UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

ESCUELA POLITÉCNICA SUPERIOR DE ELCHE

GRADO EN INGENIERÍA ELECTRÓNICA Y AUTOMÁTICA INDUSTRIAL

LOCALIZACIÓN DE UN ROBOT MÓVIL UTILIZANDO INFORMACIÓN VISUAL Y REDES NEURONALES CONVOLUCIONALES

TRABAJO FIN DE GRADO

Junio -2020

AUTOR: Orlando José Céspedes Gómez

DIRECTOR/ES: Luis Payá Castelló

Sergio Cebollada López

Biblioteca
UNIVERSITAS Miguel Hernández

Agradecimientos

En primer lugar, quiero dar las gracias a mis directores de trabajo fin de grado, Luis Payá y Sergio Cebollada, quienes confiaron en mí y me brindaron su mano siempre que yo lo necesitaba. Seguidamente, quiero agradecer al grupo de trabajo del laboratorio ARVC porque desde el primer momento me hicieron sentir parte de él.

Fuera del ámbito universitario, y de una forma muy especial, quiero dedicar este trabajo a mi familia. A mis padres, que dejaron todo atrás en nuestra hermosa Colombia para brindarnos un mejor futuro a mi hermano y a mí. Ellos, que siempre han pensado en nosotros antes que en sí mismos. Vuestra constante lucha ha dado sus frutos, lo habéis conseguido. A ti, Daniel, que con tu incansable esfuerzo siempre has dado todo por mantener unida esta familia, por muy lejos que pudiésemos encontrarnos unos de otros. Todo lo que he llegado a ser hasta el día de hoy es por vosotros. Muchas gracias.

Gracias a ti también, Ana, por tu apoyo, porque sin ti este trabajo no hubiese sido lo que es, por tus constantes ánimos y por tu paciencia, por haber estado siempre ahí y haberme empujado a llegar hasta este momento.

A todos vosotros, de nuevo, gracias. Cierro esta gran etapa universitaria rodeado de unas personas increíbles, que día a día me aportan bienestar y felicidad para seguir adelante. Y qué mejor que eso.

Biblioteca
UNIVERSITAS Miguel Hernández

Biblioteca
UNIVERSITAS Miguel Hernández

Índice general

1. Introducción **1**

2. Visión omnidireccional **11**

2.1. Cámaras omnidireccionales 11

2.2. Descriptores de apariencia global 13

2.2.1. Descriptores holísticos basados en métodos analíticos 14

2.2.2. Descriptores holísticos basados en *deep learning* 18

3. Material y métodos **23**

3.1. Base de datos empleada 23

3.2. Redes neuronales utilizadas 30

4. Experimentos **35**

4.1. Localización visual 35

4.2. Experimento 1: Comparación entre descriptores basados en *deep learning* y descriptores basados en métodos analíticos 40

4.3. Experimento 2: Comparativa entre las diferentes capas de la CNN 42

4.4. Experimento 3: Comparación entre CNNs 46

4.5. Experimento 4: Localización visual frente a efectos visuales 53

4.6. Experimento 5: Comparación entre diferentes fuentes de información 56

4.7. Experimento 6: Localización visual frente a rotaciones 63

5. Conclusiones y trabajos futuros **67**

Referencias **73**

Biblioteca
UNIVERSITAS Miguel Hernández

Índice de figuras

Figura 1.1: Múltiples sistemas robóticos. 2

Figura 1.2: Interrelación entre los conceptos que engloba la navegación integrada en robótica móvil: creación de mapas, localización y planificación de trayectorias. 4

Figura 1.3: Robot rover *Opportunity*. 5

Figura 2.1: Plataforma móvil empleada para la adquisición de las imágenes por el laboratorio de Saarbrücken. La toma de corriente portátil junto con la configuración de la cámara se muestran en la parte derecha de la imagen. ... 13

Figura 2.2: Diseño de arquitectura CNN del modelo pre-entrenado *Caffe*. 20

Figura 3.1: Imágenes adquiridas en el laboratorio de Saarbrücken. La figura (a) muestra las imágenes tomadas mediante la cámara perspectiva, mientras que la figura (b) lo hace mediante la cámara omnidireccional. Imágenes obtenidas de COLD *database* [39]. 26

Figura 3.2: Recorrido del robot móvil en el laboratorio de Saarbrücken. Imágenes obtenidas de COLD *database* [39]. 27

Figura 3.3: Imagen omnidireccional RGB. 28

Figura 3.4: Imagen omnidireccional recortada RGB. 29

Figura 3.5: Imagen panorámica en escala de grises. 29

Figura 3.6: Diseño de arquitectura CNN del modelo pre-entrenado *AlexNet*. .. 31

Figura 3.7: Diseño de arquitectura CNN del modelo pre-entrenado *GoogLeNet*. . 34

Figura 4.1: Error medio de localización de los descriptores obtenidos de las capas de clasificación de la CNN *Places* junto con los descriptores HOG y gist. 41

Figura 4.2: Tiempo medio total de cómputo de los descriptores obtenidos de las capas de clasificación de la CNN *Places* junto con los descriptores HOG y gist... 42

Figura 4.3: Error medio de localización de los descriptores obtenidos de las capas convolucionales de la CNN *Places* junto con el descriptor obtenido por la capa ‘fc6’. 44

Figura 4.4: Tiempo medio total de cómputo de los descriptores obtenidos de las capas convolucionales de la CNN *Places* junto con el descriptor obtenido por la capa ‘fc6’. 45

Figura 4.5: Error medio de localización de los descriptores obtenidos de las capas convolucionales de la CNN *AlexNet*. 48

Figura 4.6: Tiempo medio total de cómputo de los descriptores obtenidos de las capas convolucionales de la CNN *AlexNet*. 49

Figura 4.7: Error medio de localización de los descriptores obtenidos de ciertas capas convolucionales de la CNN *GoogLeNet*..... 50

Figura 4.8: Tiempo medio total de cómputo de los descriptores obtenidos de ciertas capas convolucionales de la CNN *GoogLeNet*. 50

Figura 4.9: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places*, *AlexNet* y *GoogLeNet*. 51

Figura 4.10: Tiempo medio total de cómputo de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places*, *AlexNet* y *GoogLeNet*. 52

Figura 4.11: La figura (a) muestra la imagen original capturada dentro del entorno de Saarbrücken, mientras que diversos efectos aplicados a nuestro modelo visual se observan en las siguientes: (b) ruido Gaussiano, (c) oclusión y (d) efecto *blur*. 54

Figura 4.12: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places* y *AlexNet* junto con el descriptor HOG frente a los efectos visuales de ruido (naranja), oclusiones (amarillo) y efecto *blur* (morado). 56

Figura 4.13: La figura (a) muestra la imagen panorámica en escala de grises, la figura (b) la panorámica en color, la figura (c) la omnidireccional en color y la figura (d) la omnidireccional en escala de grises. Todas ellas fueron capturadas dentro del entorno de Saarbrücken. Imágenes obtenidas de COLD *database* [39]. 57

Figura 4.14: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places*, *AlexNet* y *GoogLeNet* calculado en función del tipo de fuente utilizada: panorámica en escala de grises (azul) o en RGB (naranja) y omnidireccional en RGB (amarillo) o en escala de grises (morado). 59

Figura 4.15: Tiempo medio total de cómputo del descriptor obtenido de la capa convolucional 'conv2' de la CNN *Places*. 60

Figura 4.16: Tiempo medio total de cómputo del descriptor obtenido de la capa convolucional 'conv4' de la CNN *AlexNet*. 60

Figura 4.17: Tiempo medio total de cómputo del descriptor obtenido de la capa convolucional 'inception_3b-1x1' de la CNN *GoogLeNet*. 61

Figura 4.18: La figura (a) muestra la imagen original capturada dentro del entorno de Saarbrücken, mientras que la figura (b) representa el efecto de rotación aplicado sobre la misma. 64

Figura 4.19: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales y de clasificación de la CNN *Places* y *AlexNet* junto con el descriptor HOG. 66

Biblioteca
UNIVERSITAS Miguel Hernández

Índice de tablas

Tabla 3.1: Parámetros y configuración de las cámaras. 24

Tabla 3.2: Estancias recorridas por el robot móvil, en el laboratorio de Saarbrücken, mientras se realizaba la adquisición de las imágenes. 25

Tabla 3.3: Número de imágenes de cada *dataset*, tanto de entrenamiento como de test, creado a partir del entorno COLD – Saarbrücken. 30

Tabla 3.4: Capas utilizadas de la CNN *AlexNet*. 32

Tabla 3.5: Capas utilizadas de la CNN *Places*. 33

Tabla 3.6: Capas utilizadas de la CNN *GoogLeNet*. 34

Tabla 4.1: Tipos de distancias utilizadas entre los descriptores. 39

Tabla 4.2: Mejor canal obtenido para la tarea de localización en cada una de las capas convolucionales de la CNN *Places*. 43

Tabla 4.3: Mejor canal obtenido para la tarea de localización en cada una de las capas convolucionales de la CNN *AlexNet*. 46

Tabla 4.4: Mejor canal obtenido para la tarea de localización en ciertas capas convolucionales de la CNN *GoogLeNet*. 47

Tabla 4.5: Tiempo de cálculo necesario para obtener la conversión de las diferentes imágenes a introducir en nuestras redes neuronales convolucionales. 63

Biblioteca
UNIVERSITAS Miguel Hernández

Capítulo 1
Introducción

Desde hace décadas la robótica se ha convertido en una de las ramas de la ciencia con mayor importancia, dando pie a un gran desarrollo e investigación en todo lo relacionado a este tema, llegando a ser un campo de estudio imprescindible en la actualidad. De esta forma se capacita a máquinas para hacer tareas por sí mismas, ya que la robótica autónoma son sistemas que reemplazan a los humanos para desarrollar tareas mecánicas, rutinarias o peligrosas, y necesitan un alto grado de autonomía para llevar estas a cabo.

Si para desarrollar la tarea para la que fue diseñado, el robot necesita moverse dentro de un determinado entorno, entonces se tendrán que resolver las tareas relacionadas con la robótica móvil. Asimismo, hay que tener presente que los robots móviles tienen la capacidad de moverse a través de su entorno y no se fijan a una ubicación física, es por ello por lo que la utilización de estos en la sociedad se ha incrementado exponencialmente en los últimos años, ya que se ha experimentado un gran avance debido a la variedad de problemas que puede solucionar esta tecnología. En este sentido, el hecho de que estos no se encuentren fijos y puedan navegar por el entorno en el que se hallan permite mejorar la labor de operar en zonas amplias y heterogéneas sin necesidad de introducir cambios en la estructura del robot, provocando de esta forma el aumento de los ámbitos en los que se puede emplear la robótica. Un buen ejemplo sería en ambientes de difícil acceso para el ser humano: misiones espaciales (mantenimiento de estaciones orbitales, etc.) [1], tareas submarinas (mediciones, misiones de búsqueda y rescate, etc.) [19] o subterráneas (minería, construcción de túneles, etc.) [32], en vigilancia e intervención de seguridad (desactivación de explosivos, operación en zonas radioactivas, etc.) [14] y

aplicaciones militares [40]. En la actualidad, este tipo de robot se está utilizando también en el hogar para la realización de tareas domésticas, entretenimiento, e igualmente como asistentes robóticos para personas de mayor edad [9]. A continuación, en la figura 1.1, se muestran algunos ejemplos de los diferentes tipos de robots desarrollados actualmente.

Figura 1.1: Múltiples sistemas robóticos.

Si bien a lo largo de muchos años la robótica se centraba en aspectos tales como la capacidad de rendimiento, la precisión de repetición y la velocidad, desde finales del siglo XX se ha ido abriendo paso una innovadora investigación dentro del genérico campo de la robótica: la robótica móvil. Esta es un área de conocimiento multidisciplinar que dota a robots móviles de diferentes sensores que les permiten extraer la información necesaria del entorno para poder hacer de forma autónoma las labores para las cuales fueron diseñados. El término autonomía hace referencia al hecho de que el robot sea capaz de realizar estas tareas sin intervención humana.

Para que un robot móvil pueda realizar una determinada tarea de forma autónoma, en un medio desconocido, es necesario que construya un modelo de dicho entorno. Este modelo debe contener información suficiente para que el robot sea capaz de estimar su posición y orientación, con suficiente precisión y un coste computacional aceptable; y, a continuación, planificar una trayectoria libre de obstáculos con objeto de llegar a los puntos deseados. En estos aspectos reside la importancia de encontrar soluciones óptimas a los problemas de construcción de mapas (*mapping*) y localización.

Estos son considerados dos problemas cruciales en el desarrollo de robots móviles plenamente autónomos. Las soluciones a estos problemas son desde hace años, y siguen siendo actualmente, áreas de investigación muy activas en el campo de la robótica. En primer lugar, el término *mapping* se refiere al problema de construir modelos espaciales en entornos físicos haciendo uso de la información capturada por los sistemas de percepción montados en el entorno y/o en uno o varios robots, normalmente móviles. Por otra parte, en cuanto al concepto de localización, este consiste en estimar la pose del robot en el modelo, es decir, sus coordenadas de posición y orientación en todos los grados de libertad de movimiento que posea dicho robot; comparando la información capturada por sus sistemas de percepción con la información almacenada en el mapa previamente creado. En ocasiones, el problema de localización debe resolverse a la vez que se está construyendo el mapa. Se trata de un problema muy común puesto que cuando un robot comienza a realizar una tarea por primera vez, en un medio desconocido, debe comenzar a generar un modelo del entorno desde cero mientras estima su posición. Es a este proceso al que se le denomina localización y creación de mapas de forma simultánea, conocido comúnmente mediante sus siglas en inglés: SLAM (*Simultaneous Localization and Mapping*).

Los tres conceptos anteriores (creación de mapas, localización y planificación de trayectorias) se relacionan como se muestra en la figura 1.2. Seguidamente, seguiremos con el resto de los términos que aparecen en dicha imagen, dejando a un lado la tarea de SLAM, ya explicada con anterioridad. Por un lado, el uso simultáneo de algoritmos de creación de mapas con algoritmos de planificación de trayectorias es conocido como exploración. El objetivo de este problema será decidir la trayectoria óptima que deben seguir los robots para construir un mapa del entorno lo más rápidamente posible, pero sin perder precisión en la información que se almacena. Por otro lado, la combinación de los problemas de localización y planificación de trayectorias da lugar a la tarea de localización activa, en la que se intenta que el robot recorra ciertas trayectorias que le ayuden a perfeccionar su localización a partir de la información adquirida a lo largo de dichos recorridos. Por último, la ejecución paralela de los problemas de creación de mapas, localización y planificación de

trayectorias da lugar a lo que denominamos sistemas de navegación integrados, horizonte último que se debe plantear durante el diseño de un robot móvil autónomo [25].

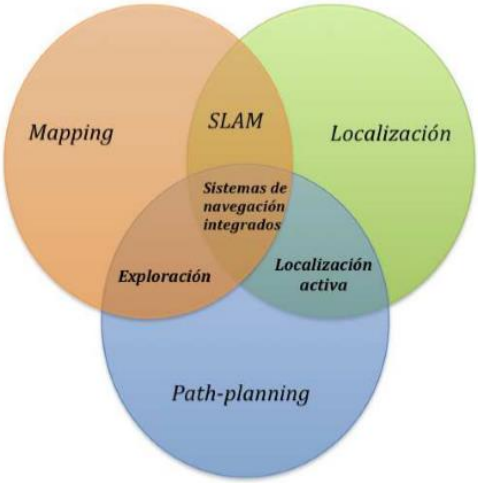


Figura 1.2: Interrelación entre los conceptos que engloba la navegación integrada en robótica móvil: creación de mapas, localización y planificación de trayectorias.

Una gran parte de la investigación y desarrollo de la robótica móvil actual se basa en técnicas que proporcionen a los robots un mayor grado de autonomía y versatilidad. Con el paso de los años se ha hecho imprescindible el uso de la tecnología de visión artificial para la tarea de navegación, ya que esta tecnología nos permite adquirir, procesar y analizar imágenes. De hecho, se puede decir que el primer intento de utilizar la visión artificial fue con el robot llamado Shakey [23] seguido del Stanford Cart [20]. Este último disponía de un sistema de visión compuesto por dos cámaras, una de las cuales podía moverse para proporcionar distintos puntos de vista para la visión estereoscópica. No obstante, un ejemplo más conocido podría ser el que se muestra en la figura 1.3, uno de los robots más importantes de la NASA pertenecientes al programa *Mars Exploration Rover* (MER). Hay que tener en cuenta que un rover es un dispositivo de exploración de superficie planetaria creado para moverse a través de la superficie de un planeta u otros cuerpos celestes. Algunos rovers han sido diseñados como vehículos terrestres para transportar a miembros de una tripulación de vuelo espacial; otros han sido robots parcial o totalmente autónomos, como es el caso

del rover *Opportunity*. Este fue diseñado para explorar algunos de los terrenos más recónditos del planeta Marte, analizando sedimentos depositados en un rango temporal superior a los mil millones de años, siendo uno de los hitos más importantes la exploración del cráter Victoria [36]. Debido a que los rovers no pueden ser controlados de forma remota en tiempo real, ya que la velocidad a la que viajan las señales de radio es demasiado lenta para tener una comunicación instantánea, estos deben operar de forma autónoma, con poca ayuda del control de tierra en lo que respecta a la navegación y adquisición de datos. Es por ello por lo que es tan importante la autonomía de estos robots móviles, el conseguir que puedan tomar grandes decisiones por sí mismos hizo posible que el rover *Opportunity* superara múltiples tormentas y varios fallos mecánicos. De esta forma se logró que estuviese activo durante 15 años en la misión para la que fue diseñado a pesar de que se construyó pensando que su vida espacial sería de tan solo 90 días.

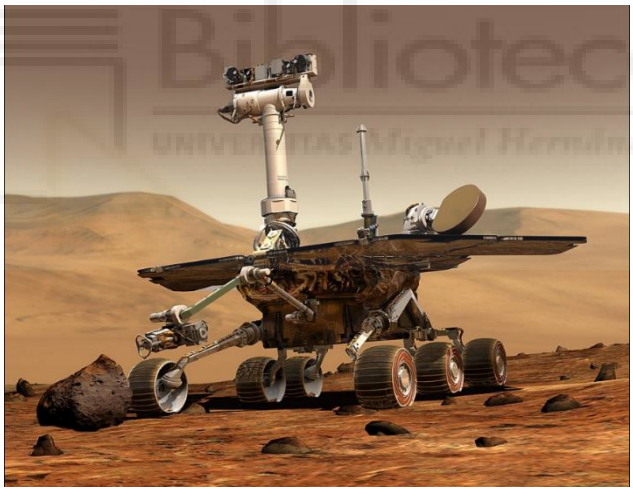


Figura 1.3: Robot rover *Opportunity*.

El hecho de que estos robots presenten navegación autónoma hace que tengan la capacidad de planificar y seguir una trayectoria de forma óptima evitando los obstáculos presentes en su entorno de trabajo [12].

El problema de la navegación se puede dividir en cuatro etapas: (a) percepción, el robot extrae información relevante mediante los sensores, (b) localización, calcula su posición en el mapa, (c) planificación de ruta, el robot define la trayectoria que debe seguir para llegar a su objetivo y (d) control de

movimiento, regula su movimiento para seguir la trayectoria deseada [18]. El sistema de percepción tiene como objetivos permitir que el robot navegue de forma segura detectando y localizando los obstáculos que pueda encontrar, modelar el entorno y conocer su posición. Algunos de los sensores que se pueden utilizar para este fin son el sonar, el ultrasónico, el láser o las cámaras.

La utilización de éstos últimos se ha visto incrementada debido a las ventajas que presentan. De hecho, se aplica en distintos ámbitos de la robótica móvil, como la localización, la navegación visual, la odometría visual, la tarea de SLAM, etc. Los sistemas de visión ofrecen abundante información sobre la escena, siendo similar a la que obtendría el ojo humano. Entre otras de sus ventajas se encuentran su bajo coste y que son fáciles de usar. Además, son más eficientes en términos de consumo que otros sensores, por lo que son una buena opción en tareas cuya realización requiere un elevado tiempo [10]. De entre la alta variedad de tipos de cámaras que ofrece el mercado hoy en día, destaca el uso de las cámaras omnidireccionales, las cuales han presentado la ventaja de ofrecer un campo de visión más amplio. De esta manera, si se incorpora una cámara omnidireccional sobre un sistema robótico, con un solo sensor se puede obtener información en 360º del entorno que rodea al robot, incluyendo así toda la escena en una única imagen.

Generalmente, en visión por computador se extrae información relevante de las imágenes. Los métodos de extracción de información más utilizados son los basados en características locales y los basados en apariencia global (también conocidos como métodos de descripción holística). El primer modo consiste en obtener un conjunto de puntos pertenecientes a cada imagen con unas determinadas características. Sin embargo, los métodos basados en la apariencia global utilizan las imágenes en su conjunto, no extraen ninguna información local. Por tanto, se obtendrá un único vector para cada imagen que contendrá información sobre su apariencia global. Este tipo de métodos presentan algunas ventajas en aquellos entornos dinámicos o mal estructurados [27].

Hay que tener en cuenta que esta información visual que se obtiene por medio de las imágenes puede verse afectada por los cambios de iluminación. Estos pueden hacer ver comprometida una correcta tarea de localización, haciendo creer al robot que se encuentra en un punto muy distinto al real debido a dichos efectos.

Todo este trabajo de procesamiento, en la tarea de saber dónde exactamente se encuentra nuestro robot, pasa a ser mucho más eficiente cuando trabajamos con técnicas de *machine learning*. Esta es una derivación de la inteligencia artificial que crea sistemas que aprenden de manera automatizada, es decir, identifican patrones complejos de entre millones de datos y predicen comportamientos mediante algoritmos. Dichos algoritmos son capaces, de manera independiente, de autoprogramarse según lo que aprenden y la experiencia que obtienen de la combinación de datos y el procesamiento de la información. Esta técnica es muy importante a la hora de la creación de mapas que representen la actividad y la navegabilidad del entorno en el que se encuentra nuestro robot móvil [42]. Estas técnicas pasan a ser más precisas cuando, de una forma más exhaustiva, tratamos con *deep learning*, el cual lleva a cabo el proceso de *machine learning* usando una red neuronal artificial que se compone de determinados niveles jerárquicos. Con ello conseguimos un aprendizaje profundo empleando modelos informativos y creando redes neuronales artificiales para la transmisión y análisis de datos. Esta técnica más reciente ha mostrado resultados sobresalientes a la hora de resolver una amplia variedad de tareas en robótica móvil, como puede ser en las áreas de percepción, planificación, localización y control [4].

Con todo ello podemos ver las grandes oportunidades que nos ofrece el tratamiento de imágenes mediante estas técnicas, debido a que estaríamos trabajando con mayor precisión, anticipándonos a los problemas que fuesen apareciendo en el camino a recorrer y aprovechando mejor nuestros recursos en un menor tiempo de actuación.

Este trabajo fin de grado tiene como principal objetivo resolver la tarea de localización de robots móviles, en entornos de interior cambiantes, mediante la

utilización de descriptores holísticos obtenidos a través de herramientas de *deep learning*. Dado que se trata de una técnica menos madura, es interesante estudiar con profundidad su comportamiento en tareas de localización y la robustez de cada uno de estos algoritmos de descripción ante determinados fenómenos que hacen cambiar el entorno real de trabajo. La manera de resolver dicha tarea es por medio del método conocido como *image retrieval* o recuperación de la imagen, el cual se explicará con más detalle en capítulos posteriores.

Durante el proceso de localización se ha realizado un estudio acerca de cuán robustos son los descriptores holísticos comprobando su eficiencia ante el cambio sustancial de la apariencia visual del entorno, ya que de estar en un ambiente de trabajo real el robot podría enfrentarse a diversos fenómenos perjudiciales, los cuales debe saber solventar con éxito para poder llevar a cabo una correcta tarea. Estos, por ejemplo, podrían ser los siguientes: un cambio de iluminación en el ambiente de trabajo con respecto a las imágenes que se tomaron inicialmente y que componen el modelo visual; una variación en el entorno, ya que puede verse parcialmente ocluido debido a la actividad humana en el espacio de trabajo; una nueva organización del medio que haya hecho cambiar el mobiliario de las estancias, etc.

Por último, el presente trabajo se encuentra estructurado de la siguiente manera:

- En el capítulo 2 se presenta una explicación más exhaustiva acerca de las cámaras omnidireccionales y el beneficio que estas aportan a la hora de ser usadas en la tarea de localización de robots móviles en entornos de interior. Por otro lado, principalmente, se presentan los descriptores de apariencia global, también llamados descriptores holísticos, haciendo previamente una pequeña mención al funcionamiento de los descriptores de apariencia local. Para finalizar, se explicarán las dos ramas que actualmente tenemos al trabajar con descriptores holísticos, ya que para que estos sean

obtenidos podemos basarnos en métodos analíticos o en herramientas de *deep learning*.

- A continuación, el capítulo 3 está centrado en la descripción de la base de datos utilizada, de la cual hemos creado nuestro modelo visual. Se detalla, también, cómo han sido obtenidas las imágenes, así como las estancias que el robot móvil ha recorrido. En este mismo capítulo, posteriormente, se especifican las redes neuronales convolucionales empleadas, *Convolutional Neural Networks* (CNN), para el tratamiento de la información. Se describe el funcionamiento de estas, tratando desde su configuración estructural hasta el tipo de imagen que estas aceptan.

- En el capítulo 4 se explica, en primer lugar, el proceso de localización de forma detallada junto al algoritmo que hace posible esta tarea y, seguidamente, se detallan los seis experimentos que se han realizado para conocer qué descriptor de cada capa de la red neuronal convolucional es más eficiente a la hora de tratar con los diferentes efectos aplicados al modelo visual.

- Finalmente, las principales conclusiones se exponen en el capítulo 5, en el cual se detallan asimismo las aportaciones más relevantes obtenidas tras haber realizado lo expuesto en la sección anterior y las futuras líneas de investigación.

Biblioteca
UNIVERSITAS Miguel Hernández

Capítulo 2
Visión omnidireccional

Para llevar a cabo la creación de mapas, localización y navegación de robots móviles es necesario hacer uso de la información percibida por los sistemas sensoriales del robot. Desde los primeros trabajos sobre robots móviles [13] los sistemas de control diseñados han hecho uso, en mayor o menor grado, de la información recogida del entorno por el que el robot se mueve. La forma de trabajar con esta información ha ido evolucionando según surgían nuevos algoritmos y la capacidad de los sistemas de percepción y de cómputo aumentaba.

El hecho de incorporar visión en robots les otorga la capacidad de percibir visualmente el medio que les rodea e interactuar con él. Asimismo, la visión se utiliza en diversas técnicas de la robótica, debido al gran contenido de información que proporcionan sobre la escena. La robótica visual hace uso de los métodos de visión por computador para llevar a cabo las tareas que debe realizar el robot. El objetivo de dichos métodos es intentar entender la escena y los objetos que se encuentran en ella, semejándose a la función que realiza el ser humano.

2.1. Cámaras omnidireccionales

Una cámara es un mecanismo de formación de imágenes como resultado de que la luz se proyecte sobre una superficie sensible a la misma. Se pueden obtener diferentes cámaras variando tres elementos: (1) la geometría de la superficie, (2) la distribución geométrica y propiedades ópticas de los fotorreceptores, y (3) la forma en la que se recoge y proyecta la luz sobre la superficie (lentes simples o múltiples). Los sistemas de visión procesan estas

imágenes para reconocer, navegar, y generalmente interactuar con el entorno [22]. Las cámaras presentan diversas ventajas en relación con otros tipos de sensores, como su bajo coste, peso y consumo de energía, siendo esta última una característica de gran utilidad para los robots autónomos, además de la alta información que ofrecen del entorno.

En lo que respecta al sensor, podemos encontrar sistemas basados en sistemas monoculares [34] y binoculares (pares estéreo) [35], aunque algunos autores también proponen soluciones basadas en configuraciones trinoculares [2]. Estas dos últimas configuraciones permiten medir la profundidad de las imágenes, pero requieren de varias de estas para obtener información completa del entorno. Asimismo, los sistemas de visión omnidireccionales pueden estar compuestos por varias cámaras apuntando hacia distintas direcciones o por una única cámara y una superficie reflexiva [27].

En los últimos años han ganado popularidad los sistemas de visión omnidireccionales [41] gracias a la gran cantidad de información que suministran, ya que tienen un campo de visión de 360º alrededor del robot; la estabilidad de las características que aparecen en la imagen, puesto que permanecen durante más tiempo en el campo de visión según el robot se mueve; su coste relativamente bajo comparado con otros sensores como, por ejemplo, los sensores láser; y su bajo consumo, que resulta de ayuda en el diseño de robots autónomos que deben funcionar con baterías durante largos periodos de tiempo. Estos sistemas de visión omnidireccionales suelen estar basados en la combinación de una cámara convencional con un espejo convexo cónico, esférico, parabólico o hiperbólico (sistemas catadióptricos). La información visual proporcionada por estas cámaras se puede representar en diversos formatos: omnidireccional, panorámico o vista en planta [26], [11]. En este trabajo haremos uso de la representación panorámica y de la omnidireccional. Como mostraremos en posteriores capítulos, ambas contienen suficiente información para poder estimar la posición del robot y su orientación cuando el movimiento del robot está restringido al plano del suelo.

Asimismo, el sensor usado en este trabajo fin de grado ha sido un sistema de visión catadióptrico montado sobre el propio robot, tal como podemos ver en la figura 2.1, que proporciona imágenes con un campo de visión de 360° alrededor del mismo. Es con este robot móvil con el que se ha creado el modelo visual del entorno a partir de las imágenes omnidireccionales capturadas.

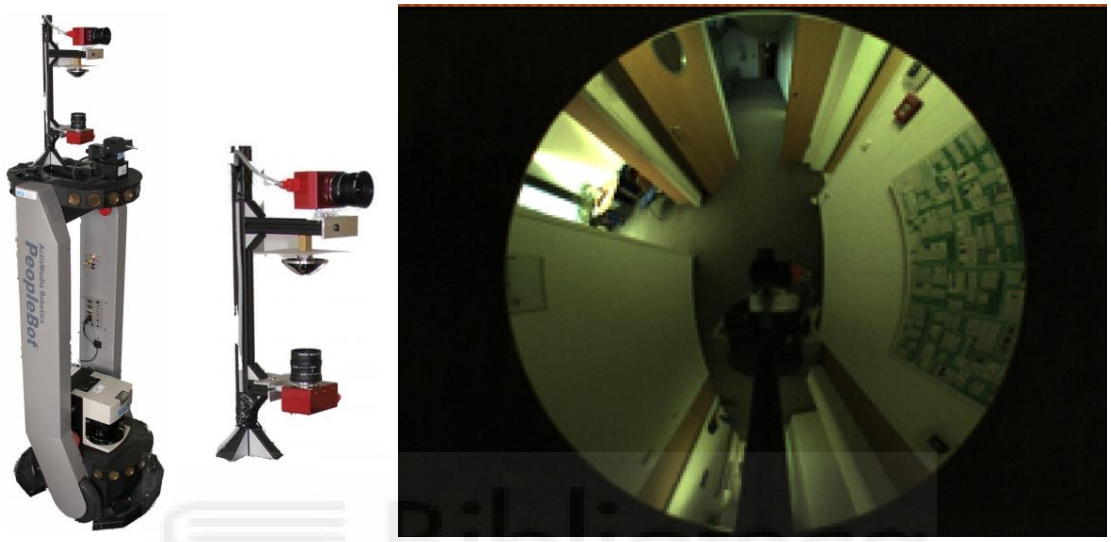


Figura 2.1: Plataforma móvil empleada para la adquisición de las imágenes por el laboratorio de Saarbrücken. La toma de corriente portátil junto con la configuración de la cámara se muestran en la parte derecha de la imagen.

2.2. Descriptores de apariencia global

El uso de información visual se lleva a cabo comúnmente mediante descriptores, es decir, métodos que extraen la información más relevante de la imagen; dentro de esta técnica hay dos grandes vertientes:

Por un lado, la utilización de descriptores de apariencia local, que se basan en la extracción de puntos, objetos o regiones características de la escena y obtienen un vector de información por cada punto seleccionado dentro de esta. Algunos de los métodos más populares y ampliamente utilizados para la obtención de estos descriptores es el *Scale Invariant Features Transform* (SIFT) [17] y el *Speeded-Up Robust Features* (SURF) [3]. Ambos se han utilizado en tareas de creación de mapas y localización con robots móviles, como se muestra, entre otros, en los trabajos de Se et al. [31] y Murillo et al. [21], respectivamente.

Por otro lado, la utilización de descriptores de apariencia global o descriptores holísticos, que se basan en la creación de un único vector por imagen que contiene información de toda la escena. El hecho de tener un único descriptor hace que los métodos de *mapping* y localización sean sencillos, basados muchas veces en la comparación de pares de imágenes. Asimismo, la aplicación a entornos estructurados es más directa y su coste computacional suele ser menor. Son estos descriptores los que han sido utilizados para el estudio de este trabajo. Estos, a su vez, se pueden subdividir en dos ramas en función de la forma en que pueden ser calculados: descriptores holísticos basados en métodos analíticos y descriptores holísticos basados en *deep learning*.

En este capítulo revisaremos los algoritmos más relevantes para describir las características globales de un conjunto de escenas y estudiaremos las propiedades que los hacen aplicables para resolver problemas con robots móviles.

2.2.1. Descriptores holísticos basados en métodos analíticos

Los descriptores de apariencia global obtenidos mediante métodos analíticos se basan, fundamentalmente, en cálculos de gradientes de orientación y/o color de los diferentes píxeles que componen la imagen. Su uso ha sido bastante frecuente para resolver problemas de robótica móvil, dado que estos métodos conducen a algoritmos de localización más directos basados en comparación entre pares de descriptores.

Entre las técnicas de descripción global de la imagen encontramos descriptores como *Principal Components Analysis* (PCA) [15], una de las primeras alternativas robustas que aparecieron. Este descriptor considera la imagen como un conjunto de información multidimensional que puede ser proyectada en dimensiones bajas del espacio reteniendo información original, pero únicamente la más importante.

Distintos autores han propuesto otro tipo de descriptores diferentes, como aplicar la transformada de Fourier a la imagen para extraer la información más relevante, de esta operación obtenemos el descriptor *Fourier Signature* (FS). Existen varias alternativas a la hora de trabajar con la transformada de Fourier. Por un lado, si se realiza la transformada a imágenes panorámicas encontramos autores que han escrito sobre los resultados obtenidos, como es el caso de Payá et al. [26]; y, por otro lado, también es posible aplicar la *Spherical Fourier Transform* (SFT) a la imagen omnidireccional, como en el trabajo propuesto por Rossi et al. [30]. En ambos casos el resultado del descriptor es la compresión de la información original en un menor número de componentes.

Seguidamente, otra serie de técnicas para realizar la descripción global de la imagen pueden ser la operación gradiente, *Histograms of Oriented Gradients* (HOG); y la convolución de filtros, gist. Estas dos son las técnicas utilizadas en este trabajo y se detallan a continuación.

Por una parte, los descriptores de Histograma de Orientación del Gradiente (HOG) son descriptores de características usados en visión por computador y en procesamiento de imágenes, normalmente para la detección de objetos. La técnica tiene en cuenta la orientación del gradiente en partes localizadas de una imagen. Estos descriptores destacan por la sencillez de su construcción y los buenos resultados que proporcionan, con un coste computacional muy aceptable. La experiencia con este tipo de descriptores en el campo de la robótica móvil es satisfactoria al haber presentado resultados interesantes para tereas de localización, tal como observamos en el trabajo de Cebollada et al. [5].

La idea que subyace en este tipo de descriptores es que tanto la apariencia de los objetos en una imagen como su forma pueden ser descritas por la distribución de la intensidad de los gradientes y la dirección de los bordes. La implementación de este descriptor puede lograrse dividiendo la imagen en pequeñas regiones conectadas, llamadas celdas, y compilando para cada una de ellas un histograma de orientación del gradiente para los píxeles dentro de dicha celda. La combinación de estos histogramas forma el descriptor.

Para mejorar el funcionamiento, los histogramas locales pueden ser normalizados calculando una medida de intensidad dentro de una región más grande de la imagen, la cual es llamada bloque; esta medida se usaría para normalizar todas las celdas dentro de dicho bloque. Se espera que esta normalización dote de mayor robustez al descriptor ante cambios en el nivel de iluminación de las escenas [25].

Hay que tener en cuenta que, en nuestro caso, el método de descripción HOG utilizado es invariante ante rotaciones, es decir, que el vector calculado será el mismo independientemente de la rotación que presente la imagen. Es por ello por lo que este es considerado como un descriptor visual de la posición del robot.

Por otra parte, los descriptores gist tratan de imitar la habilidad que tiene el sistema de percepción humano para reconocer inmediatamente una escena a través de la identificación de determinadas regiones que poseen un color, una orientación y/o una textura destacados en relación con el entorno. El concepto de gist fue introducido por Oliva et al. [24], bajo el nombre de *holistic representation of the spatial envelope*. La idea fundamental consiste en crear un descriptor de la escena, de dimensión pequeña, que evite el segmentado y procesamiento de puntos, objetos o regiones individuales. Los autores demuestran que este método de descripción genera un espacio multidimensional en el que las escenas que comparten pertenencia a una misma categoría semántica (como, por ejemplo, calles, edificios, costas, cielo, etc.) son proyectadas en puntos cercanos. La justificación del método surge de trabajos previos que sugirieron que el reconocimiento humano de escenas se inicia mediante la codificación de configuración global, ignorando la mayor parte de los detalles y la información de objetos individuales.

Matemáticamente, el método trata de codificar la información espacial realizando la transformada de Fourier 2D en varias regiones de la imagen distribuidas en una rejilla regularmente espaciada. El conjunto de descriptores de cada región es, en este momento, dimensionalmente reducido mediante el Análisis de Componentes Principales (PCA) para dar lugar a un único descriptor

de la escena de menor dimensión. Posteriormente, los mismos autores hacen uso de pirámides de wavelet orientables en lugar de la transformada de Fourier [38]. De esta forma demostraron con éxito cómo este tipo de descriptores son capaces de clasificar correctamente conjuntos de imágenes a partir de características subjetivas como el grado de naturalidad, el grado de apertura, el grado de aspereza, etc.

Otros trabajos más recientes hacen uso del concepto de prominencia, que hace referencia a las zonas de la escena que destacan significativamente respecto de sus vecinas, mientras que gist implica la acumulación de datos estadísticos de cada escena en su conjunto. Siagian et al. [33] tratan de establecer sinergias entre ambos conceptos en un único descriptor que sintetiza tanto características gist como características de prominencia y cuyo coste computacional es relativamente reducido. La construcción de este descriptor se basa en tres características fundamentales de la escena: intensidad, orientación y color. Para obtener un descriptor de prominencia de cada característica se construye para cada escena una pirámide que contiene la escena original y varias copias de menor resolución. En cada imagen de la pirámide se aplican operaciones que comparan cada píxel con sus alrededores y con los píxeles equivalentes de las imágenes de distinta escala dentro de la pirámide. Con estas operaciones conseguimos invariancia ante los cambios de iluminación debido a la comparación de los alrededores respecto del punto central.

En el presente trabajo haremos uso de los descriptores HOG y de los descriptores gist, pero introduciremos algunas modificaciones puesto que la tarea que pretendemos resolver es distinta a la tarea original para la que se plantearon estos mismos. En nuestro caso, el mapa estará creado por un conjunto de imágenes obtenidas de un entorno de interior, que presentarán muchas similitudes entre ellas. El objetivo será el de obtener un descriptor capaz de trabajar con imágenes de características similares (textura, apertura, estructuración, etc.), de modo que la distancia en el espacio descriptor refleje la distancia geométrica entre los puntos donde las imágenes fueron capturadas.

Tanto el descriptor HOG como el descriptor gist son utilizados en imágenes panorámicas en escala de grises, por lo que hay que tener en cuenta su previa conversión desde la imagen omnidireccional en color que obtenemos directamente de nuestro robot móvil para, de esta forma, poder trabajar correctamente con ellos. Esta conversión de la imagen será explicada con más detalle en el capítulo 3.

2.2.2. Descriptores holísticos basados en *deep learning*

El aprendizaje máquina o *machine learning* es un método de análisis de datos que, entre otras opciones, automatiza la construcción de descriptores de apariencia global basados en modelos analíticos. Esta técnica está incluida dentro de la rama de la inteligencia artificial y se sustenta en la idea de que el sistema puede aprender a identificar patrones partiendo exclusivamente de los datos aportados. Para el uso de esta técnica es necesario seleccionar previamente la información que se utilizará para entrenar el modelo y, posteriormente, llevar a cabo el proceso de aprendizaje, el cual, dependiendo de los requerimientos establecidos, puede llevar un mayor o menor tiempo de cómputo.

Dada la potencia de las técnicas de inteligencia artificial desarrolladas durante las últimas décadas, este trabajo ha tomado las recientes propuestas de estudiar herramientas más innovadoras, dentro del área de *machine learning*, para llevar a cabo de manera más eficiente la tarea de localización en robótica móvil. La idea es obtener nuevos descriptores holísticos que caractericen las imágenes a través de técnicas de *deep learning*. Algunos de los autores que han utilizado estas nuevas técnicas son, por ejemplo, Xu et al. [43] al proponer el uso del *autoencoder*, el cual es un tipo de red neuronal artificial que se utiliza para aprender codificaciones de datos eficientes de manera no supervisada, para obtener descriptores de apariencia global que le permitiesen detectar tumores de pecho; y Payá et al. [28] al utilizar los vectores obtenidos de capas intermedias de una red neuronal convolucional (CNN) como descriptores de apariencia global y, a partir de esta información, llevar a cabo la creación de mapas jerárquicos. En nuestro trabajo, las CNN serán utilizadas con el propósito de obtener un

descriptor holístico por imagen capturada para estudiar qué tan eficiente es con ellos la tarea de localización.

Las redes neuronales convolucionales reciben un exhaustivo entrenamiento para llevar a cabo una determinada tarea. Pero cabe señalar que su uso más extendido es el de resolver tareas de clasificación. En este sentido, en primer lugar, un conjunto de imágenes correctamente etiquetadas se recopila y se introduce en la red neuronal convolucional para abordar el proceso de aprendizaje; tras esto, la red estará preparada para resolver el problema de la clasificación: una imagen test como entrada (sin conocer a priori la etiqueta) y las salidas de la CNN como opciones de etiquetas más probables. Las redes neuronales convolucionales están compuestas por varias capas ocultas, cuyos parámetros y ponderaciones se ajustan durante el proceso de entrenamiento. Para nuestro trabajo, nos centramos en la obtención de descriptores holísticos, esto es, introducir una imagen a la CNN y obtener un vector de una de las capas intermedias, de manera que este vector sea el descriptor de apariencia global a utilizar [8]. Básicamente, la teoría para calcular dicho descriptor es la siguiente: el punto de partida es una imagen panorámica expresada como una matriz bidireccional a la cual, posteriormente, se aplicarán los cálculos matemáticos específicos para obtener un vector que caracteriza la imagen original.

Las redes neuronales convolucionales están estructuradas en tres partes claras. La primera consiste, básicamente, en una capa de entrada; la segunda, de varias capas ocultas intermedias; y, finalmente, la última parte está compuesta por una capa de salida. Dentro de las capas intermedias tenemos dos grupos claramente diferenciados: la primera fase, que consiste en capas para el aprendizaje de características; y la segunda fase, en la que hallamos capas que se encuentran completamente conectadas (*fully connected*) y generan vectores que proporcionan información para la clasificación de la imagen.

En el presente trabajo han sido utilizadas las CNN *Places*, *AlexNet* y *GoogLeNet*; las cuales serán ampliamente detalladas en el capítulo 4. No obstante, para que se pueda entender correctamente el funcionamiento de cómo

se estructura una red neuronal convolucional, visualizando lo expuesto hasta ahora, se presenta en la figura 2.2 la CNN *Caffe*. En esta imagen vemos, por un lado, las capas C1 – C5, que son capas convolucionales (denominadas en posteriores capítulos como conv); y, por otro lado, las capas FC6 – FC8, las cuales son capas *fully connected* (fc). Estas primeras capas, que son el objeto de estudio del presente trabajo, aplican varios filtros convolucionales a las imágenes de entrada con el objetivo de activar ciertas características de la imagen. Finalmente, se obtiene un descriptor de estas capas seleccionando una imagen del conjunto de datos de salida y reestructurando dicha información en un único vector, ya que es obtenida en formato matricial; es decir, si el tamaño de la imagen de salida es m × m, dependiente m de la capa convolucional utilizada, el tamaño del descriptor tendrá la siguiente estructura: $1 \times m^2$.

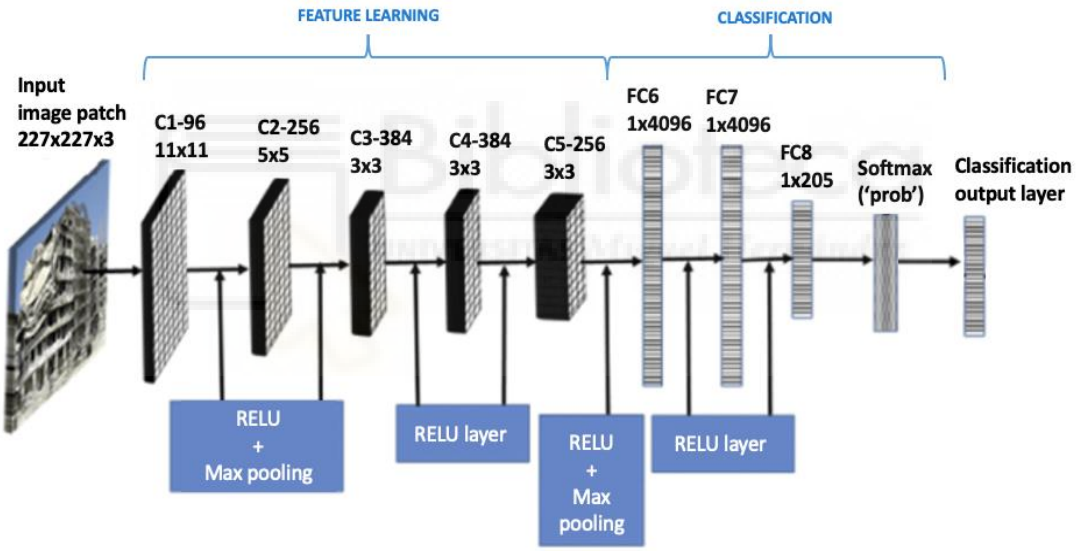


Figura 2.2: Diseño de arquitectura CNN del modelo pre-entrenado *Caffe*.

Asimismo, se ha demostrado cómo las primeras capas de las CNN logran una mejor precisión en la tarea de localización, haciendo hincapié en ellas debido a que su uso para la obtención de descriptores de apariencia global es escaso hoy en día a pesar de que muestran resultados muy favorables, tal como podemos ver en el trabajo de Cebollada et al. [7] con las capas 'conv4' y 'conv5'. Este comportamiento es razonable, ya que el objetivo de las primeras capas en una red neuronal convolucional es obtener información característica global de las imágenes, mientras que sus últimas capas se centran únicamente en

optimizar la tarea de clasificación. Teniendo en cuenta que estudios en el pasado han hecho únicamente uso de las capas *fully connected* (fc) para obtener descriptores de apariencia global, el presente trabajo, y basándonos en estudios previos, ha decidido investigar acerca de si capas convolucionales (conv) anteriores a la 'conv4' también pueden ser utilizadas para dicho propósito.

Por último, como se ha expuesto, los descriptores de apariencia global basados en herramientas de *deep learning* son más útiles a la hora de realizar tareas de localización, pero se debe considerar que la apariencia global de las imágenes depende mucho de las condiciones de iluminación. Sin embargo, se cree que la descripción global de la imagen no va a tener grandes variaciones con respecto a cambios en las condiciones lumínicas ni con pequeñas modificaciones en la escena como cambios en la posición de los objetos y puertas, pequeñas oclusiones, etc [29]. Es por ello por lo que parte de este trabajo se dedicará a realizar un estudio de este fenómeno comprobando si realmente es factible utilizar estos descriptores para resolver de una forma óptima la tarea de localización de un robot móvil, tal como podremos ver en los principales experimentos realizados en el capítulo 4.

Biblioteca
UNIVERSITAS Miguel Hernández

Capítulo 3
Material y métodos

En este capítulo se expondrá de forma detallada, por una parte, la información más relevante acerca de la base de datos empleada junto a los materiales que han sido utilizados; y, por otra parte, la estructuración de las diferentes redes neuronales convolucionales que han sido usadas para la realización de este trabajo fin de grado.

3.1. Base de datos empleada

Por un lado, como ya vimos en la figura 2.1, tenemos la plataforma robótica móvil empleada para la adquisición de imágenes en el laboratorio de Saarbrücken, espacio en el cual se basa nuestro estudio. Esta estaba equipada con escáneres láser tipo SICK, que proporcionan una medida de distancia escalar del sensor al objeto con una precisión y velocidad de escaneo óptima; y encoders en las ruedas tipo SICK, que miden el giro de las mismas y permiten calcular la posición y velocidad del robot móvil, lo cual hace posible una correcta odometría. Estos sensores posibilitan la obtención del *ground truth* (posición real) del robot en cada instante, a efectos comparativos, aunque en el presente trabajo el problema de localización se resolverá considerando únicamente información visual. Durante su estancia en el laboratorio el robot estaba controlado manualmente mediante un joystick.

La configuración de la cámara se creó mediante dos cámaras digitales *Videre Design* MDCS2; una para imágenes perspectiva y otra para imágenes omnidireccionales. Los parámetros detallados y las configuraciones de las cámaras se pueden observar en la tabla 3.1. El sistema de visión omnidireccional catadióptrico fue construido usando un espejo hiperbólico. Las dos cámaras y el

espejo se montaron juntos en un soporte portátil tal como se mostró en la figura 2.1 [39].

Robot móvil modelo ActivMedia PeopleBot Saarbrücken		
Tipo de cámara	Perspectiva	Omnidireccional
Ratio	5 imágenes / segundo	
Resolución	640 × 480 píxeles, Patrón de Bayer	
Exposición	Automática	
Campo de visión	68.9º × 54.4º	----
Altura de la cámara	140 cm	116 cm

Tabla 3.1: Parámetros y configuración de las cámaras.

Por otro lado, la base de datos con la que se ha decidido trabajar es la denominada como COLD *database* [39]. Esta base de datos ofrece tres entornos de trabajo diferentes: el *Autonomus Intelligent Systems Laboratory* en la universidad de Friburgo, Alemania; el *Visual Cognitive Systems Laboratory* en la universidad de Ljubljana, Eslovenia; y el *Language Technology Laboratory* en el Centro Alemán de Investigación en Inteligencia Artificial de Saarbrücken, Alemania. Sin embargo, de una forma más conocida, estos tres *dataset* que componen nuestra base de datos son denominados por el nombre de la ciudad donde la adquisición de las imágenes fue realizada (COLD – Saarbrücken, COLD – Friburgo y COLD – Ljubljana). La base de datos COLD es un banco de pruebas ideal, compuesto por imágenes omnidireccionales, para evaluar la robustez de los algoritmos de localización y *mapping*. Esta ofrece no solo cambios de iluminación, ya que la apariencia visual de las distintas habitaciones, en especial de las regiones que se encuentran cercanas a las ventanas, son fuertemente afectadas por las condiciones lumínicas y climáticas; sino también cambios dinámicos como la actividad humana, la variación en la distribución y aspecto del mobiliario, etc.

La elección del conjunto de imágenes a utilizar para la realización de este estudio ha sido el denominado como Saarbrücken. Esta decisión fue tomada

debido a que este *dataset* había sido menos estudiado y, por tanto, con él se podía demostrar si los descriptores de apariencia global con los que se había estado trabajando en COLD – Friburgo, como en el trabajo de Román [29], también otorgaban buenos resultados a la hora de utilizar otro conjunto diferente de imágenes.

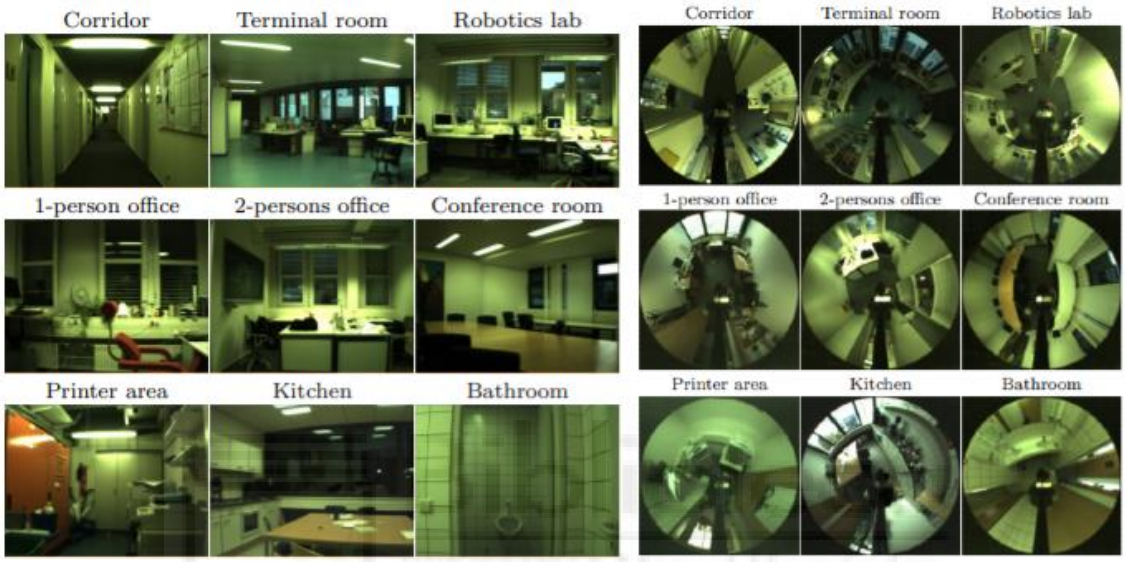
Las secuencias de imágenes en COLD – Saarbrücken se adquirieron en diferentes condiciones de iluminación, en un lapso de tiempo que varía en torno a dos/tres días. Las distintas estancias que recorrió el robot móvil para la adquisición de las imágenes se muestran en la tabla 3.2. Asimismo, un ejemplo de las imágenes capturadas durante este proceso puede ser visualizado en la figura 3.1, tanto en formato perspectiva como en omnidireccional. Como se puede observar, las habitaciones tienen diferentes propósitos y por ello estarán organizadas de una forma distinta, por lo que no serán afectadas por la actividad humana de la misma manera. A su vez, este *dataset* contiene también imágenes borrosas y que no proporcionan mucha información debido a la posición de adquisición en la que fueron tomadas. Todas estas desventajas hacen que este conjunto de imágenes sea adecuado para llevar a cabo experimentos en condiciones reales de funcionamiento, característica ideal para la realización de este estudio.

Estancias COLD – Saarbrücken	
Corridor	Printer area
Terminal room	Kitchen
Robotics lab	Bath room
1 - person office	2 - persons office
Conference room	

Tabla 3.2: Estancias recorridas por el robot móvil, en el laboratorio de Saarbrücken, mientras se realizaba la adquisición de las imágenes.

En la figura 3.2 se muestran los mapas generales de los entornos de interior de las dos partes diferentes del laboratorio que fueron consideradas por

separado. Estas pueden verse como dos entornos individuales similares, los cuales están denotados como ‘Parte A’ y ‘Parte B’. Como se puede comprobar, cada una de las partes está compuesta por dos rutas distintas; una de color rojo, en la que el recorrido del robot móvil es el más extenso, obteniendo así un *dataset* mucho más amplio, y otra de color azul, que aporta el recorrido más corto al no incluir todas las estancias del laboratorio.

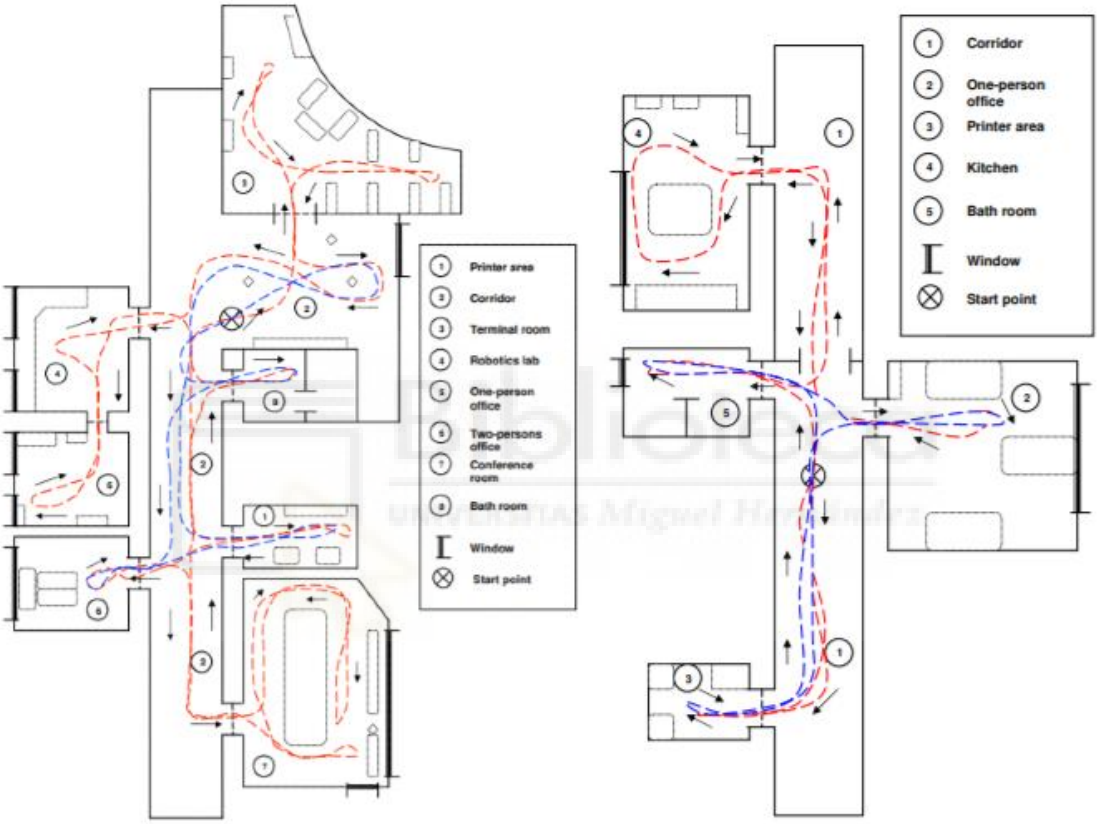


(a) (b)

Figura 3.1: Imágenes adquiridas en el laboratorio de Saarbrücken. La figura (a) muestra las imágenes tomadas mediante la cámara perspectiva, mientras que la figura (b) lo hace mediante la cámara omnidireccional. Imágenes obtenidas de COLD *database* [39].

Se ha tomado la decisión de trabajar con la ‘Parte B’ del mapa del laboratorio de Saarbrücken, ya que era el único de los dos que contenía información acerca de los tres tipos de iluminación que necesitábamos: nublado, soleado y noche. Asimismo, se ha tomado la trayectoria de color rojo, la cual nos proporciona mucha más información del entorno. Por otro lado, cabe señalar que la velocidad media de este robot móvil es de 0.3 m/s y la distancia que obtenemos entre imágenes, teniendo en cuenta que este está continuamente adquiriendo información a una velocidad de 5 fotogramas por segundo, es de aproximadamente 6 cm.

La base de datos COLD también nos ofrece la estimación de la pose del robot, tanto la posición ‘x - y’ como el ángulo de giro, para cada imagen adquirida mediante un sensor láser durante el recorrido. Esta información servirá de ayuda para estudiar el error cometido en la tarea de localización, ya que dispondremos de la posición real dónde se adquirió la imagen y de la posición estimada en la que está localizado el robot móvil. Este proceso será explicado con más detalle en el siguiente capítulo.



(a) Map of Saarbrücken Portion A (b) Map of Saarbrücken Portion B

Figura 3.2: Recorrido del robot móvil en el laboratorio de Saarbrücken. Imágenes obtenidas de COLD *database* [39].

El primer paso para poder trabajar con nuestra base de datos es tratar las imágenes de COLD – Saarbrücken, para ello se hará una conversión de formato, pasando de imagen omnidireccional RGB a panorámica en escala de grises. Esto se hace para que los datos obtenidos mediante descriptores de apariencia global basados en *deep learning*, que pueden trabajar con imágenes omnidireccionales RGB sin ningún tipo de problema, puedan ser comparados

con los datos hallados por descriptores holísticos basados en modelos analíticos, los cuales han sido desarrollados con imágenes panorámicas en escala de grises.

Esta conversión se realiza tomando la imagen omnidireccional en color (RGB) que obtenemos directamente del *dataset* de Saarbrücken, la cual se presenta en la figura 3.3, para posteriormente recortarla y eliminar toda la información posible de los bordes de la imagen que no sea de interés, quedando esta tal como podemos observar en la figura 3.4. Seguidamente, se ha implementado una función que transforma la imagen omnidireccional recortada en color en una imagen panorámica en escala de grises, únicamente introduciendo como parámetros de entrada la imagen omnidireccional recortada RGB y el centro de esta, el cual será el mismo para todas las imágenes a convertir en este trabajo. El resultado final de la conversión de la imagen queda reflejado tal como vemos en la figura 3.5. Dado que las imágenes RGB están formadas por tres capas de intensidad, con valores comprendidos entre 0 - 255, y las imágenes en escala de grises por tan solo una de ellas, podemos ver reducido considerablemente el coste computacional al trabajar con estas.

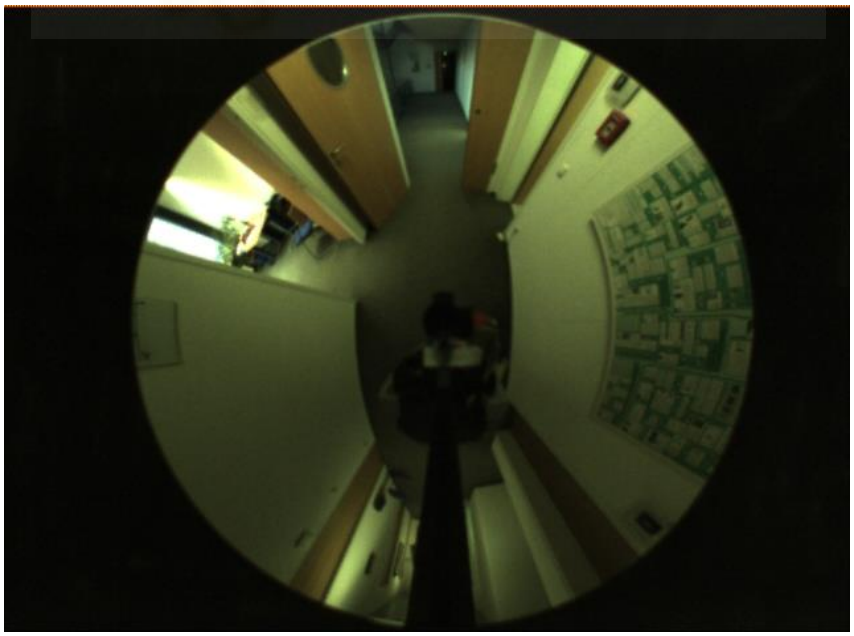


Figura 3.3: Imagen omnidireccional RGB.

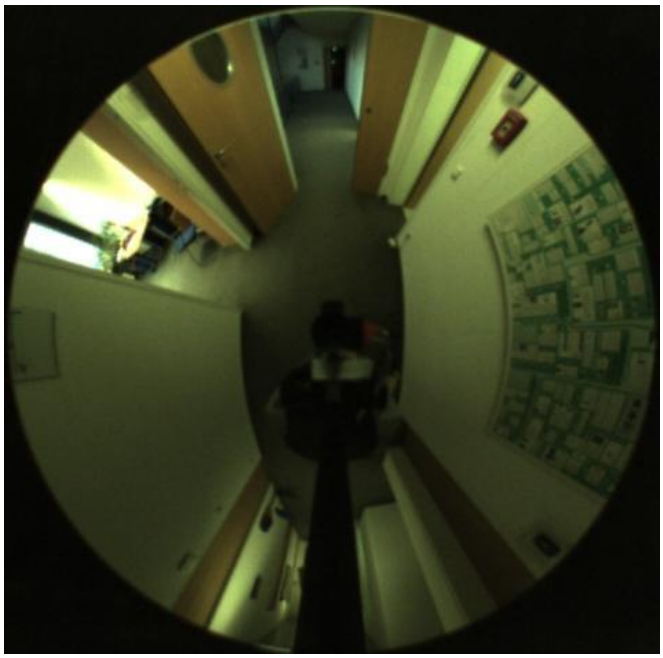


Figura 3.4: Imagen omnidireccional recortada RGB.

Figura 3.5: Imagen panorámica en escala de grises.

Una vez se ha realizado esta conversión para todas las imágenes de nuestra base de datos –nublado, soleado y noche (imágenes test)–, obtendremos un cuarto *dataset* de entrenamiento, el cual sería el obtenido de haber realizado la tarea de *mapping*, y con el que compararemos las imágenes test en nuestra tarea de localización. Para la creación de este modelo se decidió coger una de cada cinco imágenes del *dataset* adquirido durante un día nublado, ya que este es el que resulta menos afectado por los cambios de iluminación del entorno, de este modo nuestro modelo está creado por la adquisición de una imagen cada 30 cm, aproximadamente. La tabla 3.3 muestra la información acerca de los detalles de los cuatro *dataset* con los que realizaremos nuestros experimentos en el capítulo 4. Finalmente, una vez llegados a este punto, ya es posible comenzar a trabajar con los algoritmos de descripción de apariencia global de las imágenes.

Dataset	**Número de imágenes**
Test (nublado)	1064
Test (soleado)	1036
Test (noche)	1027
Entrenamiento (nublado)	213

Tabla 3.3: Número de imágenes de cada *dataset*, tanto de entrenamiento como de test, creado a partir del entorno COLD – Saarbrücken.

3.2. Redes neuronales utilizadas

Las redes neuronales convolucionales, comúnmente conocidas como CNN, son actualmente la herramienta más popular entre las técnicas de aprendizaje profundo, ya que han dado resultados exitosos en muchas aplicaciones prácticas. Estas son un tipo especializado de red neuronal para procesar datos que presentan una topología ya conocida, siendo comúnmente diseñadas para recibir imágenes como parámetros de entrada y teniendo diferentes aplicaciones como la clasificación o la detección de objetos.

Las CNN consisten en conexiones locales entre neuronas y transformaciones jerárquicamente organizadas de los datos. Básicamente, estas redes están compuestas principalmente por tres tipos de capas: convolucionales (conv), de agrupación (*pooling*) y completamente conectadas –*fully connected* (fc)–. Cada una de estas transforma la entrada y genera una salida de acuerdo con los parámetros establecidos. Este proceso de conexión se aborda mediante la transferencia de información a través de todas sus capas y finaliza en la última de ellas, la cual es una capa completamente conectada que genera un vector de características 1D, que proporciona la predicción más probable.

Las redes neuronales convolucionales que han sido usadas en este trabajo son la CNN *Places*, la CNN *AlexNet* y la CNN *GoogLeNet*. Todas ellas son utilizadas directamente con el entrenamiento previo realizado por los creadores, por lo tanto, no es necesario volver a entrenarlas para obtener descriptores holísticos de sus capas. A continuación, estas serán detalladas estructuralmente.

Por un lado, la CNN *AlexNet* fue presentada por Krizhevsky et al. [16]. Esta red fue entrenada con alrededor de 1.2 millones de imágenes para clasificar 1000 posibles tipos de objetos. Asimismo, de esta red neuronal convolucional fue creada la conocida CNN *Caffe*. En la figura 3.6 hemos tratado la arquitectura de *AlexNet* de una forma detallada para que con esta información se pueda entender de forma correcta cómo está organizada internamente. Esta red neuronal convolucional consta de 25 capas diferentes, entre ellas una capa de entrada, la cual toma una imagen RGB de 227 × 227 × 3, por lo tanto, cada imagen se debe normalizar antes del entrenamiento y/o clasificación; ocho capas intermedias, cinco de ellas convolucionales y tres completamente conectadas; tres capas de agrupación; un *softmax* final de 1000 categorías posibles de objetos como, por ejemplo, un teclado, un bolígrafo, una mesa o una variedad de animales; y una capa de salida, que tiene la función de indicar qué clase de objeto tenemos como imagen.

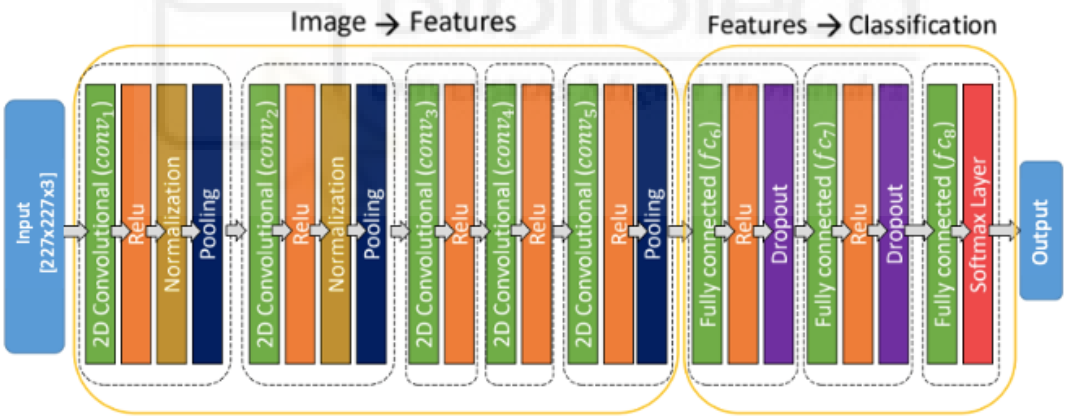


Figura 3.6: Diseño de arquitectura CNN del modelo pre-entrenado *AlexNet*.

Las capas convolucionales son las únicas de la CNN que están compuestas por canales, esto quiere decir que están formadas a su vez por múltiples capas internas; todas ellas han sido objeto de estudio en el presente trabajo. Las capas utilizadas de esta red neuronal convolucional en los experimentos del capítulo 4 se detallan en la tabla 3.4 junto con los canales que las componen.

Nombre de la capa	Canales
conv1	96
conv2	256
conv3	384
conv4	384
conv5	256
fc6	–

Tabla 3.4: Capas utilizadas de la CNN *AlexNet*.

Por otro lado, la red neuronal convolucional *Places* [44], que presenta una arquitectura similar a la ya vista en la figura 3.6 por la CNN *AlexNet*, fue entrenada con alrededor de 2.5 millones de imágenes para clasificar 205 posibles tipos de escenas. Asimismo, esta red nace de la CNN *Caffe*, la cual fue entrenada para clasificar diferentes tipos de objetos, por lo que ambas comparten la misma arquitectura de red y tienen fortalezas complementarias en tareas centradas tanto en escenas como en objetos. La red neuronal convolucional *Places* está compuesta por una capa de entrada, la cual toma una imagen RGB de 227 × 227 × 3, por lo tanto, cada imagen se debe normalizar antes del entrenamiento y/o clasificación; cinco capas convolucionales, las cuales tienen una función de aprendizaje o caracterización; tres capas totalmente conectadas, que tienen una labor de clasificación; tres capas de agrupación; una capa *softmax*, que indica la probabilidad de que la imagen de entrada corresponda a cada una de las 205 estancias posibles; y una capa de salida, la cual otorgará el lugar probable en el que se haya tomado la imagen.

Las capas utilizadas de esta red neuronal convolucional en los experimentos del capítulo 4 se detallan en la tabla 3.5 junto con los canales que las componen.

Nombre de la capa	Canales
conv1	96
conv2	256
conv3	384
conv4	384
conv5	256
fc6	–
fc7	–
fc8	–
prob	–

Tabla 3.5: Capas utilizadas de la CNN *Places*.

Finalmente, la CNN *GoogLeNet* fue propuesta por Szegedy et al. [37]. Esta red está compuesta por 144 capas, de las cuales 22 son convolucionales, y, al igual que la CNN *AlexNet*, también está entrenada para la clasificación de objetos, pero con la diferencia de que usa 12 veces menos parámetros que ella y es 6 veces más grande. Asimismo, a diferencia de las otras dos redes, esta arquitectura toma como entrada una imagen RGB de 224 × 224 × 3, por lo tanto, cada imagen se debe normalizar antes del entrenamiento y/o clasificación. Como podemos observar en la figura 3.7, esta red neuronal convolucional tiene una estructura muy compleja, nada que ver con las estudiadas hasta el momento, es por ello por lo que únicamente se ha tomado una parte de estas capas para nuestro estudio, las correspondientes a la parte más inicial de esta CNN, las que tienen puramente una función de aprendizaje de características (*feature learning*). Estas capas, que han sido utilizadas en los experimentos del capítulo 4, se detallan en la tabla 3.6 junto con los canales que las componen.

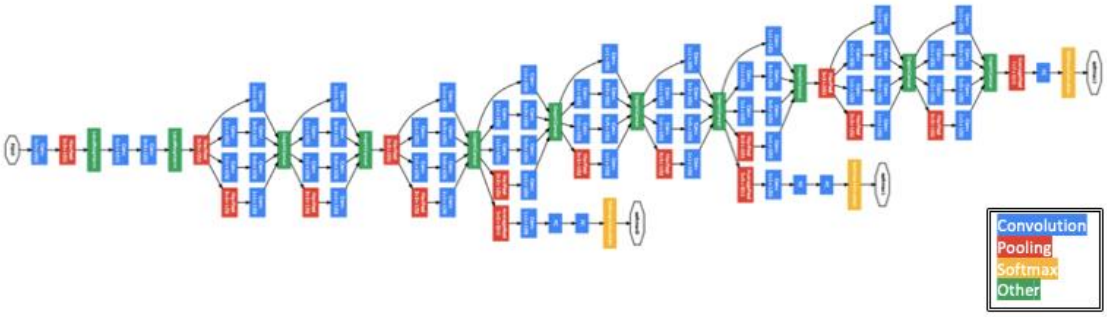


Figura 3.7: Diseño de arquitectura CNN del modelo pre-entrenado *GoogLeNet*.

Nombre de la capa	Canales
conv2-3x3	192
inception_3a-1x1	64
inception_3b-1x1	128
inception_4a-1x1	192
inception_4b-1x1	160
inception_4c-1x1	128
inception_4d-1x1	112
inception_4e-1x1	256

Tabla 3.6: Capas utilizadas de la CNN *GoogLeNet*.

La idea de estudiar estas tres redes neuronales convolucionales es que podamos conocer cuál de ellas, trabajando en un entorno de interior, estima con mayor precisión la posición en la que se encuentra el robot móvil. Es por ello por lo que presentamos un estudio sobre el uso de las capas intermedias de estas redes neuronales convolucionales para calcular descriptores holísticos y resolver la tarea de localización mediante el problema de recuperación de la imagen (*image retrieval*), es decir, recuperar la posición donde se obtuvo la imagen de entrada.

Capítulo 4
Experimentos

En este capítulo se estudia la eficacia de los descriptores de apariencia global en la tarea de localización de un robot móvil mediante la comparativa entre los múltiples descriptores obtenidos por las distintas CNN seleccionadas en este trabajo. Asimismo, para la realización de este estudio, se han hecho seis experimentos para analizar la robustez de estos descriptores ante diferentes efectos aplicados al modelo visual. Estos son un cambio de variación en la iluminación del entorno, una perturbación en la imagen debido a efectos como el *blur*, el ruido o las oclusiones; una variación de la estructura de la imagen a la entrada de la red neuronal convolucional y una posible rotación de la imagen debido a la orientación con la que el robot hubiese tomado la instantánea.

4.1. Localización visual

La tarea de localización consiste en un problema de *image retrieval*, en el cual se obtiene la imagen que representa una mayor similitud en relación con la nueva imagen capturada. Para dicho propósito, previamente se ha tenido que realizar un proceso de *mapping* en el que el robot ha obtenido información visual del medio, es decir, descriptores de apariencia global que se calculan a partir de las imágenes de entrenamiento capturadas desde diferentes posiciones del entorno.

Una vez que se ha construido la representación del medio, es necesario dotar de funcionalidad a dicho modelo visual, comprobando si permite al robot móvil realizar correctamente el proceso de localización mediante la estimación de su posición a través de la comparación de la información visual que captura en un instante determinado con la información almacenada en el modelo. Sin

embargo, en los entornos de interior que estudiaremos en este trabajo fin de grado, se plantea el problema fundamental de similitud visual de zonas que se encuentran geométricamente alejadas, pero que el robot móvil estima en una misma región; este problema se conoce como *visual aliasing*. Además, para que el robot pueda llevar a cabo tareas de forma autónoma en dicho entorno, resulta fundamental que sea capaz de estimar su posición y orientación. De este modo, podrá realizar una planificación de trayectorias y será capaz de calcular las acciones de control necesarias que le lleven al punto o puntos de destino.

El problema de localización se puede plantear de manera absoluta, comparando el descriptor de la imagen capturada por el robot móvil con los descriptores almacenados en el modelo visual. En el presente trabajo estudiaremos los procesos de localización topológica global como un proceso de *image retrieval,* en el que la posición del robot móvil es la del vecino más cercano del mapa, sin hacer uso ni de la información métrica relativa a la posición en el plano de trabajo en que se capturaron las imágenes del modelo, ni de la información de la odometría interna del robot [25].

Una vez creado el modelo visual, la tarea de localización se resuelve mediante los siguientes pasos: (1) el robot captura una nueva imagen omnidireccional desde una posición desconocida, la cual denominaremos de test. (2) Esa imagen se transforma en panorámica y, seguidamente, se calcula el descriptor correspondiente. (3) Una vez que el descriptor está disponible, el robot calcula la distancia coseno entre el descriptor de test y cada descriptor del modelo visual. Cabe señalar que esta fue seleccionada en el trabajo de Cebollada et al. [5] como un método eficiente de distancia para los descriptores de apariencia global. Posteriormente, (4) un vector de distancias es obtenido, para esa posición desconocida, con tamaño igual al número de imágenes de entrenamiento. Finalmente, (5) el nodo que presenta la distancia mínima corresponde a la posición estimada del robot en el mapa.

Hecha ya una introducción en lo que respecta a la localización visual, procederemos a detallar nuestro propio proceso de localización realizado mediante la herramienta de trabajo Matlab.

Antes de empezar con el proceso, hay que definir los tres parámetros que utilizaremos para medir la eficacia de la tarea de localización realizada a lo largo de los seis experimentos llevados a cabo en el presente trabajo. En primer lugar, tenemos (1) el tiempo de cálculo del descriptor, el cual nos indica el tiempo que se necesita para hallar el descriptor correspondiente para cada una de las imágenes de nuestro *dataset*. Seguidamente, encontramos (2) el tiempo de cálculo de la estimación de la posición, que nos proporciona el tiempo necesario para obtener la posición estimada del robot móvil en nuestro modelo visual desde el momento en el que se dispone del descriptor holístico. Finalmente, tenemos (3) el error de localización, el cual nos indica la diferencia que presentan nuestros resultados entre la posición donde se capturó la imagen (ofrecida por el *dataset* a través de la información láser) y la posición estimada por el robot móvil mediante el método de localización visual propuesto.

Para empezar nuestro proceso de localización tenemos que recordar que, tal como expusimos en el capítulo 3, partimos de nuestro conjunto de imágenes test (nublado, soleado y noche) y nuestro *dataset* de entrenamiento. Todas estas imágenes han sido convertidas a panorámicas en escala de grises, siendo esta conversión ya detallada en dicho capítulo, con un tamaño final de la imagen de 128 × 512 × 1. También hay que recordar que tenemos la posición de cada una de las imágenes, tanto de test como de entrenamiento, debido a que esta información era proporcionada por la base de datos COLD.

La tarea de localización se ha llevado a cabo, primero que todo, tomando una a una las imágenes de entrenamiento y calculando sus respectivos descriptores para poder realizar, en pasos posteriores, la comparativa entre estos y los descriptores de las imágenes de test. Para esto hemos tenido que tomar individualmente cada una de las imágenes de entrenamiento y realizar un cambio de tamaño de la instantánea, ya que las redes neuronales convolucionales únicamente permiten una imagen de entrada de 227 × 227 × 3, en el caso de utilizar las CNN *Places* y *AlexNet*, y de 224 × 224 × 3, de tratarse de la CNN *GoogLeNet*.

Las redes neuronales convolucionales seleccionadas están diseñadas para trabajar únicamente con imágenes RGB, que contienen los tres canales de color, y es este número tres el que aparece junto al tamaño permitido por las CNN. No obstante, las imágenes en escala de grises están compuestas exclusivamente por un único canal de información, de ahí el número uno en el tamaño obtenido tras la conversión a nuestra imagen panorámica. Es por ello por lo que, para trabajar con estas imágenes, previamente hay que triplicar su único canal.

Una vez tenemos redimensionada correctamente la imagen panorámica ya podemos introducirla en la CNN y obtener un descriptor de ella. Nótese que al introducir la imagen a la red neuronal convolucional hay que indicar también con qué capa de esta vamos a extraer nuestro descriptor. El proceso de localización es el mismo para cualquier capa de la CNN que vayamos a utilizar. Es de esta forma, por tanto, como obtenemos todos los descriptores de las imágenes de entrenamiento.

A continuación, se presentarán las líneas de código utilizadas para obtener el descriptor. En ellas se ha ejemplificado el proceso para la CNN *AlexNet* y la capa de clasificación 'fc6'. No obstante, cabe señalar que este procedimiento es el mismo para todas las capas y redes neuronales convolucionales con las que hemos trabajado.

```
net = alexnet;
layer = 'fc6';
descriptor = activations(net,imagen,layer);
```

Llegados a este punto, comenzaremos a trabajar con las imágenes de test para obtener sus descriptores, realizando el mismo procedimiento explicado hasta el momento, ya que todas las imágenes serán tratadas de la misma forma. Por un lado, el tiempo de cálculo del descriptor es el comprendido entre el momento que introducimos la imagen a la CNN, junto con la capa de la cual queremos obtener el descriptor, hasta el momento en que la red neuronal convolucional nos devuelve el correspondiente descriptor en forma de vector. Por otro lado, el tiempo de cálculo de la estimación de la posición está

comprendido entre dos momentos clave. El primero se refiere a la comparación del descriptor de test con cada uno de los descriptores de las imágenes de entrenamiento. Esta, a su vez, conlleva a la formación de un único descriptor compuesto por distancias que contiene en cada una de sus posiciones sus respectivos resultados. Asimismo, cabe señalar que se utilizará la distancia '*cosine*' para realizar esta comparación con los descriptores obtenidos de todas las capas de las CNN salvo la *softmax* ('prob'), que necesitará de la distancia '*spearman*', pues esta capa trabaja en términos probabilísticos y, por tanto, este tipo de distancia es el más adecuado para ella. Por otra parte, el segundo momento clave que constituye el fin del tiempo de cálculo de la estimación de la posición tiene lugar cuando se halla el valor mínimo en alguna de las posiciones del vector constituido por distancias y se toma la posición 'x - y' de esta como la estimada por el robot móvil, lo cual se corresponde con el proceso de *image retrieval*. Finalmente, teniendo ya la posición real y estimada, podemos calcular el error de nuestra tarea de localización mediante la distancia '*euclidean*' entre estos dos resultados. Las diferentes distancias usadas en este trabajo son descritas en la tabla 4.1 presentada a continuación.

Distancia	**Descripción**
Euclidean	$d_{st}^2 = (x_s - x_t)(x_s - x_t)'$
Cosine	$d_{st} = 1 - \frac{x_s x'_t}{\sqrt{(x_s x'_s)(x_t x'_t)}}$
Spearman	$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}}$

Tabla 4.1: Tipos de distancias utilizadas entre los descriptores.

Terminaremos con esta tarea de localización almacenando el valor medio correspondiente a cada uno de estos tres parámetros, al finalizar de trabajar con cada conjunto de imágenes, para su representación mediante gráficas en cada uno de los experimentos realizados. Una vez obtenidas estas, realizaremos su

posterior comparación y discusión, con la intención de conocer qué descriptor presenta los mejores resultados en dicha tarea.

4.2. Experimento 1: Comparación entre descriptores basados en *deep learning* y descriptores basados en métodos analíticos

Como se ha explicado en anteriores capítulos, los trabajos previos con redes neuronales convolucionales únicamente han hecho uso de sus capas de clasificación para obtener información de los descriptores de apariencia global de las imágenes que describen el entorno. Estos han sido comparados con algunos descriptores holísticos basados en métodos analíticos, HOG y gist, para comprobar su eficacia. Es por ello por lo que, para comenzar nuestro estudio, hemos decidido partir de este mismo punto para así poder comparar nuestros resultados con estudios realizados anteriormente.

En este primer experimento, por un lado, se ha trabajado con las capas de clasificación de la red neuronal convolucional *Places* (‘fc6’, ‘fc7’, ‘fc8’, ‘prob’), las cuales están compuestas por un único canal de información; y, por otro lado, se han empleado también los descriptores holísticos HOG y gist, que tienen un tamaño de descriptor resultante de 1024 y 256, respectivamente. De esta forma podemos comparar los resultados en la tarea de localización entre descriptores de apariencia global basados tanto en métodos analíticos como en herramientas de *deep learning* cuando se aplican diferentes efectos de iluminación en el entorno.

La configuración elegida para poder trabajar con los descriptores HOG y gist es la que figura en el trabajo de Cebollada et al. [6] como aquella que optimiza los resultados de localización. Una vez se tiene clara la estructura de estos descriptores y las capas de las redes neuronales convolucionales que utilizaremos, hemos procedido a resolver la tarea de localización como se ha explicado al comienzo de este capítulo. A partir de ella, hemos obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal

como se muestra en la figura 4.1, y el tiempo medio total de cómputo en dicha tarea, como podemos ver en la figura 4.2, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición.

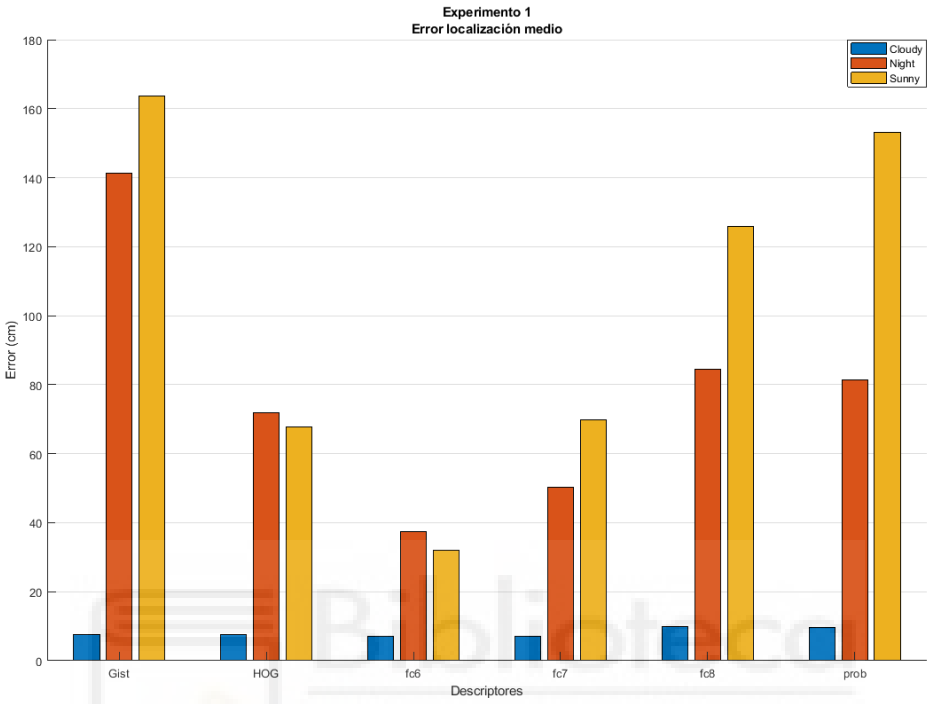


Figura 4.1: Error medio de localización de los descriptores obtenidos de las capas de clasificación de la CNN *Places* junto con los descriptores HOG y gist.

Por un lado, en lo que respecta al error medio de localización en el conjunto de imágenes de nublado, entre los descriptores de apariencia global basados en métodos analíticos, HOG obtiene mejores resultados que gist, como podemos observar en la figura 4.1. Asimismo, en los descriptores holísticos obtenidos mediante la CNN, podemos ver cómo el error de posicionamiento de nuestra tarea de localización en cada una de sus capas de clasificación es mayor en cualquiera de los *dataset* utilizados a medida que nos acercamos a la capa de salida de la red neuronal convolucional. Por otro lado, en lo que se refiere al tiempo medio total de cálculo, en la figura 4.2 observamos que, de los descriptores holísticos basados en métodos analíticos, es necesario un tiempo de cálculo notablemente mayor en el caso del descriptor gist en contraposición a HOG. A su vez, en los descriptores de apariencia global obtenidos mediante la

CNN, podemos observar cómo el tiempo medio de cálculo del descriptor es muy similar en todas sus capas, diferencia que se hace más notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es mucho mayor en las dos primeras, ‘fc6’ y ‘fc7’, ya que estas tienen un tamaño de descriptor de 4096 frente al de 205 que poseen las últimas dos capas de la CNN, ‘fc8’ y ‘prob’, que son las que menos tiempo necesitan para finalizar el proceso.

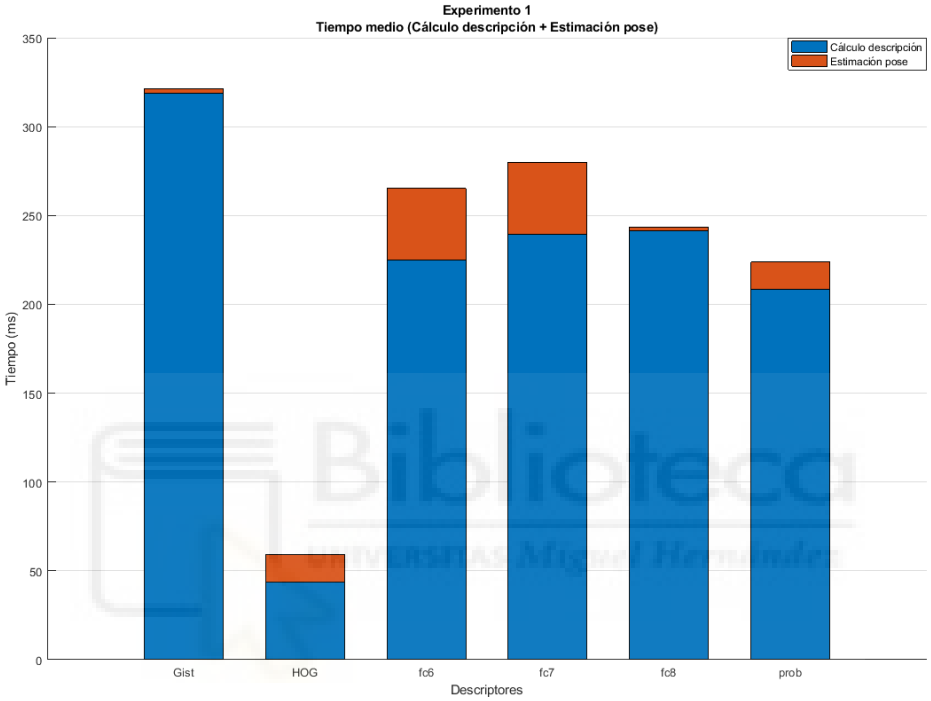


Figura 4.2: Tiempo medio total de cómputo de los descriptores obtenidos de las capas de clasificación de la CNN *Places* junto con los descriptores HOG y gist.

4.3. Experimento 2: Comparativa entre las diferentes capas de la CNN

En el siguiente experimento, al igual que en el anterior, estudiaremos la tarea de localización ante diferentes efectos de iluminación en el entorno. Para ello, haremos uso de descriptores holísticos obtenidos únicamente por las capas de la CNN *Places* pertenecientes a la fase de aprendizaje de características (*feature learning*), que está formada por las capas convolucionales ‘conv1’, ‘conv2’, ‘conv3’, ‘conv4’ y ‘conv5’. Estas están compuestas por varios canales y

no proporcionan a la salida un único vector, sino que devuelven una matriz por cada canal. Por tanto, no podemos utilizar directamente la salida de estas capas como descriptor holístico.

Como solución a este problema, investigaremos cuál de los canales de cada una de las capas nos proporciona un menor error de localización para, seguidamente, quedarnos con el resultante y reordenar la matriz obtenida en un vector. De esta manera, al igual que ocurre a la hora de trabajar con las capas de clasificación de la red neuronal convolucional, obtendremos un único descriptor de apariencia global. Para ello, trabajaremos de forma exhaustiva solamente con las imágenes de test del *dataset* nublado, ya que para los otros cambios de iluminación (tanto soleado como noche) estudiaremos únicamente la tarea de localización con el mejor canal que hemos hallado del conjunto de imágenes de nublado. Por tanto, una vez finalizado este estudio, podremos observar en la tabla 4.2 el canal que presenta un menor error de localización en nuestra tarea obtenido de cada una de las capas convolucionales de la CNN.

Nombre de la capa	**Mejor canal**	**Tamaño del descriptor**
conv1	69	3025
conv2	83	729
conv3	14	169
conv4	188	169
conv5	75	169

Tabla 4.2: Mejor canal obtenido para la tarea de localización en cada una de las capas convolucionales de la CNN *Places*.

Una vez aclarados los canales a estudiar en cada una de las capas de las redes neuronales convolucionales seleccionadas, hemos procedido a resolver la tarea de localización como se ha explicado al comienzo de este capítulo. Con ella, hemos obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal como se muestra en la figura 4.3, y el tiempo medio total

de cómputo en dicha tarea, como podemos ver en la figura 4.4, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición.

Asimismo, se han comparado los resultados obtenidos con los de la capa 'fc6', la cual fue seleccionada como la mejor para realizar la tarea de localización entre todas las capas de clasificación de esta red neuronal convolucional. De esta forma podemos contrastar los resultados de los descriptores holísticos basados en *deep learning* obtenidos a partir de capas convolucionales con los resultados de aquellos adquiridos a partir de las capas de clasificación cuando se aplican diferentes efectos de iluminación en el entorno.

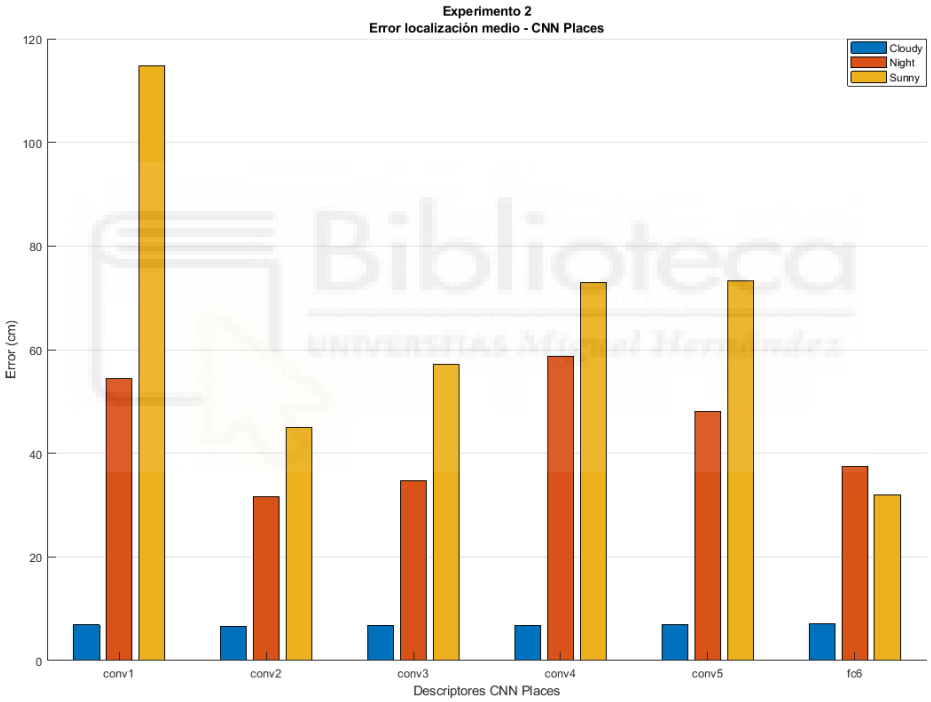


Figura 4.3: Error medio de localización de los descriptores obtenidos de las capas convolucionales de la CNN *Places* junto con el descriptor obtenido por la capa 'fc6'.

Por un lado, en lo que respecta al error medio de localización en el conjunto de imágenes de nublado, entre los descriptores holísticos hallados por las capas convolucionales, la 'conv2' obtiene los mejores resultados con respecto al resto, como podemos observar en la figura 4.3. Asimismo, también vemos cómo a partir de esta los errores de posicionamiento en nuestra tarea de localización se

incrementan en cualquiera de los *dataset* utilizados. En cuanto a la capa de clasificación ‘fc6’, los resultados que esta obtiene son muy similares a los de la capa convolucional ‘conv2’, aunque estos los mejoran cuando se utilizan las imágenes de un tiempo soleado. Por otro lado, en lo que se refiere al tiempo medio total de cálculo, en la figura 4.4 observamos que este aumenta en cada una de las capas convolucionales a medida que avanzamos a través de la CNN y nos acercamos a las de clasificación. A su vez, podemos observar cómo el tiempo medio de cálculo del descriptor también se incrementa a medida que pasamos de capa, diferencia que se hace más notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es mucho mayor en las dos primeras, ‘conv1’ y ‘conv2’, ya que estas tienen un mayor tamaño de descriptor frente al que poseen las últimas tres capas convolucionales de la CNN. En lo que se refiere a la capa de clasificación ‘fc6’, esta necesita de un tiempo medio de cálculo notablemente mayor a cualquier capa convolucional previa, debido a que está enfocada a un trabajo muy especializado y posee un tamaño de descriptor superior.

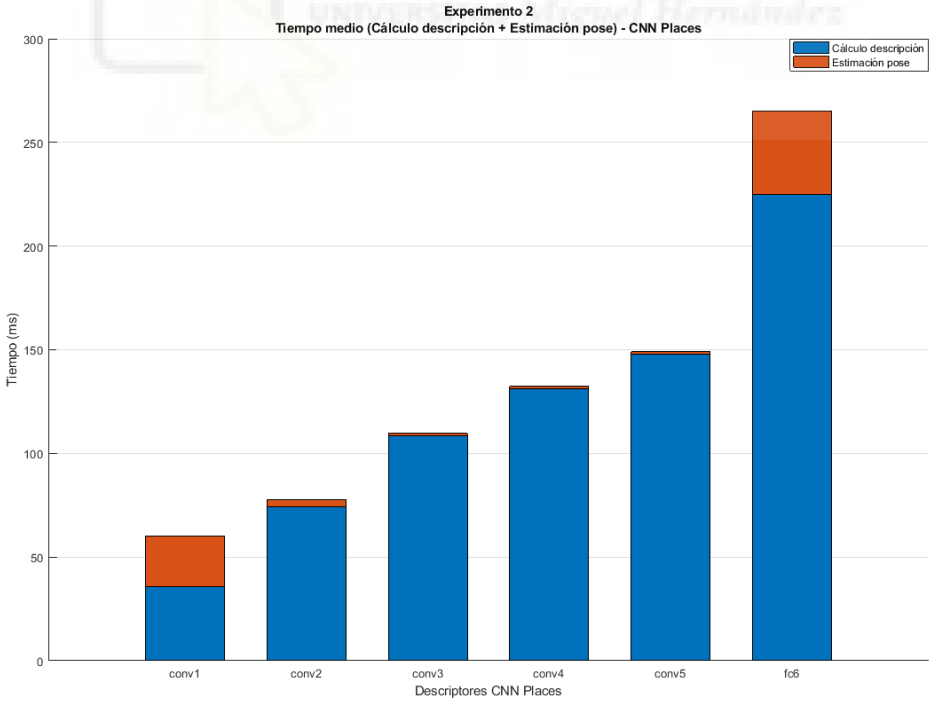


Figura 4.4: Tiempo medio total de cómputo de los descriptores obtenidos de las capas convolucionales de la CNN *Places* junto con el descriptor obtenido por la capa ‘fc6’.

4.4. Experimento 3: Comparación entre CNNs

En este experimento, estudiaremos la tarea de localización ante diferentes efectos de iluminación en el entorno, pero esta vez utilizaremos redes neuronales convolucionales diferentes. Para ello, haremos uso de descriptores holísticos obtenidos únicamente de las capas convolucionales de la CNN *AlexNet* y de algunas de las que componen la CNN *GoogLeNet*. El nombre de las capas utilizadas en el presente experimento junto con los canales que las conforman se presentó ya en la tabla 3.4 para la red neuronal convolucional *AlexNet* y en la tabla 3.6 para *GoogLeNet*. Se ha realizado con ellas el mismo procedimiento que se determinó en el experimento anterior para conocer qué canal de cada una de sus capas otorga mejores resultados en la posición del robot móvil.

Por tanto, una vez finalizado este estudio, podremos observar en la tabla 4.3 el canal obtenido de cada una de las capas convolucionales de la CNN *AlexNet* que presenta un menor error de localización en nuestro proceso. Asimismo, en la tabla 4.4, se exponen los resultados hallados para la red neuronal convolucional *GoogLeNet*.

Nombre de la capa	Mejor canal	Tamaño del descriptor
conv1	54	3025
conv2	134	729
conv3	220	169
conv4	79	169
conv5	223	169

Tabla 4.3: Mejor canal obtenido para la tarea de localización en cada una de las capas convolucionales de la CNN *AlexNet*.

Nombre de la capa	Mejor canal	Tamaño del descriptor
conv2-3x3	185	3136
inception_3a-1x1	3	784
inception_3b-1x1	85	784
inception_4a-1x1	116	196
inception_4b-1x1	4	196
inception_4c-1x1	21	196
inception_4d-1x1	53	196
inception_4e-1x1	236	196

Tabla 4.4: Mejor canal obtenido para la tarea de localización en ciertas capas convolucionales de la CNN *GoogLeNet*.

Una vez aclarados los canales a estudiar en cada una de las capas de las dos redes neuronales convolucionales seleccionadas, hemos procedido a resolver la tarea de localización como se ha explicado al comienzo de este capítulo. En primer lugar, se ha resuelto dicha tarea para la red neuronal convolucional *AlexNet*, en la cual se ha obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal como se muestra en la figura 4.5, y el tiempo medio total de cómputo en dicha tarea, como podemos ver en la figura 4.6, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición. Seguidamente, se ha realizado el mismo procedimiento para la CNN *GoogLeNet*. En este hemos obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal como se muestra en la figura 4.7, y el tiempo medio total de cómputo en dicha tarea, como podemos ver en la figura 4.8, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición.

Finalmente, se han comparado únicamente los resultados obtenidos de las mejores capas de estas dos redes neuronales convolucionales con el mejor descriptor seleccionado de la CNN *Places*, que fue calculado, en el anterior experimento, mediante la capa convolucional ‘conv2’. Con ello, hemos agrupado estos tres descriptores resultantes en una única gráfica que muestra el error

medio de posicionamiento, tal como se observa en la figura 4.9, y en otra que representa el tiempo medio total de cómputo necesario para poder llevar a cabo la tarea de localización, como podemos ver en la figura 4.10, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición. De esta forma podemos contrastar los resultados de diversos descriptores de apariencia global obtenidos de tres CNN diferentes cuando se aplican distintos efectos de iluminación en el entorno.

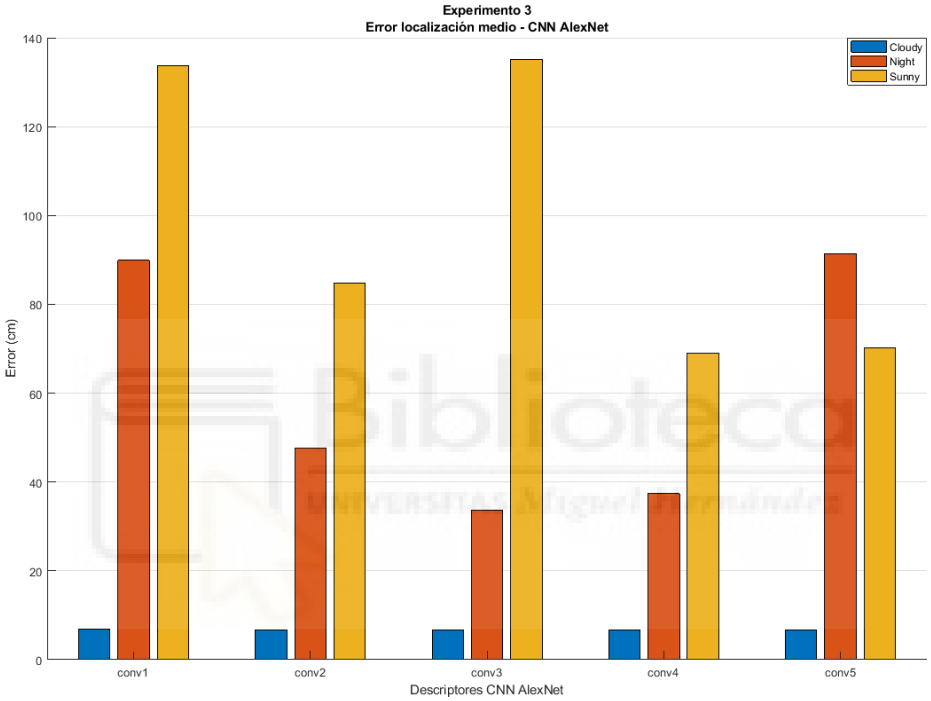


Figura 4.5: Error medio de localización de los descriptores obtenidos de las capas convolucionales de la CNN *AlexNet*.

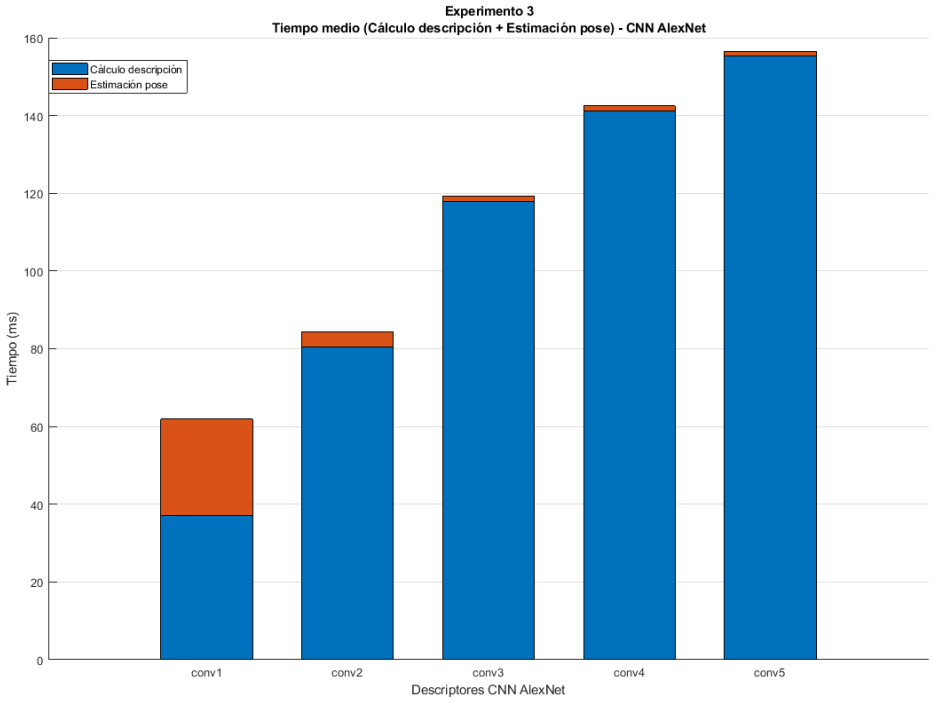


Figura 4.6: Tiempo medio total de cómputo de los descriptores obtenidos de las capas convolucionales de la CNN *AlexNet*.

Por un lado, con respecto a los datos obtenidos por la red neuronal convolucional AlexNet, en la figura 4.5 podemos observar que, entre los descriptores holísticos hallados por las capas convolucionales, la 'conv4' es la que posee los mejores resultados del error medio de localización en el conjunto de imágenes de nublado. Asimismo, también vemos cómo los errores de posicionamiento en nuestra tarea de localización, por parte del resto de las capas, se incrementan en cualquiera de los *dataset* utilizados. Por otro lado, en lo que se refiere al tiempo medio total de cálculo, en la figura 4.6 observamos que este aumenta en cada una de las capas convolucionales a medida que avanzamos a través de la CNN y nos acercamos a las de clasificación. A su vez, podemos ver cómo el tiempo medio de cálculo del descriptor también se incrementa a medida que pasamos de capa, diferencia que se hace más notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es mucho más elevado en las dos primeras, 'conv1' y 'conv2', ya que estas presentan un tamaño de descriptor mayor que el que poseen las últimas tres capas convolucionales de la CNN.

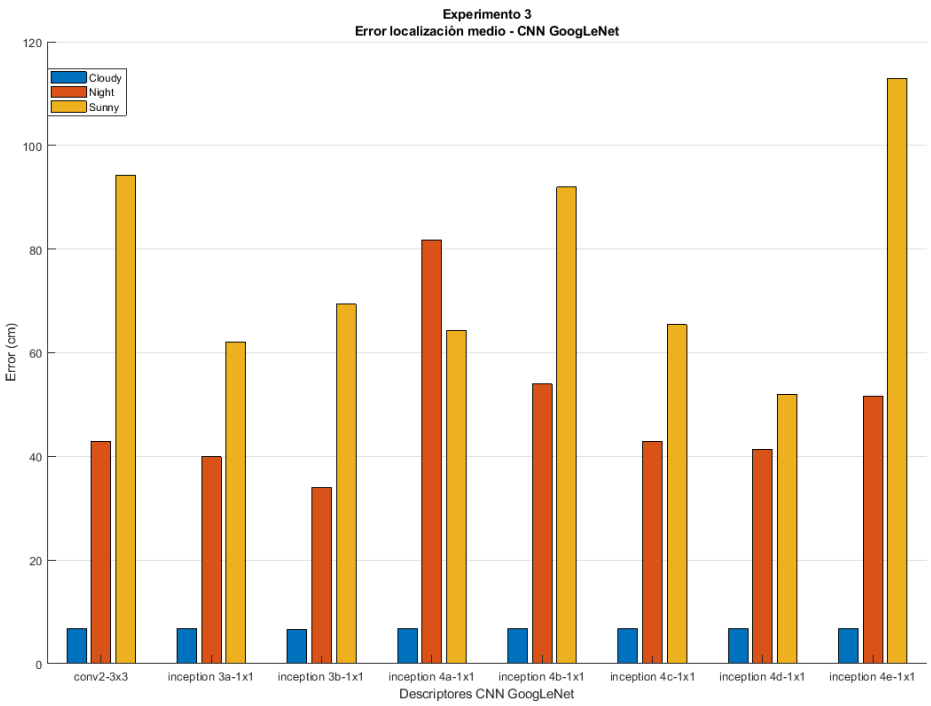


Figura 4.7: Error medio de localización de los descriptores obtenidos de ciertas capas convolucionales de la CNN *GoogLeNet*.

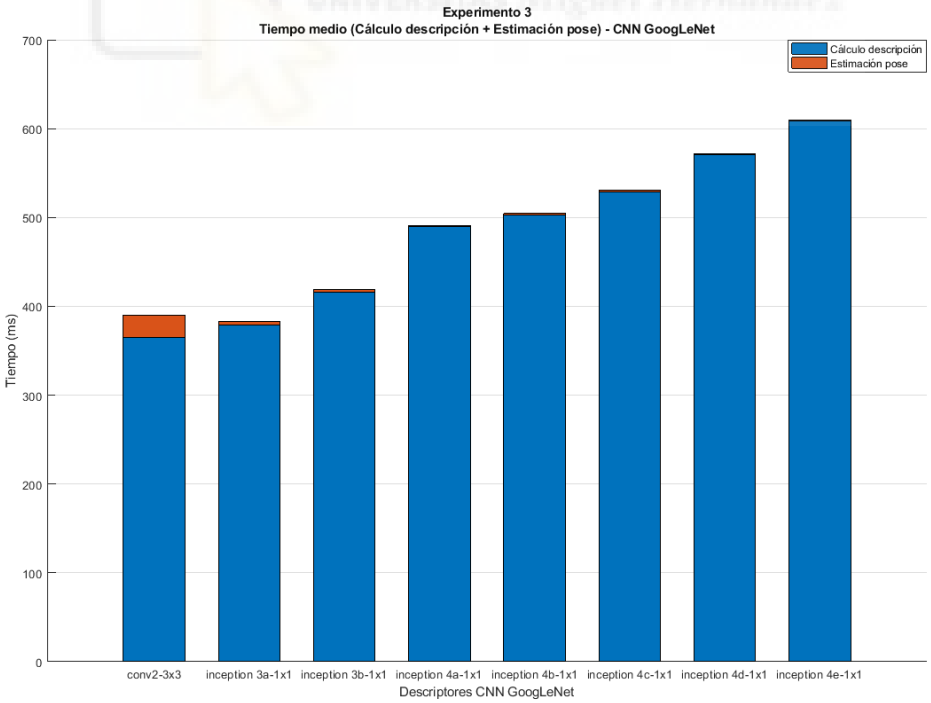


Figura 4.8: Tiempo medio total de cómputo de los descriptores obtenidos de ciertas capas convolucionales de la CNN *GoogLeNet*.

Por una parte, en lo que respecta a los datos obtenidos por la red neuronal convolucional GoogLeNet, la figura 4.7 muestra que, entre los descriptores holísticos hallados por las capas convolucionales, la ‘inception_3b-1x1’ posee los mejores resultados del error medio de localización en el conjunto de imágenes de nublado. Asimismo, también vemos cómo los errores de posicionamiento en nuestra tarea de localización, por parte del resto de las capas, se incrementan en cualquiera de los *dataset* utilizados. Por otra parte, en lo que se refiere al tiempo medio total de cálculo, en la figura 4.8 observamos que este aumenta en cada una de las capas convolucionales a medida que avanzamos a través de la CNN y nos acercamos a las de clasificación. A su vez, podemos observar cómo el tiempo medio de cálculo del descriptor también se incrementa a medida que pasamos de capa, diferencia que se hace más notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es mucho mayor en las tres primeras, ‘conv2-3x3’, ‘inception_3a-1x1’ e ‘inception_3b-1x1, ya que estas tienen un mayor tamaño de descriptor frente al que poseen las otras cinco capas convolucionales seleccionadas de esta CNN.

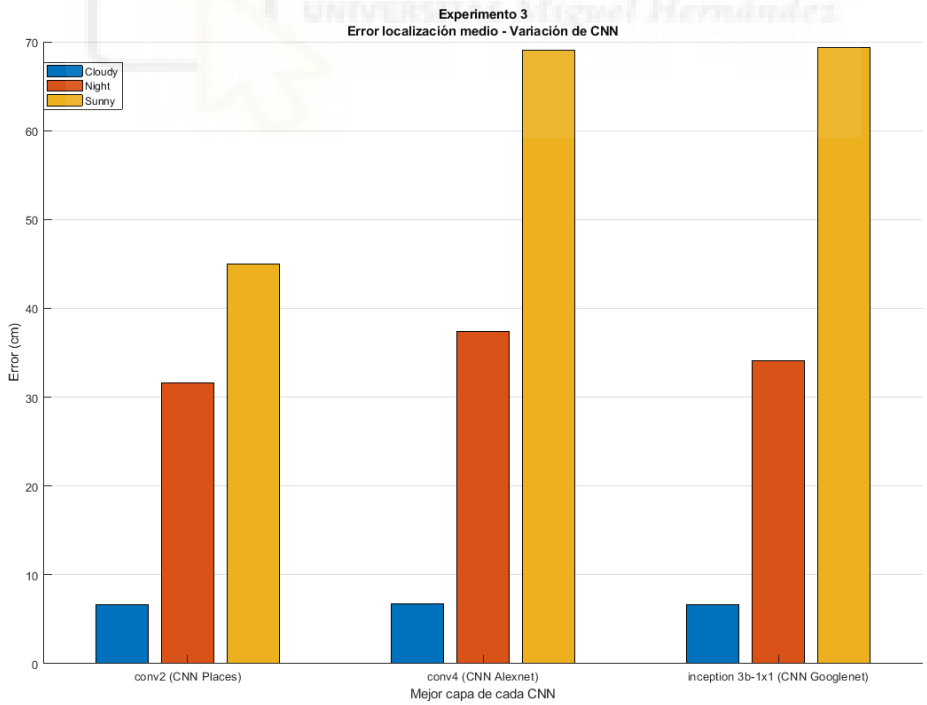


Figura 4.9: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places*, *AlexNet* y *GoogLeNet*.

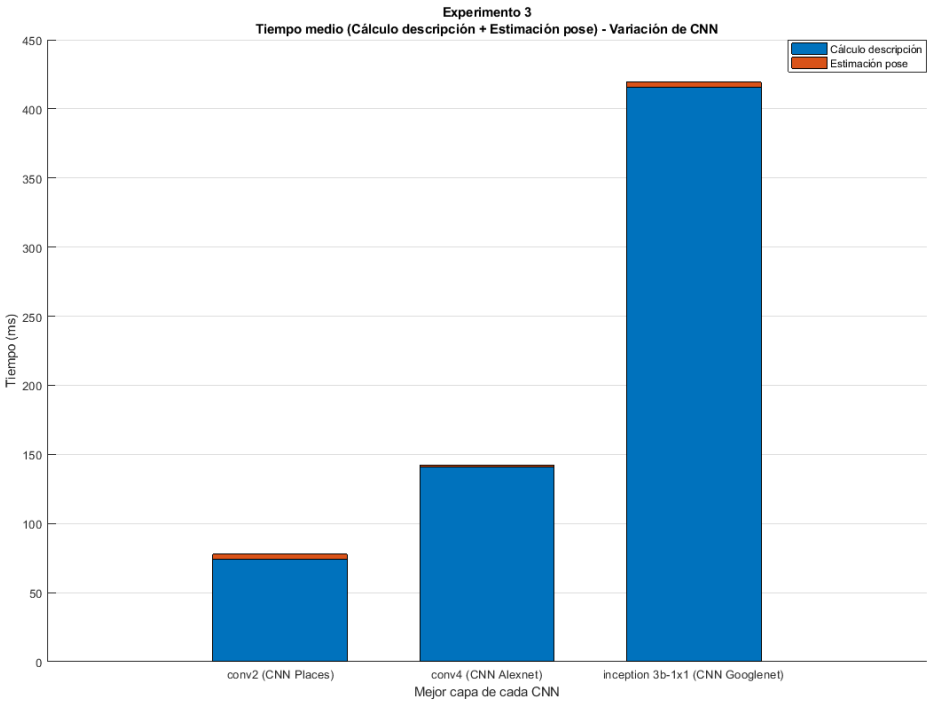


Figura 4.10: Tiempo medio total de cómputo de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places*, *AlexNet* y *GoogLeNet*.

Finalmente, una vez obtenidos los mejores descriptores de cada una de las tres redes neuronales convolucionales utilizadas en el presente experimento, en la figura 4.9 podemos ver que, entre los descriptores holísticos hallados por estas tres capas, la convolucional 'conv2' de la CNN *Places* obtiene los mejores resultados en lo que respecta al error medio de localización. Asimismo, también vemos, por parte de los dos descriptores obtenidos de las capas convolucionales de las CNN *AlexNet* y *GoogLeNet*, que los errores de posicionamiento en la tarea de localización son muy similares con respecto a los de la capa 'conv2' cuando se trabaja con los conjuntos de imágenes de nublado y noche, pero se incrementan en el *dataset* de soleado. A su vez, en lo que se refiere al tiempo medio total de cálculo, en la figura 4.10 observamos que la capa más adecuada vuelve a ser la seleccionada por la red neuronal convolucional *Places*. Esto es debido a que el tiempo de cómputo necesario para finalizar el proceso se duplica en la CNN *AlexNet* y se quintuplica al trabajar con *GoogLeNet*. De igual forma, podemos observar cómo el tiempo medio de cálculo del descriptor aumenta en

la capa ‘conv4’ con respecto a la ‘conv2’ y se hace mucho más grande en el caso de ‘inception_3b-1x1’, diferencia que se hace menos notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es similar entre las tres capas convolucionales.

4.5. Experimento 4: Localización visual frente a efectos visuales

Llegados a este punto, se han finalizado todos los estudios en nuestra tarea de localización correspondientes a la aplicación de efectos de iluminación en nuestro modelo visual de nublado, como se ha podido comprobar al trabajar con los *dataset* de soleado y noche. Es por ello por lo que en este experimento aplicaremos nuevos efectos a nuestras imágenes de test, los cuales pueden darse en situaciones reales de funcionamiento. Entre ellos, en primer lugar, cabe señalar el de oclusión, que simula los casos en que algunas zonas de la imagen están ocultas, ya sea por cómo está dispuesta la configuración del sensor en ese instante o por alguna otra circunstancia como, por ejemplo, que una persona esté frente al robot móvil. Este efecto se aplica al introducir objetos grises geométricos sobre partes aleatorias de la imagen. En segundo lugar, también encontramos el efecto *blur*, el cual produce, a su vez, un efecto de desenfoque cuando la imagen se captura mientras la cámara está en movimiento, obteniendo así una imagen borrosa. Por último, cabe mencionar el efecto de ruido, que agrega a la imagen un ruido de tipo Gaussiano.

La figura 4.11 muestra los ejemplos de estos tres efectos aplicados sobre una imagen de test. En ella podemos observar cuatro imágenes: la primera, obtenida directamente del conjunto de datos de nublado, es la original, mientras que las otras tres son el resultado de haber aplicado un determinado efecto visual a la anterior. Hay que tener en cuenta que dichos efectos se utilizan únicamente en imágenes que no poseen previamente ningún otro efecto visual, ya que no es óptimo para nuestro proceso aplicar varios de forma simultánea en una misma instantánea, porque lo que se busca con el presente trabajo es estudiar cómo se comporta nuestra tarea ante diferentes efectos visuales planteados de manera

individual. Es por ello por lo que estos siempre han sido aplicados al *dataset* de nublado.

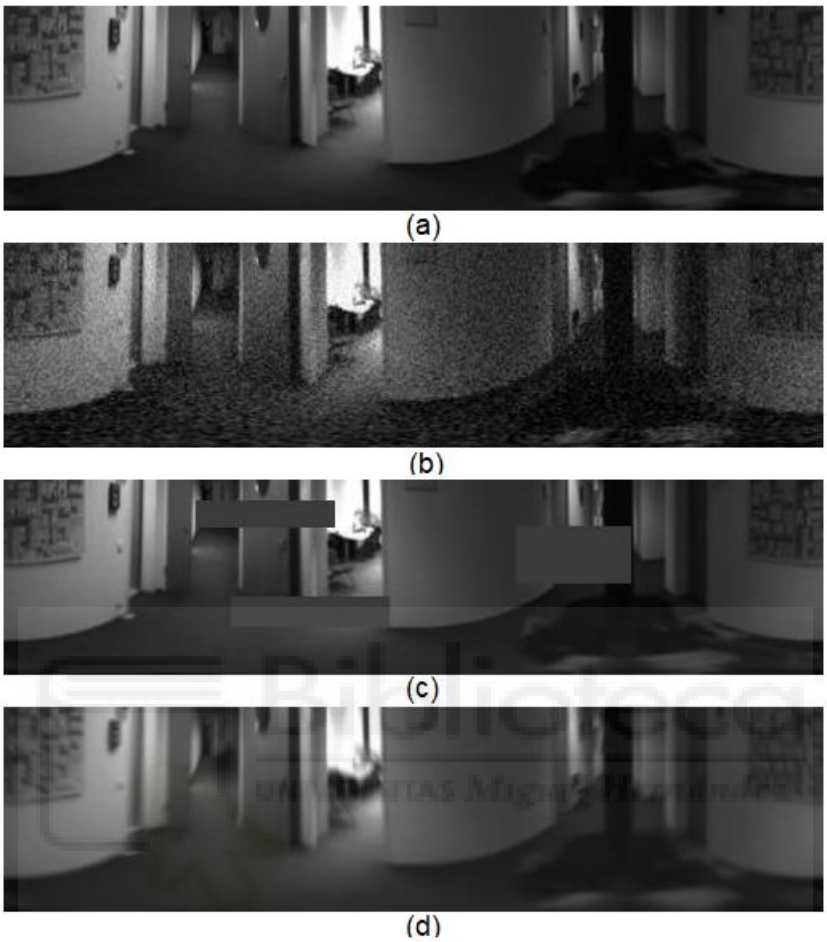


Figura 4.11: La figura (a) muestra la imagen original capturada dentro del entorno de Saarbrücken, mientras que diversos efectos aplicados a nuestro modelo visual se observan en las siguientes: (b) ruido Gaussiano, (c) oclusión y (d) efecto *blur*.

En este experimento, al igual que en los anteriores, realizaremos los mismos estudios en la tarea de localización, pero con la diferencia de que en este trabajaremos únicamente con los dos mejores descriptores obtenidos en experimentos previos de las redes neuronales convolucionales *Places* y *AlexNet* junto con el mejor descriptor holístico basado en métodos analíticos: HOG. Como vemos, se ha dejado fuera de este experimento tanto a la CNN *GoogLeNet* como al descriptor gist, debido a que estos produjeron peores resultados tanto en el error de localización como en el tiempo total de cómputo del proceso.

Una vez aclarados cuáles son los descriptores de apariencia global seleccionados, hemos procedido a resolver la tarea de localización como se ha explicado al comienzo de este capítulo. Con ella, hemos obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal como se muestra en la figura 4.12. De esta forma podemos contrastar, cuando se aplican diferentes efectos en el entorno, los resultados de los descriptores holísticos basados en *deep learning* obtenidos a partir de capas convolucionales con los resultados de los descriptores de apariencia global basados en métodos analíticos.

El tiempo medio total de cómputo en dicha tarea, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición, no se ha representado en este experimento. Esto se debe a que este no iba a aportar nuevos resultados en el estudio, ya que todas estas capas han sido trabajadas previamente, por lo que conocemos cuánto tiempo de cómputo necesitan para resolver la tarea de localización. No obstante, como hemos comentado anteriormente, la capa de la red neuronal convolucional que cuenta con un menor tiempo de cálculo para realizar el proceso es la denominada ‘conv2’, de la CNN *Places*, y no la capa ‘conv4’, perteneciente a la CNN *AlexNet*. Asimismo, aun siendo similar el tiempo de cómputo entre este descriptor y el obtenido por HOG, es este último el que realiza la tarea con mayor rapidez.

Finalmente, en lo que respecta al error medio de localización, entre los descriptores holísticos hallados por estas dos capas convolucionales y HOG, la ‘conv2’ obtiene los mejores resultados en comparación con el resto, como podemos observar en la figura 4.12. No obstante, podemos ver cómo los resultados entre las CNN son similares, aunque empeoran notablemente los obtenidos por *AlexNet* para los efectos de ruido y oclusión. Asimismo, observamos también que los de HOG son parecidos a los de la capa ‘conv2’ siempre y cuando no se esté trabajando con imágenes afectadas por el ruido, ya que de ser así se obtienen resultados sustancialmente peores.

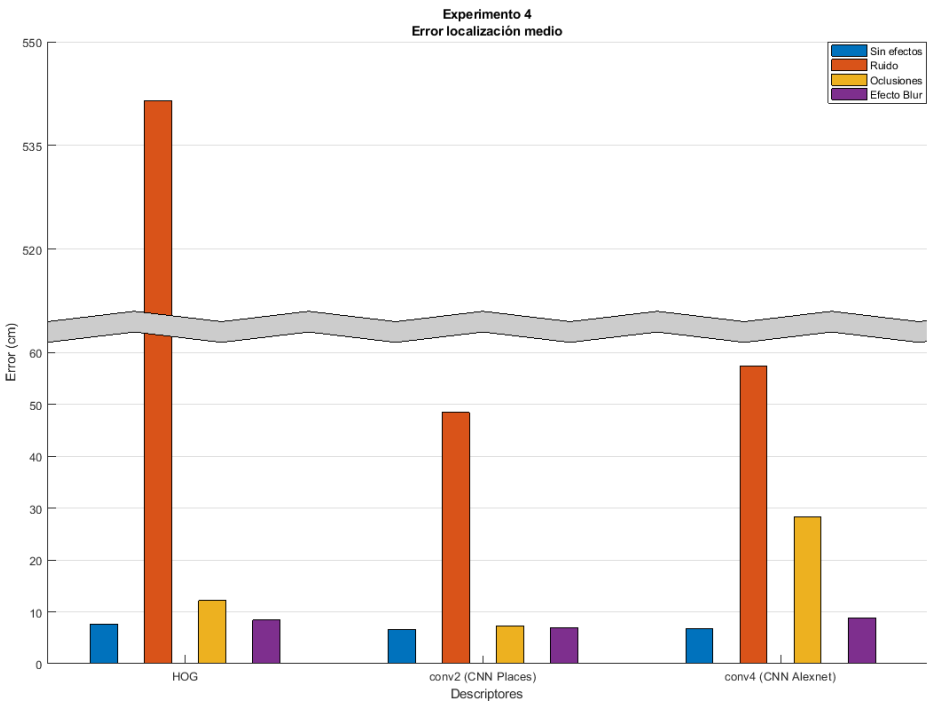


Figura 4.12: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places* y *AlexNet* junto con el descriptor HOG frente a los efectos visuales de ruido (naranja), oclusiones (amarillo) y efecto *blur* (morado).

4.6. Experimento 5: Comparación entre diferentes fuentes de información

Antes de empezar este estudio, cabe recordar que, con la finalidad de contrastar los resultados de los descriptores holísticos basados en métodos analíticos con los obtenidos de los descriptores de apariencia global basados en *deep learning*, hemos partido hasta este momento de la misma imagen panorámica en escala de grises para llevar a cabo la tarea de localización de nuestro robot móvil y así poder realizar de forma más precisa esta comparación entre estos dos métodos. No obstante, para este experimento, incluiremos en los algoritmos de apariencia global la información contenida en los canales de color (RGB), ya que hemos pensado que es posible que estos aporten información más rica al descriptor y permitan mejorar el proceso de localización. Asimismo, trabajaremos con las imágenes originales del *dataset* de nublado, ya que en este

experimento no estamos estudiando cuán precisa es nuestra tarea ante efectos de iluminación o perturbación, sino cómo afecta en dicho proceso un tratamiento diferente de la imagen introducida a la red neuronal convolucional.

La figura 4.13 muestra los diferentes tipos de imágenes con las que se trabajará en este experimento en las redes neuronales convolucionales. En ella podemos observar cuatro: las dos primeras panorámicas, una en escala de grises y otra en RGB, mientras que las restantes son imágenes omnidireccionales, tanto en escala de grises como en RGB.

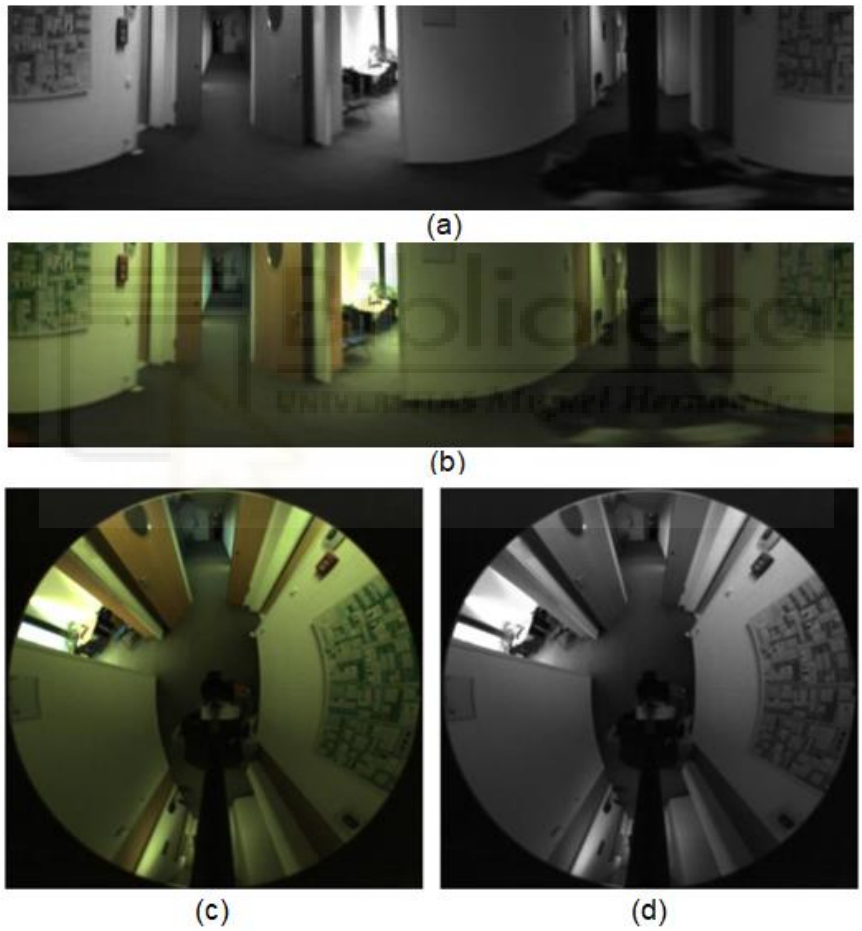


Figura 4.13: La figura (a) muestra la imagen panorámica en escala de grises, la figura (b) la panorámica en color, la figura (c) la omnidireccional en color y la figura (d) la omnidireccional en escala de grises. Todas ellas fueron capturadas dentro del entorno de Saarbrücken. Imágenes obtenidas de COLD *database* [39].

En este experimento, al igual que en los anteriores, realizaremos los mismos estudios en la tarea de localización, pero con la diferencia de que en este trabajaremos únicamente con los tres mejores descriptores holísticos obtenidos del experimento número 3 de cada una de estas redes neuronales convolucionales: *Places*, *AlexNet* y *GoogLeNet*.

Una vez aclarados cuáles son los descriptores de apariencia global seleccionados, hemos procedido a resolver la tarea de localización como se ha explicado al comienzo de este capítulo. Por un lado, de ella hemos obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal como se muestra en la figura 4.14. Por otro lado, en cuanto al tiempo medio total de cómputo compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición, este ha sido obtenido para tres redes neuronales convolucionales diferentes: la CNN *Places*, como podemos ver en la figura 4.15, la CNN *AlexNet*, como podemos observar en la figura 4.16, y la CNN *GoogLeNet*, como se representa en la figura 4.17.

Por un lado, en lo que respecta al error medio de localización, entre los descriptores holísticos hallados por estas tres capas convolucionales, la ‘conv4’ obtiene los mejores resultados con respecto al resto, como podemos observar en la figura 4.14. No obstante, podemos ver cómo los resultados entre las CNN son similares, aunque empeoran notablemente los obtenidos por *Places* y *GoogLeNet* cuando trabajamos en el proceso con imágenes omnidireccionales, ya sean en escala de grises o en RGB. Asimismo, observamos también que los resultados hallados por el descriptor de la capa ‘conv2’, de la CNN *Places*, son los peores de toda nuestra tarea de localización siempre y cuando se trabaje con imágenes omnidireccionales.

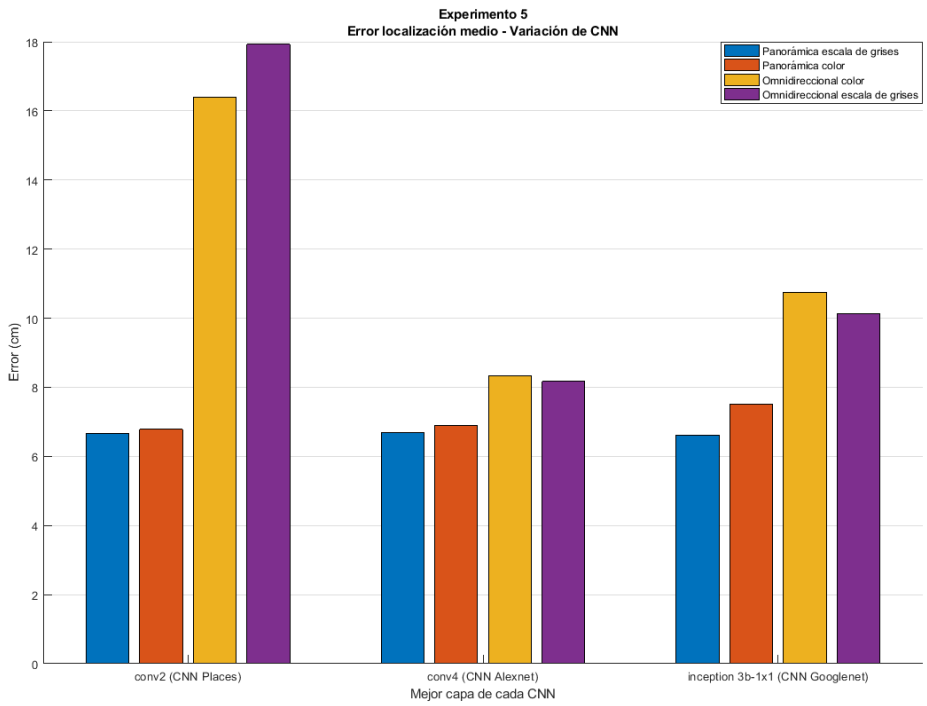


Figura 4.14: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales de la CNN *Places*, *AlexNet* y *GoogLeNet* calculado en función del tipo de fuente utilizada: panorámica en escala de grises (azul) o en RGB (naranja) y omnidireccional en RGB (amarillo) o en escala de grises (morado).

Por otro lado, en lo que se refiere al tiempo medio total de cálculo, entre las distintas imágenes utilizadas por el descriptor holístico obtenido a partir de la capa convolucional 'conv2' de la CNN *Places*, en la figura 4.15 observamos que se necesita un menor tiempo de procesamiento en la tarea de localización si se trabaja con imágenes panorámicas en escala de grises. A su vez, podemos observar cómo el tiempo medio de cálculo del descriptor se incrementa si realizamos el proceso con imágenes omnidireccionales, diferencia que se hace menos notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es similar entre todas las imágenes utilizadas en el experimento.

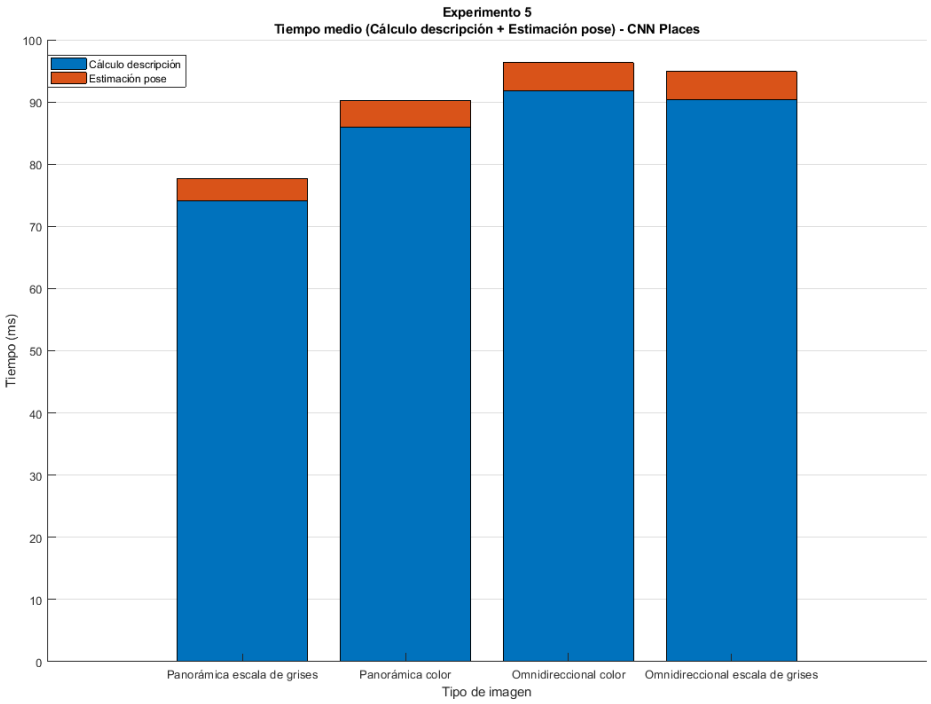


Figura 4.15: Tiempo medio total de cómputo del descriptor obtenido de la capa convolucional ‘conv2’ de la CNN *Places*.

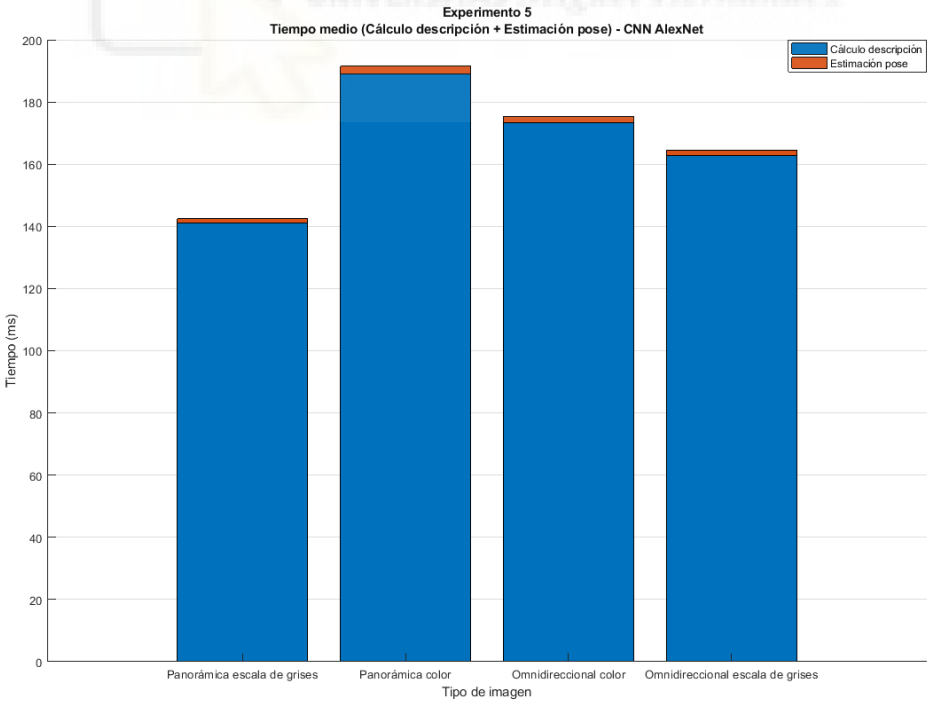


Figura 4.16: Tiempo medio total de cómputo del descriptor obtenido de la capa convolucional ‘conv4’ de la CNN *AlexNet*.

Seguidamente, en lo que se refiere al tiempo medio total de cálculo, entre las distintas imágenes utilizadas por el descriptor holístico obtenido a partir de la capa convolucional ‘conv4’ de la CNN *AlexNet*, en la figura 4.16 observamos que se necesita un menor tiempo de procesamiento en la tarea de localización si se trabaja con imágenes panorámicas en escala de grises. A su vez, podemos observar cómo el mayor tiempo medio de cálculo del descriptor es el correspondiente a realizar el proceso con imágenes panorámicas en color, diferencia que se hace menos notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es similar entre todas las imágenes utilizadas en el experimento.

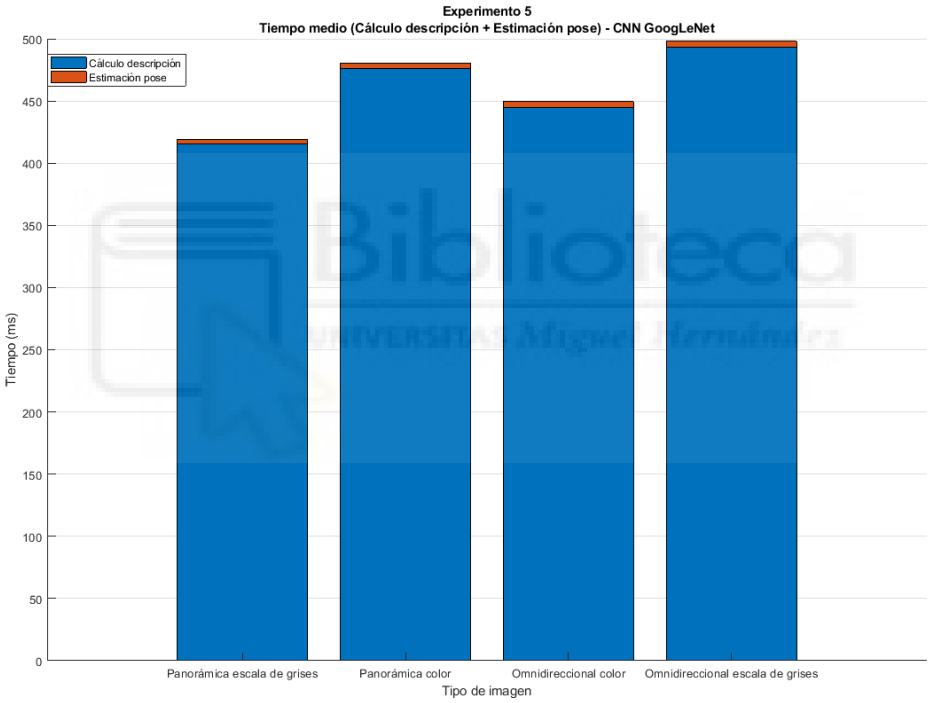


Figura 4.17: Tiempo medio total de cómputo del descriptor obtenido de la capa convolucional ‘inception_3b-1x1’ de la CNN *GoogLeNet*.

A continuación, en lo que se refiere al tiempo medio total de cálculo, entre las distintas imágenes utilizadas por el descriptor holístico obtenido a partir de la capa convolucional ‘inception_3b-1x1’ de la CNN *GoogLeNet*, en la figura 4.17 observamos que se necesita un menor tiempo de procesamiento en la tarea de localización si se trabaja con imágenes panorámicas en escala de grises. A su

vez, podemos observar cómo el mayor tiempo medio de cálculo del descriptor es el correspondiente a realizar el proceso con imágenes omnidireccionales en escala de grises, diferencia que se hace menos notable si nos fijamos en el tiempo de cómputo de la estimación de la pose, el cual es similar entre todas las imágenes utilizadas en el experimento.

Finalmente, recordando lo expuesto en el presente trabajo, hay que tener en cuenta que las redes neuronales convolucionales, tal como se explicó en el capítulo 3, están diseñadas para trabajar con imágenes en color y con un determinado tamaño, ya sea 227 × 227 × 3, en el caso de *Places* y *AlexNet*, o de 224 × 224 × 3, si se trata de *GoogLeNet*, por lo que siempre que se cumplan estas restricciones no habrá ningún problema en que la imagen introducida a la CNN sea panorámica u omnidireccional. Como consecuencia, el interés de este experimento viene dado a la hora de querer optimizar los tiempos de procesamiento en la tarea de localización mediante la imagen seleccionada para el proceso, ya que hasta este momento se trabajaba con una imagen panorámica en escala de grises convertida previamente desde una omnidireccional en color, lo cual necesitaba de un tiempo de conversión elevado.

Asimismo, también queremos hacer hincapié en que los tiempos medios totales de cómputo calculados en el presente trabajo no han tenido en cuenta el tiempo necesario de conversión de la imagen obtenida por el robot móvil desde que esta es tomada en formato omnidireccional hasta que se transforma en cada una de las deseadas en este experimento. Por esta razón, hemos decidido hallar los tiempos necesarios de conversión para cada una de las imágenes utilizadas, tal como muestra la tabla 4.5. Los siguientes resultados son independientes de la red neuronal convolucional a utilizar.

Imagen resultante	Tiempo medio empleado (ms)
Panorámica en color	310.7185
Panorámica en escala de grises	120.3564
Omnidireccional en color	28.4492
Omnidireccional en escala de grises	27.5685

Tabla 4.5: Tiempo de cálculo necesario para obtener la conversión de las diferentes imágenes a introducir en nuestras redes neuronales convolucionales.

Téngase en cuenta que, tal como vemos, el tiempo de procesamiento en cada una de las imágenes es mayor cuando se trata de las de color. Esto se debe a que, por un lado, la conversión a imágenes panorámicas en color se tiene que llevar a cabo para cada una de sus tres capas, acción que disminuye el tiempo de cálculo si trabajamos con imágenes en escala de grises, ya que únicamente se tiene que realizar esta transformación para una capa. Por otro lado, en las imágenes omnidireccionales, se tiene que desechar la información de la imagen adquirida por el robot móvil que no sea de importancia, acción que implica un menor tiempo de cómputo si utilizamos imágenes en escala de grises, debido a que este corte en la información únicamente se realiza para un canal y no para tres, como en el caso de una imagen RGB.

4.7. Experimento 6: Localización visual frente a rotaciones

Llegados a este experimento, se aplicará un nuevo efecto visual a nuestras imágenes de test, que puede darse en situaciones reales de funcionamiento. Este es el denominado efecto de rotación, el cual puede ocurrir cuando el robot móvil visita una determinada zona, pero con una rotación diferente a cuando esta fue inspeccionada en el proceso de *mapping*. En este experimento se ha aplicado una rotación aleatoria entre 10 y 350 grados sobre la imagen omnidireccional, lo que se traduce a un desplazamiento horizontal en la imagen panorámica.

La figura 4.18 muestra el ejemplo de este efecto aplicado sobre una imagen de test. En ella podemos observar, a su vez, dos imágenes: la primera, obtenida directamente del conjunto de datos de nublado, es la original, mientras que la segunda es el resultado de haber aplicado este determinado efecto visual a la anterior. Hay que tener en cuenta que dicho efecto se utiliza únicamente en imágenes que no poseen previamente ningún otro efecto visual, ya que no es óptimo para nuestro proceso aplicar varios de forma simultánea en una misma instantánea, porque lo que se busca con el presente trabajo es estudiar cómo se comporta nuestra tarea ante diferentes efectos visuales planteados de manera individual. Es por ello por lo que estos siempre han sido aplicados al *dataset* de nublado.

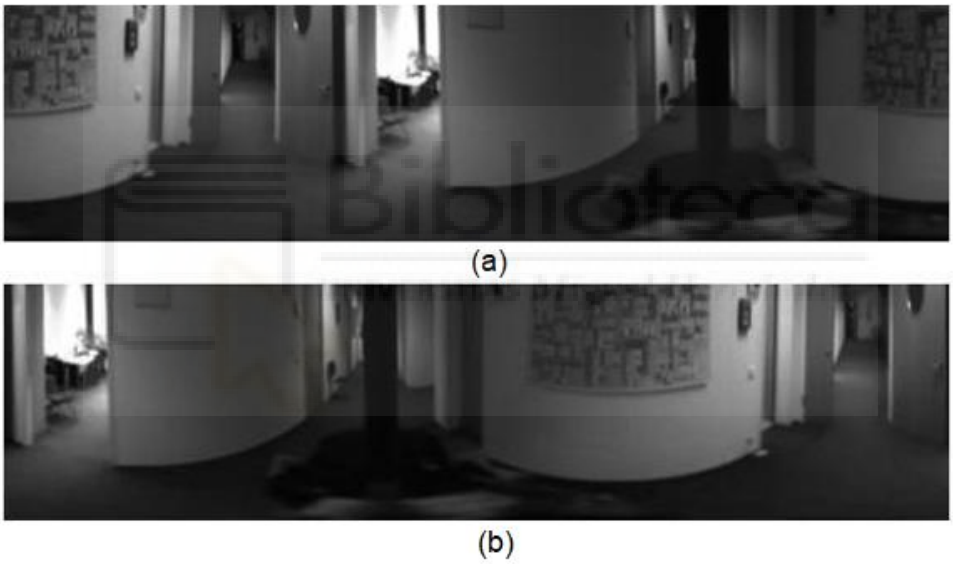


Figura 4.18: La figura (a) muestra la imagen original capturada dentro del entorno de Saarbrücken, mientras que la figura (b) representa el efecto de rotación aplicado sobre la misma.

En este experimento, al igual que en los anteriores, realizaremos los mismos estudios en la tarea de localización, pero con la diferencia de que en este trabajaremos únicamente con los cuatro mejores descriptores obtenidos en experimentos previos de las redes neuronales convolucionales *Places* y *AlexNet*, dos correspondientes a las capas convolucionales y dos a las de clasificación, junto con el mejor descriptor holístico basado en métodos analíticos: HOG. Este último, en nuestro estudio, es invariante ante rotaciones de la imagen; por esta

misma razón en este experimento buscamos ratificar dicha propiedad. Como vemos, se ha dejado fuera de este experimento tanto a la CNN *GoogLeNet* como al descriptor gist, debido a que estos produjeron peores resultados tanto en el error de localización como en el tiempo total de cómputo del proceso.

Una vez aclarados cuáles son los descriptores de apariencia global seleccionados, hemos procedido a resolver la tarea de localización como se ha explicado al comienzo de este capítulo. Con ella, hemos obtenido para cada uno de los descriptores empleados el error medio de posicionamiento, tal como se muestra en la figura 4.19. De esta forma podemos contrastar, cuando se aplican diferentes efectos de rotación en el entorno, los resultados de los descriptores holísticos basados en *deep learning* obtenidos a partir de capas convolucionales y de clasificación con los resultados de los descriptores de apariencia global basados en métodos analíticos.

El tiempo medio total de cómputo en dicha tarea, compuesto por el tiempo de cálculo del descriptor y el tiempo de cálculo de la posición, no se ha representado en este experimento. Esto se debe a que este no iba a aportar nuevos resultados en el estudio, ya que todas estas capas se han trabajado previamente, por lo que conocemos cuánto tiempo de cómputo necesitan para resolver la tarea de localización. No obstante, como hemos comentado anteriormente, la capa de la red neuronal convolucional que cuenta con un menor tiempo de cálculo para realizar el proceso es la denominada ‘conv2’, de la CNN *Places*, y no la capa ‘conv4’, perteneciente a la CNN *AlexNet*. Asimismo, aun siendo similar el tiempo de cómputo entre este descriptor y el obtenido por HOG, es este último el que realiza la tarea con mayor rapidez. Por último, una vez más, los descriptores obtenidos mediante las capas de clasificación ‘fc6’ vuelven a ser los que más tiempo de cómputo necesitan para llevar a cabo la tarea, debido a que estas capas están enfocadas a un trabajo muy especializado y poseen un tamaño de descriptor superior, todo ello teniendo en cuenta que el tiempo medio total de cálculo de la CNN *Places* es ligeramente superior al necesario por la CNN *AlexNet*.

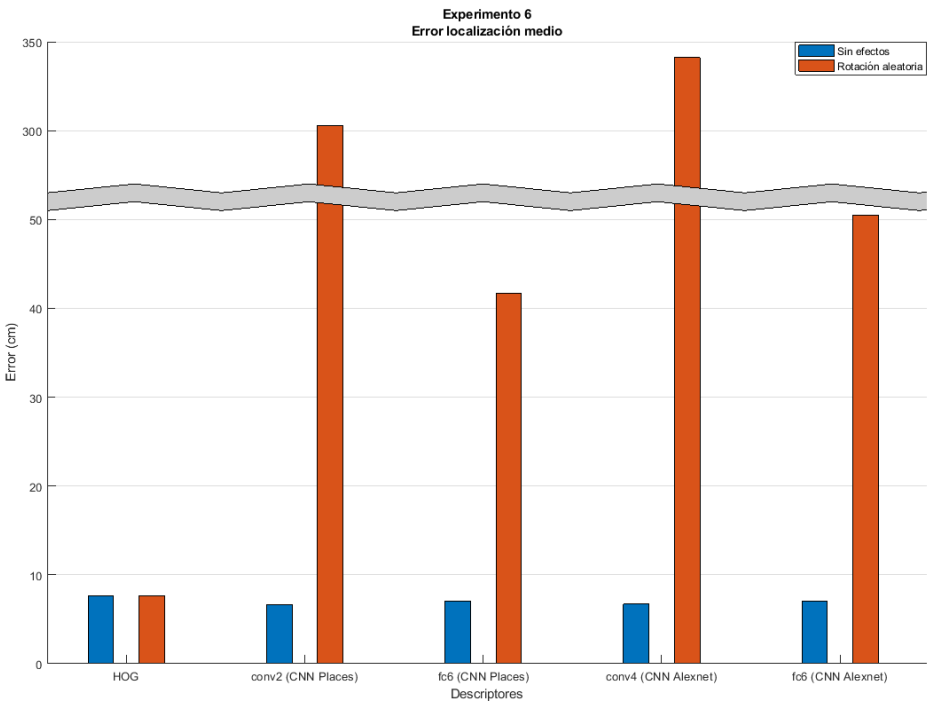


Figura 4.19: Error medio de localización de los descriptores seleccionados obtenidos de las capas convolucionales y de clasificación de la CNN *Places* y *AlexNet* junto con el descriptor HOG.

Finalmente, en lo que respecta al error medio de localización, vemos cómo los resultados obtenidos por el descriptor HOG son los mejores en nuestra tarea de localización. Así pues, podemos corroborar que este es invariante ante efectos de rotación en la imagen. No obstante, entre los descriptores holísticos hallados por estas dos capas de clasificación, la 'fc6' de la CNN *Places* obtiene los mejores resultados con respecto a la 'fc6' de la CNN *AlexNet*, como podemos observar en la figura 4.19. A su vez, vemos que los resultados obtenidos por las capas convolucionales de ambas CNN son desalentadores, ya que empeoran de forma notable cuando se aplican efectos de rotación al modelo visual.

Capítulo 5
Conclusiones y trabajos futuros

Este trabajo está dedicado al estudio de los problemas en la tarea de localización cuando se aplican distintos efectos al modelo visual y solo se dispone de una cámara omnidireccional como sensor de medida. El problema se ha abordado realizando una serie de pruebas en las que se ha resuelto dicha tarea cuando existen cambios de iluminación en el entorno, una perturbación en la imagen debido a efectos como el *blur*, el ruido o las oclusiones; una variación de la estructura de la imagen a la entrada de la red neuronal convolucional y una posible rotación de la fotografía debido a la orientación con la que el robot hubiese tomado la instantánea. El robot móvil que realizaba esta tarea estaba equipado por un sistema de visión catadióptrico montado sobre él y las imágenes que capturaba eran la única información que se proporcionaba para resolver el problema de localización.

El punto de partida es un conjunto de imágenes del Centro Alemán de Investigación en Inteligencia Artificial de Saarbrücken contenidas en la base de datos COLD. Estas imágenes han sido captadas por un sensor de visión en formato omnidireccional. Para la extracción de la información de la imagen se han utilizado descriptores de apariencia global basados en métodos analíticos, como son HOG y gist, y descriptores holísticos basados en herramientas de *deep learning*, como son los obtenidos mediante redes neuronales convolucionales. Tras describir las imágenes, se ha realizado la tarea de localización y se han obtenido los parámetros de tiempo y error de nuestro proceso. Posteriormente, se han observado los resultados y se han comparado los datos obtenidos por parte de todos los descriptores empleados. A continuación, se presentan las conclusiones de cada uno de los seis experimentos realizados.

Experimento nº 1

En este experimento, por una parte, podemos corroborar que el mejor descriptor de apariencia global basado en métodos analíticos es el denominado HOG, debido a que este tiene una mayor precisión en la posición estimada del robot móvil y necesita de un tiempo de cálculo menor para resolver la tarea de localización. Por otra parte, concluimos que la capa de clasificación de la CNN *Places* más adecuada para nuestra tarea es la denominada 'fc6', puesto que su error en la estimación de la posición es mucho menor que en el resto de las capas de la red neuronal convolucional. Asimismo, observamos cómo el tiempo de cómputo del proceso es menos elevado en las capas 'fc8' y 'prob', ya que estas presentan un menor tamaño de descriptor frente al que poseen 'fc6' y 'fc7', que son las dos primeras capas de clasificación de la CNN.

Experimento nº 2

Con el experimento número dos podemos concluir, por un lado, que el descriptor holístico obtenido de las capas convolucionales de la CNN *Places* más adecuado para resolver nuestra tarea de localización es el resultante de utilizar la capa denominada 'conv2'. Por otro lado, entre los descriptores de apariencia global obtenidos mediante las capas convolucionales y de clasificación ('fc6'), se ha decidido que el mejor de ellos corresponde al convolucional. Esto se debe a que, por similares que sean sus resultados en los errores de posición de nuestro robot móvil, el descriptor 'fc6' necesita tres veces más el tiempo total de cálculo para realizar la misma tarea que el 'conv2'.

Experimento nº 3

En este experimento se ha podido estudiar a fondo cada una de las tres redes neuronales convolucionales empleadas en el presente trabajo, por lo que a partir de los resultados obtenidos se ha llegado a la conclusión de que el mejor descriptor para realizar una correcta tarea de localización es el correspondiente a la capa 'conv2' de la CNN *Places*. Esto se debe a dos motivos: el primero es que, a pesar de que se obtienen resultados parecidos en la posición estimada de nuestro móvil para los *dataset* de nublado y noche, en el de soleado presenta una clara ventaja el descriptor 'conv2'; en cuanto al segundo motivo, este

necesita de la mitad del tiempo total de cálculo del proceso de ‘conv4’ y de una quinta parte del de ‘inception_3b-1x1’.

Experimento nº 4

En este experimento hemos podido analizar los resultados en la tarea de localización cuando trabajamos con imágenes afectadas por el ruido, las oclusiones y el *blur*. Así pues, concluimos que el descriptor seleccionado para realizar dicha tarea de una forma más precisa es el obtenido por la capa ‘conv2’ de la CNN *Places*, ya que otorga unos resultados más estables aun trabajando con imágenes afectadas por el ruido, que ha sido el efecto más dañino de todos los aplicados a las imágenes de test.

Experimento nº 5

En este experimento se ha analizado el efecto en nuestro proceso de localización al variar el tipo de imagen introducida en las tres redes neuronales convolucionales seleccionadas. De esta forma, llegamos a la conclusión de que el descriptor con una mayor precisión en dicha tarea corresponde al obtenido por la capa ‘conv4’ de la CNN *AlexNet*, ya que obtenemos unos resultados muy buenos aun trabajando con imágenes omnidireccionales, que han sido las más perjudiciales de todas las introducidas a las redes neuronales convolucionales. No obstante, una vez hallado el tiempo de conversión de las imágenes utilizadas en este estudio, el cual es muy elevado para las panorámicas y pequeño para las omnidireccionales, podemos concluir que una buena propuesta de investigación sería utilizar estas últimas en las CNN *AlexNet* y *GoogLeNet*, quedando excluida *Places* por obtener los peores resultados. Estas imágenes omnidireccionales, a pesar de que poseen un mayor error de posicionamiento con respecto a las panorámicas (un 20% más en el caso de la red neuronal convolucional *AlexNet* y uno de en torno al 45% en el caso de *GoogLeNet*), estarían logrando reducir el tiempo total de cálculo de la tarea de localización en unas 5 veces si trabajamos con imágenes omnidireccionales en escala de grises y en unas 10 veces si utilizamos imágenes omnidireccionales en color.

Experimento nº 6

Para finalizar, en este experimento hemos podido analizar los resultados en la tarea de localización cuando trabajamos con imágenes afectadas por la rotación. Así pues, concluimos que el descriptor seleccionado para realizar dicha tarea de una forma más adecuada es el obtenido por la capa 'fc6' de la CNN *Places*, el cual obtiene resultados más razonables que los hallados por las mejores capas convolucionales de las CNN utilizadas en el presente trabajo. Asimismo, se ha comprobado que nuestro descriptor HOG, tal como definíamos, es invariante ante rotaciones en la imagen.

Los métodos de descripción visual basados en redes neuronales convolucionales buscan minimizar el efecto de *visual aliasing*, pero este es irremediable al haber utilizado en este trabajo únicamente información visual para resolver la tarea de localización. Es por ello por lo que los errores obtenidos en la resolución de esta tarea son producidos principalmente por dicho efecto. Con respecto al tiempo de cálculo del descriptor, ninguna de las capas de las redes neuronales convolucionales requiere de valores altos y, como era de esperar, cuanto más alejada esté la capa correspondiente con respecto a la entrada de la CNN, mayor será el tiempo de cómputo. Asimismo, el tiempo de cálculo para estimar la posición es directamente proporcional al tamaño del descriptor.

Debido al valor de estos resultados experimentales, los cuales animan a profundizar en este campo de estudio por su gran precisión en la estimación de la posición del robot móvil en el modelo visual, se pueden plantear nuevas líneas de trabajo que permitan avanzar en este campo, ya que se ha comprobado que los descriptores de apariencia global son una alternativa robusta en la resolución de la tarea de localización. Este hecho puede suponer interesantes aplicaciones en el mundo de la robótica móvil.

En cuanto a las posibles líneas futuras de investigación, existen varios estudios que se podrían realizar para ampliar este trabajo fin de grado, pero que no se han podido llevar a cabo debido a los límites de extensión establecidos por la normativa. Estos serían los siguientes:

- Realizar los mismos experimentos en entornos de exterior para resolver la tarea de *mapping* y localización. En ellos, los efectos lumínicos son mucho más perjudiciales que los estudiados en entornos de interior.

- Aplicar este método en una base de datos alternativa, lo cual podría ser de interés para contrastar los resultados obtenidos.

- Desarrollar nuevos descriptores de apariencia global basados en técnicas de *deep learning* que permitan mejorar el valor de la precisión en la tarea de localización de un robot móvil.

- Implementar un sistema de navegación totalmente autónomo mediante algoritmos de control visual que permitan navegar al robot haciendo uso del mapa. En esta línea de investigación tomaría gran importancia la tarea de SLAM, ya que sería necesario realizar de manera simultánea las tareas de *mapping* y localización.

Biblioteca
UNIVERSITAS Miguel Hernández

Referencias

[1] R. O. Ambrose, H. Aldridge, R. S. Askew, R. R. Burridge, W. Bluethmann, M. Diftler, et al., “Robonaut: NASA's Space Humanoid”, *IEEE Intelligent Systems*, vol. 15, no. 4, pp. 57-63, August 2000.

[2] N. Ayache and F. Lustman, “Trinocular stereo vision for robotics”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 1, pp. 73-85, January 1991.

[3] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, “Speeded-Up Robust Features (SURF)”, *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, June 2008.

[4] A. Carrio, C. Sampedro, A. Rodríguez and P. Campoy, “A Review of Deep Learning Methods and Applications for Unmanned Aerial Vehicles”, *Journal of Sensors*, vol. 2017, no. 2, pp. 1-13, August 2017.

[5] S. Cebollada, L. Payá, W. Mayol and Ó. Reinoso, “Evaluation of Clustering Methods in Compression of Topological Models and Visual Place Recognition Using Global Appearance Descriptors”, *Applied Sciences*, vol. 9, no. 3, pp. 1-30, January 2019.

[6] S. Cebollada, L. Payá, V. Román and Ó. Reinoso, “Hierarchical Localization in Topological Models Under Varying Illumination Using Holistic Visual Descriptors”, *IEEE Access*, vol. 7, no. 1, pp. 49580-49595, April 2019.

[7] S. Cebollada, L. Payá, D. Valiente, X. Jiang and Ó. Reinoso, “An Evaluation between Global Appearance Descriptors based on Analytic Methods and Deep Learning Techniques for Localization in Autonomous Mobile Robots”, in *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics*, Prague, Czech Republic, 2019, pp. 284-291.

[8] S. Cebollada, V. Román, L. Payá, M. Flores, L. M. Jiménez and Ó. Reinoso, “Uso de técnicas de machine learning para realizar mapping en robótica móvil”, in *Proceedings of the XL Jornadas de Automática*, Ferrol, La Coruña, España, 2019, pp. 686-693.

[9] G. Dudek and M. Jenkin, *Computational Principles of Mobile Robotics*. New York: Cambridge University Press, 2010.

[10] M. Flores, “Desarrollo de algoritmos de odometría visual mediante una cámara de 360 grados”, Trabajo de Fin de Máster, Ingeniería de Sistemas y Automática, Universidad Miguel Hernández de Elche, Elche, Alicante, España, 2018.

[11] J. Gaspar, N. Winters and J. Santos-Victor, “Vision-based navigation and environmental representations with an omnidirectional camera”, *IEEE Transactions on Robotics and Automation*, vol. 16, no. 6, pp. 890-898, December 2000.

[12] A. Gil, Ó. Reinoso, M. Ballesta, M. Juliá and L. Payá, “Estimation of Visual Maps with a Robot Network Equipped with Vision Sensors”, *Sensors*, vol. 10, no. 5, pp. 5209-5232, May 2010.

[13] G. Giralt, R. Sobek and R. Chatila, “A multi-level planning and navigation system for a mobile robot: a first approach to HILARE”, in *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, Tokyo, Japan, 1979, pp. 335-337.

[14] A. Iborra, J. A. Pastor, B. Álvarez, C. Fernández and J. M. F. Merono, “Robots in radioactive environments”, *IEEE Robotics and Automation Magazine*, vol. 10, no. 4, pp. 12-22, December 2003.

[15] M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. New York: Wiley-Interscience, 2000.

[16] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097-1105, January 2012.

[17] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, November 2004.

[18] T. T. Mac, C. Copot, D. T. Tran and R. De Keyser, “Heuristic approaches in robot path planning: A survey”, *Robotics and Autonomous Systems*, vol. 86, pp. 13-28, August 2016.

[19] G. Marani, S. K. Choi and J. Yuh, “Underwater autonomius manipulation for intervention missions AUVs”, *Ocean Engineering*, vol. 36, no. 1, pp. 15-23, January 2009.

[20] H. P. Moravec, “Towards automatic visual obstacle avoidance”, in *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, Cambridge, MA, USA, 1977, p. 584.

[21] A. C. Murillo, J. J. Guerrero and C. Sagues, “SURF features for efficient robot localization with omnidirectional images”, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Roma, Italy, 2007, pp. 1-8.

[22] J. Neumann, C. Fermuller and Y. Aloimonos, “Eyes from eyes: new cameras for structure from motion”, in *Proceedings of the IEEE Workshop on Omnidirectional Vision. Held in conjunction with ECCV'02*, Copenhagen, Denmark, 2002, pp. 19-26.

[23] N. J. Nilsson, “Shakey the Robot”, SRI International, Menlo Park, California, USA, Technical Note 323, 1984.

[24] A. Oliva and A. Torralba, “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope”, *International Journal of Computer Vision*, vol. 42, no. 1, pp. 145-175, May 2001.

[25] L. Payá, “Técnicas de descripción de la apariencia global de escenas: Aplicación a la creación de mapas y localización de robots móviles”, Tesis Doctoral, Ingeniería de Sistemas y Automática, Universidad Miguel Hernández de Elche, Elche, Alicante, España, 2014.

[26] L. Payá, L. Fernández, Ó. Reinoso, A. Gil and D. Úbeda, “Appearance-based Dense Maps Creation - Comparison of Compression Techniques with Panoramic Images”, in *Proceedings of the 6th International Conference on Informatics in Control, Automation and Robotics*, Milan, Italy, 2009, pp. 250-255.

[27] L. Payá, A. Gil and Ó. Reinoso, “A State-of-the-Art Review on Mapping and Localization of Mobile Robots Using Omnidirectional Vision Sensors”, *Journal of Sensors*, vol. 2017, pp. 1-20, April 2017.

[28] L. Payá, A. Peidró, F. Amorós, D. Valiente and Ó. Reinoso, “Modeling Environments Hierarchically with Omnidirectional Imaging and Global-Appearance Descriptors”, *Remote Sensing*, vol. 10, no. 4, pp. 522, March 2018.

[29] V. Román, “Localización de un robot móvil mediante visión por computador ante grandes cambios de iluminación”, Trabajo de Fin de Máster, Ingeniería de Sistemas y Automática, Universidad Miguel Hernández de Elche, Elche, Alicante, España, 2017.

[30] F. Rossi, A. Ranganathan, F. Dellaert and E. Menegatti, “Toward topological localization with spherical Fourier transform and uncalibrated camera”, in *Proceedings of the International Conference on Simulation, Modeling and Programming for Autonomous Robots*, Venice, Italy, 2008, pp. 319-330.

[31] S. Se, D. Lowe and J. Little, “Vision-based mobile robot localization and mapping using scale-invariant features”, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Seoul, South Korea, 2001, pp. 2051-2058.

[32] G. Shaffer and A. Stentz, “A Robotic System for Underground Coal Mining”, in *Proceedings of the IEEE International Conference on Robotics and Automation*, Nice, France, 1992, pp. 633-638.

[33] C. Siagian and L. Itti, “Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300-312, February 2007.

[34] R. Sim and G. Dudek, “Effective exploration strategies for the construction of visual maps”, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, 2003, pp. 3224-3231.

[35] R. Sim and J. J. Little, “Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters”, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006, pp. 2082-2089.

[36] S. W. Squyres, A. H. Knoll, R. E. Arvidson, J. W. Ashley, J. F. Bell, W. M. Calvin, et al., “Exploration of Victoria Crater by the Mars Rover Opportunity”, *Science*, vol. 324, no. 5930, pp. 1058-1061, May 2009.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., “Going deeper with convolutions”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1-9.

[38] A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin, “Context-based vision system for place and object recognition”, in *Proceedings of the 9th IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1023-1029.

[39] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo and P. Jensfelt, “The COLD Database”, KTH Royal Institute of Technology, Stockholm, Sweden, Technical Report TRITA-CSC-CV 2007:1, 2007.

[40] D. Voth, “A New Generation of Military Robots”, *IEEE Intelligent Systems*, vol. 19, no. 4, pp. 2-3, July/August 2004.

[41] N. Winters, J. Gaspar, G. Lacey and J. Santos-Victor, “Omni-directional vision for robot navigation”, in *Proceedings of the IEEE Workshop on Omnidirectional Vision*, Hilton Head Island, SC, USA, 2000, pp. 21-28.

[42] D. F. Wolf and G. S. Sukhatme, “Semantic Mapping Using Mobile Robots”, *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 245-258, April 2008.

[43] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, et al., “Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images”, *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 119-130, January 2016.

[44] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, “Learning Deep Features for Scene Recognition using Places Database”, in *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 487-495.

Biblioteca
UNIVERSITAS Miguel Hernández