mayor cantidad de tiempo para la comparación: 120 ms. El vocabulario BRIEFROT fue el segundo en consumir la mayor cantidad de tiempo de comparación con 98.5 ms. Mientras que el vocabulario *Malla de color de imagen completa* requirió la menor cantidad de tiempo con 36 ms.



(a) μ_1



(b) μ_2

Figura 4.13: Niveles de confianza obtenidos por los vocabularios visuales para la base de datos Bicocca25b.

Tabla 4.5: Porcentajes de resultados correctos para la base de datos Bicocca25b.

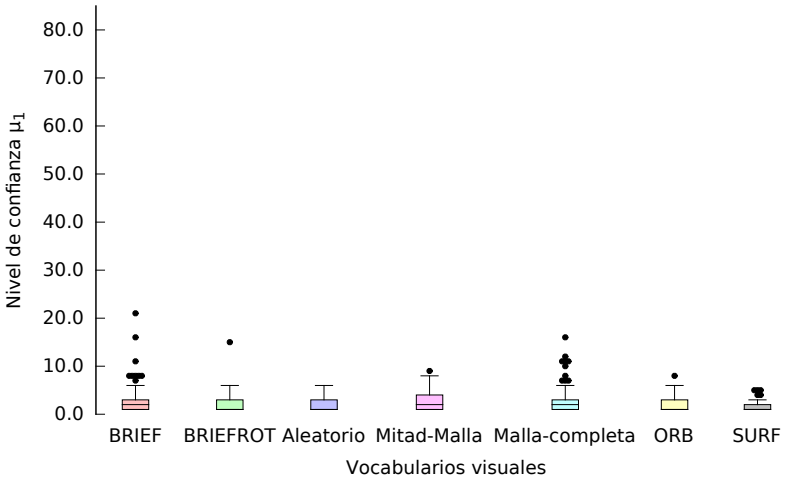
Vocabulario	Número de pruebas	Pruebas correctas	Efectividad (%)	Tiempos promedio de comparación (ms)
BRIEF	120	60	50	73.4
BRIEFROT	120	111	92.5	98.5
Aleatorio	120	111	92.5	72.7
Mitad Malla	120	57	47.5	36.0
Malla Completa	120	59	49.17	79.4
ORB	120	110	91.67	60.2
SURF	120	111	92.5	120.0

4.3.4. New College

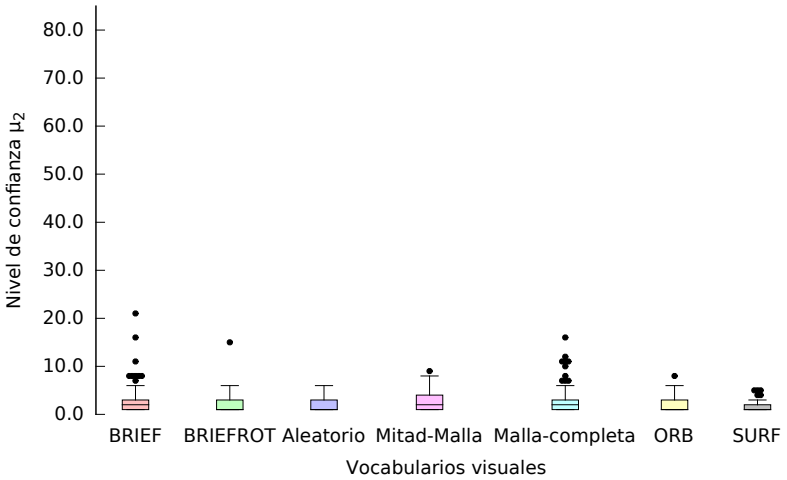
La base de datos exterior, New College [Smith et al., 2009], fue tomada al exterior de la Universidad de Oxford y se usaron 122 imágenes como memoria visual, además de 117 imágenes como prueba. Se eligió esta base de datos, para probar los vocabularios visuales en exteriores. En la Figura 4.14, se pueden observar los resultados obtenidos con esta base de datos y en la Tabla 4.6, se aprecian los porcentajes de los resultados correctos además de los tiempos promedio de comparación. El vocabulario SURF obtuvo los mejores resultados con un 63.25 % de efectividad, seguido por BRIEFROT con un 59.83 % de eficiencia. El vocabulario *Malla de color de media imagen* alcanzó los peores resultados con 29.06 % de efectividad y también consumió la menor cantidad de tiempo de comparación con 60.4 ms. El vocabulario SURF requirió la mayor cantidad de tiempo promedio de comparación: 302.3 ms.

4.3.5. Málaga

La base de datos exterior, Málaga [Blanco et al., 2009], fue tomada en el estacionamiento de la Escuela de Ciencias de la Computación de la Universidad de Málaga. Se usaron 104 imágenes tanto para la memoria visual como para las pruebas. Se eligió esta base de datos, ya que es un ambiente exterior y el tamaño de las imágenes es mayor que las



(a) μ_1



(b) μ_2

Figura 4.14: Niveles de confianza obtenidos por los vocabularios visuales para la base de datos New College.

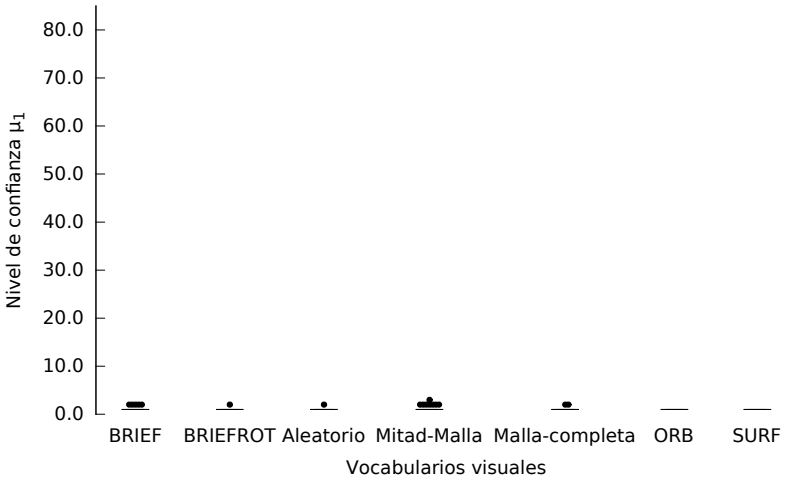
Tabla 4.6: Porcentajes de resultados correctos para la base de datos New College.

Vocabulario	Número de pruebas	Pruebas correctas	Efectividad (%)	Tiempos promedio de comparación (ms)
BRIEF	117	68	58.12	105.9
BRIEFROT	117	70	59.83	134.6
Aleatorio	117	48	41.02	110.9
Mitad Malla	117	34	29.06	60.4
Malla Completa	117	44	37.61	110.5
ORB	117	69	58.97	107.0
SURF	117	74	63.25	302.3

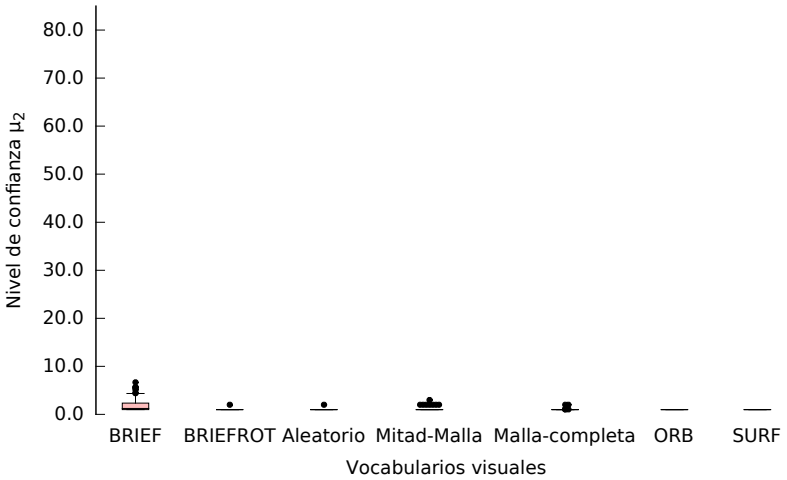
bases de datos anteriores. En la Figura 4.15, se pueden observar los resultados obtenidos con esta base de datos y en la Tabla 4.7, se aprecian los porcentajes de los resultados correctos. En este caso, todos los vocabularios obtuvieron buenas eficiencias. Los vocabularios SURF y ORB lograron los mejores resultados con un 100 % de efectividad, seguido por BRIEFROT y *Parches aleatorios de color* con un 99.04 % de eficiencia. El vocabulario *Malla de color de media imagen* alcanzó los peores resultados con 92.31 % de efectividad y requirió de 116.6 ms para la comparación.

El vocabulario SURF empleó la mayor cantidad de tiempo de comparación: 1348.2 ms. Mientras que el vocabulario *Malla de color de media imagen* requirió el menor tiempo con 116.6 ms. Como las imágenes de esta base son de mayor resolución que las imágenes de las bases de datos anteriores, el tiempo requerido para la extracción de las características es mayor y por lo tanto, el tiempo de comparación se incrementa.

De los resultados anteriores, se puede apreciar que los resultados son mejores para condiciones mas estables de iluminación como pasa en CIMAT-NAO-B (donde no hay ventanas al exterior) en comparación con CIMAT-NAO-A. También, como era de esperarse, se observa que es más difícil la localización de robots humanoides que de robots con ruedas. Además, se observa la dificultad de la localización en bases de datos de exteriores para imágenes de baja resolución, sin embargo, esto se puede solventar con imágenes de buena resolución, aunque el tiempo de cómputo requerido para la comparación



(a) μ_1



(b) μ_2

Figura 4.15: Niveles de confianza obtenidos por los vocabularios visuales para la base de datos Málaga.

Tabla 4.7: Porcentajes de resultados correctos para la base de datos Málaga.

Vocabulario	Número de pruebas	Pruebas correctas	Efectividad (%)	Tiempos promedio de comparación (ms)
BRIEF	104	99	95.19	212.8
BRIEFROT	104	103	99.04	225.6
Aleatorio	104	103	99.04	212.7
Mitad Malla	104	96	92.31	116.6
Malla Completa	104	102	98.08	130.1
ORB	104	104	100	225.3
SURF	104	104	100	1348.2

se eleva.

4.4. Evaluación de las características visuales sin ground truth

En esta sección, se muestran los resultados obtenidos durante las pruebas sin emplear el ground truth, es decir, la posición del robot no se conocía de antemano. Se requería evaluar la salida de los vocabularios por medio del nivel de confianza κ_1, explicado en la sección 2.4.1. A partir de estos resultados, se propuso un umbral para el valor de κ_1 que asegure la correcta localización del robot sin el conocimiento previo de la misma. Para encontrar el umbral, se realizaron gráficas de histogramas de frecuencias para la κ_1 de los resultados correctos e incorrectos. Los resultados correctos e incorrectos se obtuvieron empleando el ground truth, pero sólo se hizo de esta manera para encontrar el umbral y una vez encontrado, ya no se usa el ground truth. Lo ideal sería que las gráficas obtenidas a partir de resultados correctos e incorrectos estuvieran separadas para seleccionar un umbral que indique una localización exitosa.

En la Figura 4.16, se muestra una gráfica de histograma para la confianza κ_1 de pruebas correctas e incorrectas para la base de datos CIMAT-NAO-A. En dicha figura, se

aprecian las gráficas sobrepuestas de los resultados correctos e incorrectos, y no se observa un umbral que indique la localización correcta.

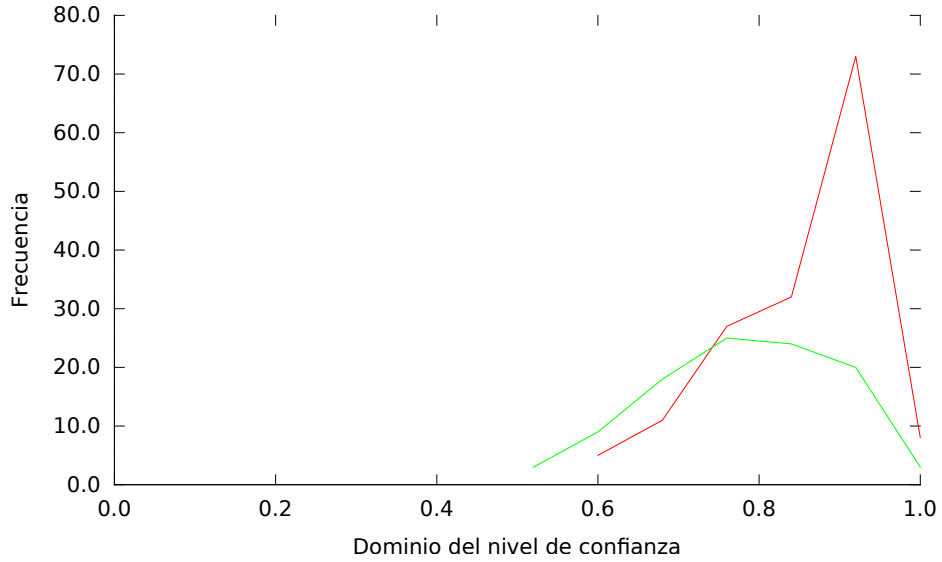


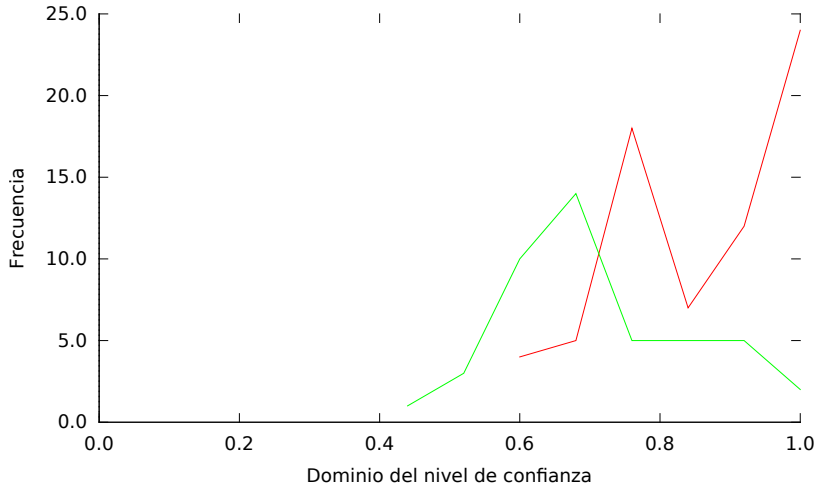
Figura 4.16: κ_1 para la base de datos CIMAT-NAO-A. La línea roja muestra los resultados correctos.

La base de datos CIMAT-NAO-B (ver Figura 4.17) muestra a las gráficas de κ_1 empalmadas y aún no se puede determinar algún umbral que permita identificar las imágenes correctamente. A partir de estos resultados, se planteó un nuevo nivel de confianza como:

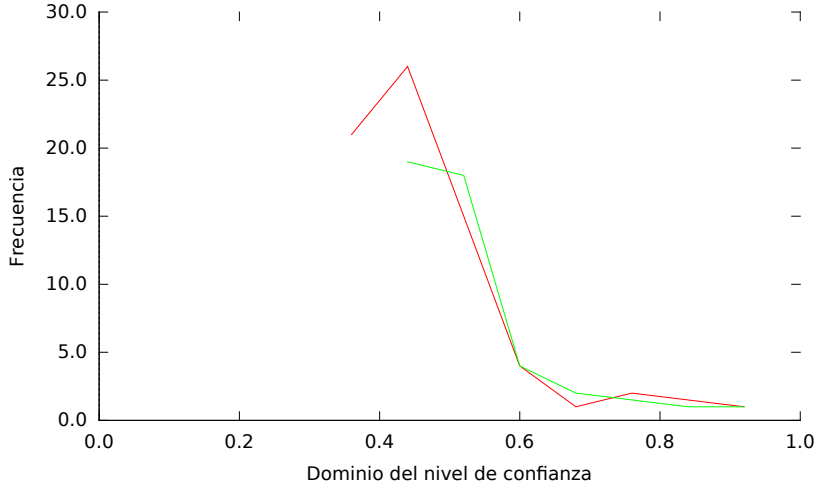
$$\kappa_2 = \frac{\epsilon}{\epsilon + \sum_l s_l}, \tag{4.1}$$

donde ϵ es una constante propuesta manualmente durante las pruebas y se fijó a un valor de 0.55; $\sum_l s_l$ es la sumatoria, a partir de la segunda posición, de los resultados de la prueba. La idea de esta fórmula es que cuando la sumatoria sea igual a cero, se indique una correspondencia perfecta, es decir igual a uno. Cuando el valor de la sumatoria sea diferente de cero: si el valor de κ_2 es menor que ϵ, el resultado es incorrecto; si el valor de κ_2 es mayor que ϵ, la detección es correcta. La Figura 4.17 no indica una mejoría (ni para κ_1 ni para κ_2) para la selección del umbral buscado.

Por otra parte, la base de datos Bicocca25b, proporciona las gráficas de la Figura 4.18 y muestran mejores resultados de κ_1 para la selección de un umbral de localización que las bases de datos anteriores. Se puede inferir un primer umbral con un valor de 0.85,



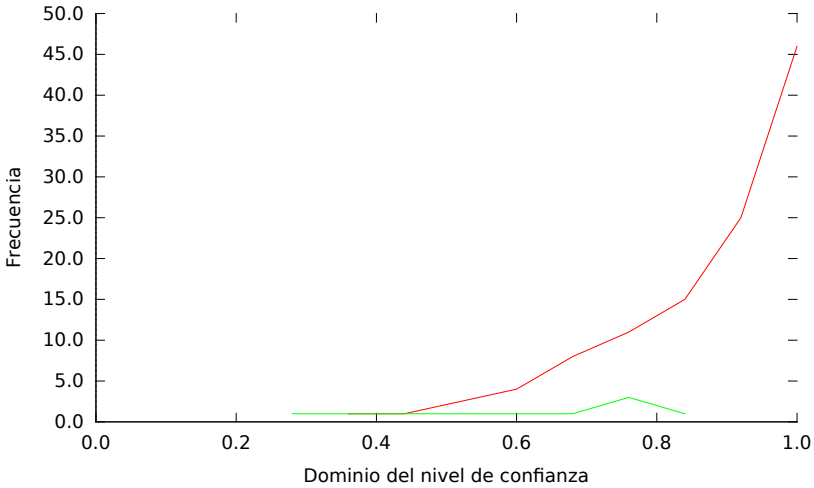
(a) κ_1



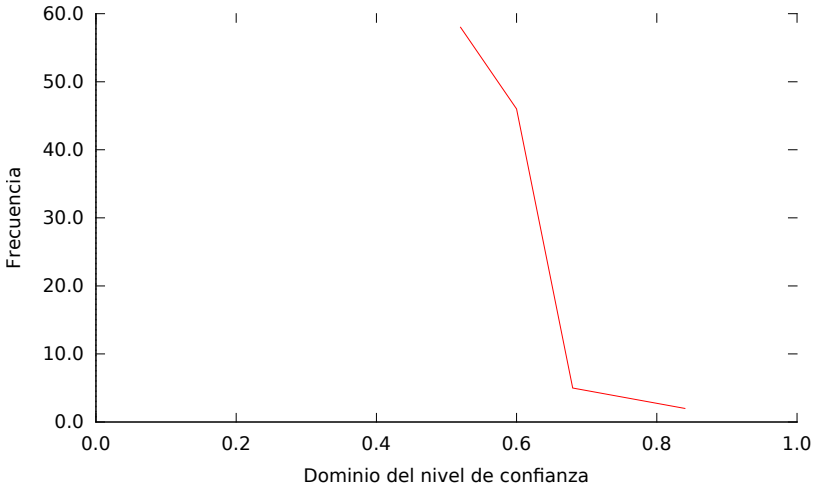
(b) κ_2

Figura 4.17: Niveles de confianza κ_1 y κ_2 para la base de datos CIMAT-NAO-B. La línea roja muestra los resultados correctos.

pero dicho valor es rebasado por algunos niveles de confianza de pruebas incorrectas de las bases de datos anteriores. Por lo cual, este no es un valor final, es sólo una primera aproximación. El valor de κ_2 no se puede inferir de esta gráfica.



(a) κ_1

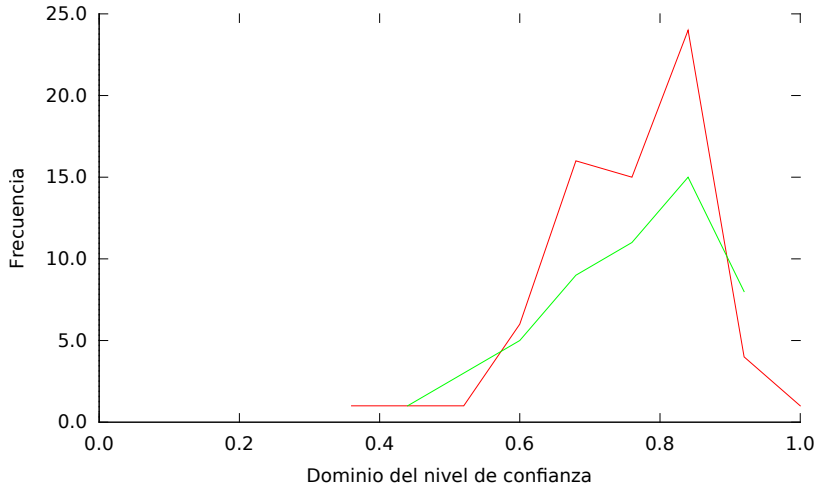


(b) κ_2

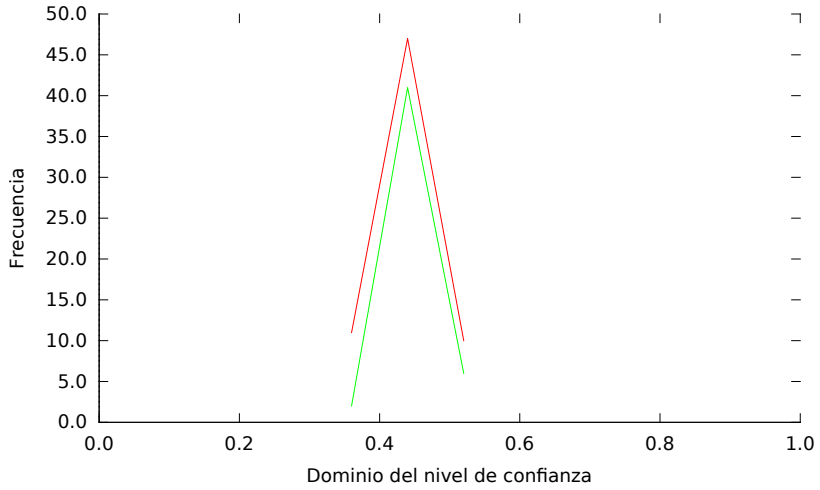
Figura 4.18: Niveles de confianza κ_1 y κ_2 para la base de datos Bicocca25b. La línea roja muestra los resultados correctos.

La base de datos New College sugiere que el umbral se puede desplazar hacia un valor de 0.94, como se ve en la Figura 4.19. Dicho valor es más conservador que el propuesto por los resultados con la base de datos Bicocca25b, por lo cual, se deja como una nueva aproximación. Nuevamente, no se puede inferir un valor adecuado para κ_2 debido a la

superposición de las gráficas.



(a) κ_1

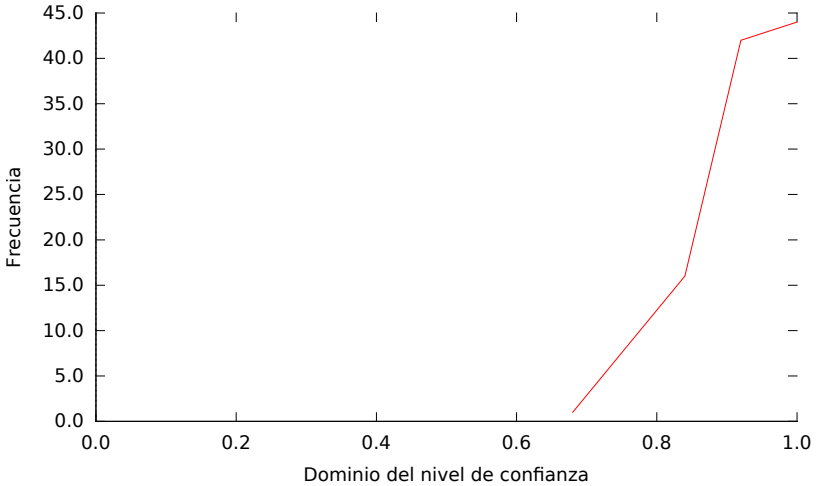


(b) κ_2

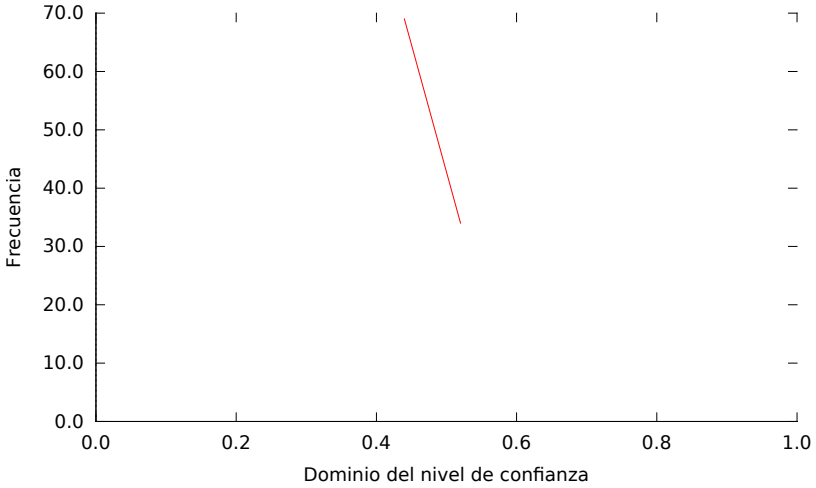
Figura 4.19: Niveles de confianza κ_1 y κ_2 para la base de datos New College. La línea roja muestra los resultados correctos.

Finalmente, la base de datos Málaga, no arroja valores tanto para κ_1 como para κ_2 que puedan dar un valor de umbral para la detección correcta de una imagen de entrada, por lo que no se tomó en cuenta para la decisión final. En la Figura 4.20 se aprecian los resultados para esta base de datos.

Con base en los resultados anteriores, se sugiere optar por un umbral con valor de



(a) κ_1



(b) κ_2

Figura 4.20: Niveles de confianza κ_1 y κ_2 para la base de datos Málaga. La línea roja muestra los resultados correctos.

0.95. Este valor es un poco conservador pero se asegura la correcta localización. Cabe mencionar que algunos niveles de confianza de las bases de datos del CIMAT, rebasan este valor, por lo que para estos casos, la prueba se debería de repetir para tener la certeza de una correcta localización. También, se puede apreciar que la sugerencia de un nuevo valor de confianza no ayudó en mucho a la selección del umbral. Como trabajo futuro se pueden proponer otras expresiones para calcular el nivel de confianza de tal forma que pueda determinarse con mayor claridad un umbral de decisión de si la localización es correcta o no.

Capítulo 5

Conclusiones y Trabajo Futuro

Conclusiones

Aplicando la metodología para la localización del robot y llevando a cabo los ajustes de parámetros para la creación de los vocabularios y la localización del robot, se obtuvieron resultados favorables. Se presentan a continuación las conclusiones logradas del procedimiento experimental propuesto y el análisis de resultados.

Se crearon siete vocabularios visuales para la localización del robot. Cuatro de los vocabularios, tienen descriptores escalares mientras que el resto usa vectores descriptores binarios. Tres de los vocabularios escalares, están basados en color, específicamente, se utilizaron los canales de Matiz (H) y Saturación (S) del espacio de color HLS. El canal L no se puso con la finalidad de proporcionar robustez ante cambios de iluminación.

Un vocabulario se diseñó para su uso especifico en robots humanoides: BRIEFROT. Dicho vocabulario, se creó a partir de otro empleado en robots tipo carro, en los cuales no existen variaciones drásticas de rotación. El vocabulario BRIEFROT es robusto a pequeños cambios de rotación que se dan en los movimientos naturales de las plataformas robóticas humanoides.

Los parámetros elegidos para el árbol fueron 8 grupos y 8 niveles de profundidad. Esta cantidad de grupos y niveles se eligió debido a que arrojaron la mayor de resultados cercanos al valor ideal y además, los resultados atípicos no se alejaron demasiado de dicho nivel. La medida de similitud χ^2 fue la que, en unión con los parámetros del árbol, obtuvo

los mejores resultados. Esto pudo ser debido a que esta métrica usa el cuadrado de la diferencia de los elementos de los vectores. Para los vocabularios de Malla de color, se usaron ventanas empalmadas por la mitad de su área, ya que fueron los que arrojaron la mayor cantidad de resultados correctos.

En la evaluación de características empleando el ground truth, se usaron cinco bases de datos para probar el algoritmo de localización. En la base de datos CIMAT-NAO-A, el vocabulario BRIEFROT obtuvo un 60.85 % de efectividad y requirió 132.4 ms para la comparación, siendo el mejor vocabulario en este caso. Mientras que el vocabulario *Malla de color de media imagen* fue el último lugar con 40.31 % de aciertos y con un tiempo de comparación de 93.1 ms. El vocabulario SURF se comportó de la mejor manera en la base de datos CIMAT-NAO-B con el 74.46 % de resultados correctos, pero fue el que requirió mayor tiempo de ejecución: 267.9 ms. El vocabulario *Parches aleatorios de color* logró el peor desempeño con un 65.96 % y necesito del menor tiempo con 109.9 ms. Para la base Biococca25b, BRIEFROT junto con los vocabularios SURF y *Parches aleatorios de color* alcanzaron la mejor puntuación 92.5 %. El vocabulario SURF fue el mejor en la base de datos New Collwge con el 63.25 % de eficiencia y 302.3 ms de tiempo de ejecución, seguido por BRIEFROT con 59.83 %. ORB y SURF produjeron las máximas puntuaciones para la base de datos Málaga con el 100 % de efectividad, seguidas muy de cerca por BRIEFROT con el 99.04 % de aciertos. El vocabulario SURF necesito la mayor cantidad de tiempo para la comparación con 1348.2 ms. En estas últimas tres bases de datos el vocabulario *Malla de color de media imagen* obtuvo el peor desempeño con 47.5 %, 29.06 % y 92.31 % respectivamente.

El vocabulario BRIEFROT cosechó buenos resultados en las bases de datos utilizadas. En tres de ellas, fue el que mejor resultados arrojó mientras que en las dos restantes, alcanzó la segunda posición en cuanto a mejores resultados logrados se refiere. Esto permite que el vocabulario BRIEFROT sea una buena opción para la localización de robots humanoides. En cambio, el vocabulario *Malla de color de media imagen* produjo los peores resultados en cuatro de las cinco bases de datos empleadas y fue la segunda peor ubicada en la restante base de datos. Esto se debe a que dicho vocabulario sólo usa la mitad de la información de las imágenes.

Las bases de datos obtenidas por robots móviles tipo carro obtuvieron mejores resultados que las adquiridas por robots humanoides. Esto de debe a que las imágenes obtenidas por los robots móviles no tienen tanto ruido y rotación como las de los robots

humanoides. Además, las bases de datos que tienen mayor resolución también arrojaron buenos resultados, pero el tiempo requerido para la comparación fue mayor. Esto es debido a que el tamaño de las imágenes es mayor y por lo tanto, se requiere mayor cantidad de tiempo para la extracción de las características de las imágenes.

En la localización sin el ground truth, se obtuvo un umbral para decidir si la salida del algoritmo es correcto. Dicho umbral, se alcanzó a partir de las pruebas realizadas con las bases de datos mencionadas anteriormente. Específicamente del valor de κ_1, ya que la nueva κ_2 propuesta no mostró ser una buena opción para la elección del umbral. El valor del umbral es conservador debido a que se debe asegurar la correcta localización. Algunos valores de κ_1 rebasan dicho umbral por lo cual, cuando una salida obtenga un valor mayor a este umbral, la prueba se debe repetir para garantizar que la salida del algoritmo sea la correcta. La prueba se puede repetir con una nueva imagen de entrada desde una localización cercana y calcular la κ_2 para esta posición y hacer lo mismo para otra posición cercana. Si el valor de esta κ_1 es semejante en las repeticiones y además mayor que el umbral, la prueba puede ser evaluada como válida, con esto se aseguraría que la salida del algoritmo es confiable.

Trabajo futuro

Como primera mejora al algoritmo de localización, se propone la combinación de los vocabularios visuales con el fin de tener una mejor localización del robot. La idea es que cuándo se tenga una imagen de entrada, se realice la comparación con dos o tres vocabularios y decidir, a partir de los resultados, cual vocabulario logra la localización correcta.

Otra mejora es la obtención de un nivel de confianza para el caso donde no se conoce la localización previa del robot. Definir una κ que permita obtener un umbral más adecuado para decidir si la localización es acertada.

El algoritmo de localización puede usarse para detectar cruce de caminos visuales y mejorar la construcción inicial de la memoria visual mediante la unión de nodos en el grafo que codifica a la memoria visual, los cuales representan la misma localización.

Referencias

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proc. of ACM-SIAM Symposium on Discrete algorithms*, pages 1027–1035.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM press, New York.

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

[Blanco et al., 2009] Blanco, J.-L., Moreno, F.-A., and González, J. (2009). A collection of outdoor robotic datasets with centimeter-accuracy ground truth. *Autonomous Robots*, 27(4):327–351.

[Bonarini et al., 2006] Bonarini, A., Burgard, W., Fontana, G., Matteucci, M., Sorrenti, D. G., and Tardos, J. D. (2006). Rawseeds: Robotics advancement through web-publishing of sensorial and elaborated extensive data sets. In *Proc.of IROS*, volume 6.

[Botterill, 2010] Botterill, T. (2010). *Visual Navigation for Mobile Robots Using the Bag-of-Words Algorithm: A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy, University of Canterbury, Christchurch, New Zealand*. PhD thesis, Christchurch, University of Canterbury.

[Calonder et al., 2010] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). Brief: Binary robust independent elementary features. In *Proc. of European Conf. on Computer Vision*, pages 778–792.

[Canny, 1986] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

[Courbon et al., 2009] Courbon, J., Mezouar, Y., and Martinet, P. (2009). Autonomous navigation of vehicles from a visual memory using a generic camera model. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):392–402.

[Davison et al., 2007] Davison, A., Reid, I., Molton, N., and Stasse, O. (2007). Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067.

[Diosi et al., 2011] Diosi, A., Segvic, S., Remazeilles, A., and Chaumette, F. (2011). Experimental evaluation of autonomous driving based on visual memory and image-based visual servoing. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):870–883.

[Dissanayake et al., 2001] Dissanayake, M. G., Newman, P., Clark, S., Durrant-Whyte, H. F., and Csorba, M. (2001). A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241.

[Filliat, 2007] Filliat, D. (2007). A visual bag of words method for interactive qualitative localization and mapping. In *Proc. of IEEE International Conference on Robotics and Automation*, pages 3921–3926.

[Gálvez-López, 2013] Gálvez-López, D. (2013). *Place and Object Recognition for Real-time Visual Mapping*. PhD thesis, Zaragoza, Universidad de Zaragoza, Zaragoza. Presentado: 18-03-2013.

[Gálvez-López and Tardós, 2012] Gálvez-López, D. and Tardós, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197.

[Goedemé et al., 2007] Goedemé, T., Nuttin, M., Tuytelaars, T., and Van Gool, L. (2007). Omnidirectional vision based topological navigation. *International Journal of Computer Vision*, 74(3):219–236.

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Proc. of Alvey vision conference*, volume 15, page 50.

[Ido et al., 2009] Ido, J., Shimizu, Y., Matsumoto, Y., and Ogasawara, T. (2009). Indoor navigation for a humanoid robot using a view sequence. *The International Journal of Robotics Research*, 28(2):315–325.

[Jogan and Leonardis, 1999] Jogan, M. and Leonardis, A. (1999). Panoramic eigenimages for spatial localisation. In *Proc. of Computer Analysis of Images and Patterns*, pages 558–567.

[Kosecka and Yang, 2004] Kosecka, J. and Yang, X. (2004). Global localization and relative pose estimation based on scale-invariant features. In *Proc. of International Conference on Pattern Recognition*, volume 4, pages 319–322.

[Lowe, 2004a] Lowe, D. G. (2004a). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[Lowe, 2004b] Lowe, D. G. (2004b). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[Matsumoto et al., 1999] Matsumoto, Y., Ikeda, K., Inaba, M., and Inoue, H. (1999). Visual navigation using omnidirectional view sequence. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 317–322.

[Matsumoto et al., 1996] Matsumoto, Y., Inaba, M., and Inoue, H. (1996). Visual navigation using view-sequenced route representation. In *Proc. of IEEE International Conference on Robotics and Automation*, volume 1, pages 83–88.

[Menegatti et al., 2004] Menegatti, E., Maeda, T., and Ishiguro, H. (2004). Image-based memory for robot navigation using properties of omnidirectional images. *Robotics and Autonomous Systems*, 47(4):251–267.

[Murillo et al., 2007] Murillo, A., Sagüés, C., Guerrero, J. J., Goedemé, T., Tuytelaars, T., and Van Gool, L. (2007). From omnidirectional images to hierarchical localization. *Robotics and Autonomous Systems*, 55(5):372–382.

[Nister and Stewenius, 2006] Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168.

[Pajdla and Hlaváč, 1999] Pajdla, T. and Hlaváč, V. (1999). Zero phase representation of panoramic images for image based localization. In *Proc. of Computer Analysis of Images and Patterns*, pages 550–557.

[Rosin, 1999] Rosin, P. L. (1999). Measuring corner properties. *Computer Vision and Image Understanding*, 73(2):291–307.

[Rosten and Drummond, 2006] Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In *Proc. of European Conf. on Computer Vision*, pages 430–443.

[Royer et al., 2007] Royer, E., Lhuillier, M., Dhome, M., and Lavest, J.-M. (2007). Monocular vision for mobile robot localization and autonomous navigation. *International Journal of Computer Vision*, 74(3):237–260.

[Rublee et al., 2011] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: an efficient alternative to sift or surf. In *Proc. of IEEE International Conference on Computer Vision*, pages 2564–2571.

[Šegvić et al., 2009] Šegvić, S., Remazeilles, A., Diosi, A., and Chaumette, F. (2009). A mapping and localization framework for scalable appearance-based navigation. *Computer Vision and Image Understanding*, 113(2):172–187.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video google: a text retrieval approach to object matching in videos. In *Proc. of IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2.

[Smith et al., 2009] Smith, M., Baldwin, I., Churchill, W., Paul, R., and Newman, P. (2009). The new college vision and laser data set. *The International Journal of Robotics Research*, 28(5):595–599.

[Stricker et al., 2001] Stricker, D., Dähne, P., Seibert, F., Christou, I., Almeida, L., Carlucci, R., and Ioannidis, N. (2001). Design and development issues for archeoguide: An augmented reality based cultural heritage on-site guide. In *Proc. of Int. Conf. Augmented Virtual Environments and 3D Imaging*, pages 1–5.

[Tamimi et al., 2006] Tamimi, H., Andreasson, H., Treptow, A., Duckett, T., and Zell, A. (2006). Localization of mobile robots with omnidirectional vision using particle filter and iterative sift. *Robotics and Autonomous Systems*, 54(9):758–765.

[Ulrich and Nourbakhsh, 2000] Ulrich, I. and Nourbakhsh, I. (2000). Appearance-based place recognition for topological localization. In *Proc. of IEEE International Conference on Robotics and Automation*, volume 2, pages 1023–1029.

[van Veen et al., 1998] van Veen, H. A., Distler, H. K., Braun, S. J., and Bülthoff, H. H. (1998). Navigating through a virtual city: Using virtual reality technology to study human action and perception. *Future Generation Computer Systems*, 14(3):231–242.