Capítulo 1

Introducción

En este capítulo se introducirá de forma general el marco en el que se encuadra el presente trabajo de fin de máster (TFM). Concretamente, se explicará qué es la Visión Artificial y, en particular, el campo de la auto-localización visual en entornos desconocidos mediante el uso del algoritmo de SLAM. Por otra parte, se expondrá qué es la robótica, así como su fuerte relación con la Visión Artificial. Por último, se expondrá la estructura que da forma al documento.

1.1. Visión Artificial

La Visión Artificial es la rama de la Inteligencia Artificial orientada a la captura y procesamiento de imágenes. En primer lugar, la captura es posible mediante el desarrollo de sensores capaces de recibir, procesar y almacenar parte del espectro electromagnético, como la luz visible o el infrarrojo, obteniendo imágenes con distintos tipos de información. En segundo lugar, el procesamiento engloba toda la parte del desarrollo de algoritmos capaces de interpretar el contenido de dichas imágenes.

Las imágenes contienen una gran cantidad de información útil para numerosos problemas, sin embargo, extraer y procesar dicha información no es un proceso sencillo. De hecho, la propia representación de las imágenes, que suele darse en forma de matriz numérica, ya es confusa. Como se puede apreciar en la Figura 1.1, nuestros sentidos no interpretan de igual modo una imagen (izquierda) que su representación (derecha).

Los orígenes de este campo se remontan a la década de los 60, cuando se conectó por primera vez una cámara a un computador con el fin de obtener y analizar la información. Larry Roberts fue una de las primeras personas en desarrollar un experimento en el cual, no solo se capturaban imágenes, sino que se analizaba su contenido (una estructura de bloques) para posteriormente reproducirlo desde otra perspectiva.

En sus inicios, las técnicas de visión artificial estuvieron limitadas por la capacidad de cómputo de los ordenadores de la época. Sin embargo, en los últimos años, con el avance de

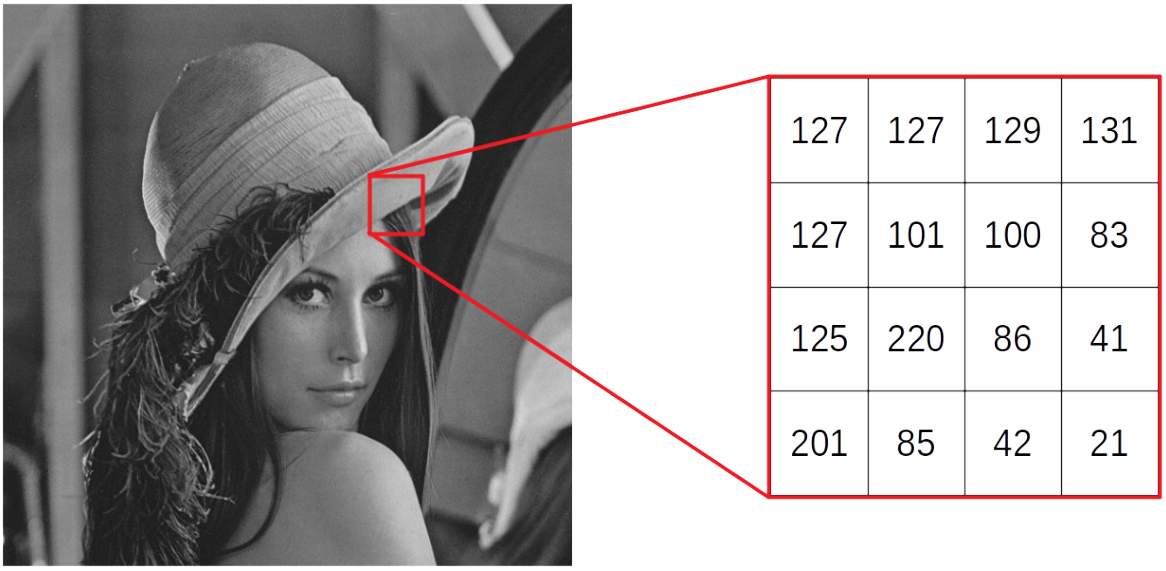

Figura 1.1: Comparación entre la información visual y su representación numérica.

la tecnología y la reducción del coste económico del *hardware*, se ha producido un gran avance en este campo. Esto ha permitido la aparición de algoritmos capaces de trabajar en tiempo real y estando presente en campos como la vigilancia, la robótica, la medicina e incluso los videojuegos. Algunas de las aplicaciones clásicas de este campo son la siguientes:

- **Detección de objetos.** Es posible identificar si un objeto se encuentra en una imagen realizando búsquedas por rasgos característicos del mismo, como la forma, el color, textura o patrones presentes en él.
- **Seguimiento de objetos.** Como extensión de la tarea anterior, es interesante poder realizar un seguimiento a un mismo objeto, de forma inequívoca, a lo largo del tiempo.
- **Reconocimiento de caracteres.** Esta técnica, conocida como OCR por sus siglas en inglés (*Optical Character Recognition*), permite la detección e identificación de los caracteres en un documento, ya sean digitales o manuscritos, con la finalidad de digitalizar el contenido de los mismos.

El rendimiento y precisión de estas tareas ha aumentado en los últimos años debido al uso de las redes neuronales, más concretamente al uso de técnicas de *Deep Learning*. La mejora ha sido tal, que algunos sistemas de reconocimiento de patrones visuales obtienen tasas de error inferiores a las humanas. El primer sistema «sobrehumano» fue obtenido en 2011 en la competición de reconocimiento de señales de tráfico IJCNN, donde el sistema vencedor obtuvo una tasa de error dos veces mejor que la obtenida por sujetos de prueba humanos [2].

Sin embargo, hay campos donde las técnicas de *Deep Learning* aún no han descatado especialmente respecto a técnicas más clásicas de Visión Artificial, como por ejemplo la auto-

localización. Este problema consiste en la estimación de la ubicación de la cámara en el entorno (conocido o desconocido) que la rodea. Esta técnica es esencial para campos como la robótica, donde es imprescindible reconocer y ubicarse en el entorno para evitar choques y/o comportamientos anómalos; o al reciente campo de la realidad aumentada.

En la próxima sección se profundiza en un algoritmo de auto-localización llamado Visual SLAM.

1.2. Visual SLAM

Localización y mapeado simultáneo o SLAM por sus siglas en inglés (*Simultaneous Localization And Mapping*) es un algoritmo de auto-localización. SLAM es el término empleado para describir el proceso por el cual, a partir de información obtenida de sensores, es posible generar un mapa del entorno que lo rodea, al mismo tiempo que se localiza en él. Tiene sus orígenes en el campo de la robótica, siendo esta una de sus principales áreas de investigación, dónde cobra especial importancia en robots autónomos que operan en entornos desconocidos.

Cuando el sensor principal de SLAM son una o más cámaras nos referimos al modelo como Visual SLAM. Pese a ser un algoritmo que aún está en desarrollo, dado que no hay una solución exacta para el problema que plantea, es empleado en un numerosas aplicaciones. Algunas de ellas son las siguientes:

- **Robot aspirador:** estos dispositivos autónomos son equipados con cámaras (u otros sensores similares como los LIDAR) siendo capaces de generar un mapa de los hogares navegando por ellos. La ventaja en este caso de conocer el entorno y su ubicación es la capacidad de generar y optimizar las rutas de limpieza, al mismo tiempo que evitan obstáculos en su trayectoria.

- **UAV:** los vehículos aéreos no tripulados (UAV por sus siglas en inglés) como los drones pueden emplear SLAM para visualizar el entorno que los rodea y tomar decisiones en tiempo real.

- **Realidad aumentada:** SLAM no solo tiene cabida en el mundo de la robótica, en el campo de la realidad aumentada también es posible hacer uso de este algoritmo para relacionar de un mejor modo el mundo real con el virtual. Estas aplicaciones emplean el mapa obtenido para introducir los elementos visuales de forma más realista. Esto solo es posible si se conoce la posición del sensor y el entorno. Un ejemplo de realidad aumentada puede observarse en la Figura 1.2, donde se incluye mobiliario en en una habitación de forma más genuina.

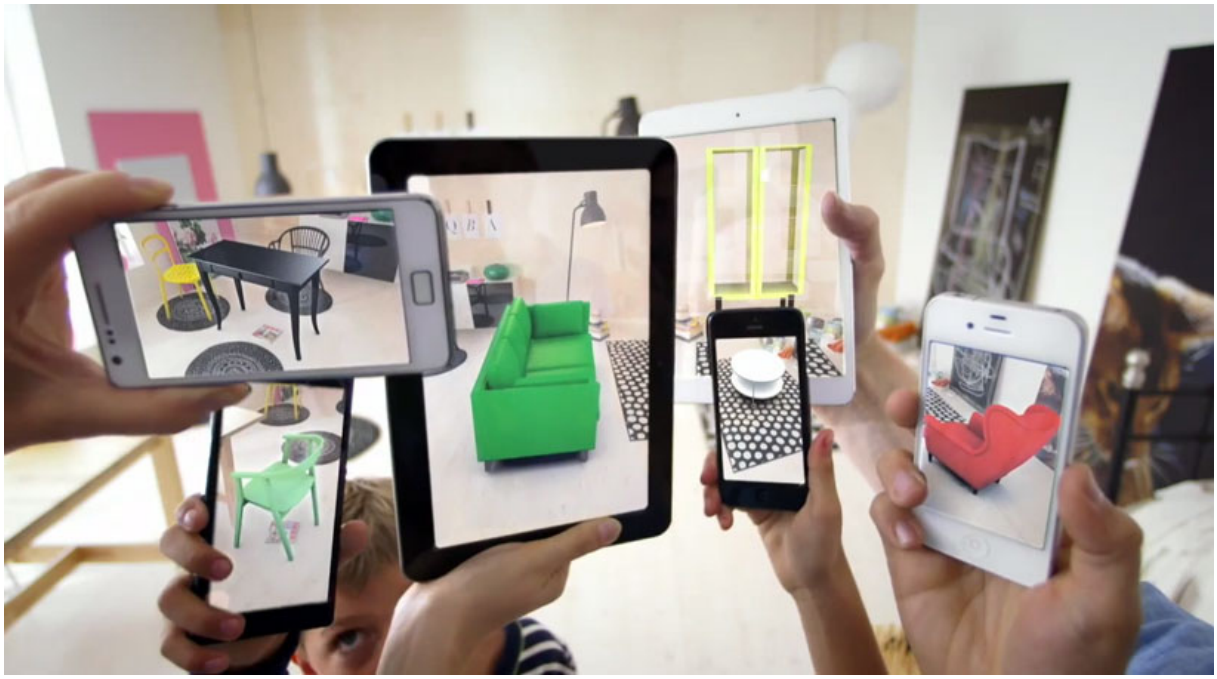

Figura 1.2: Ejemplo de aplicación de Realidad Aumentada.

Visual SLAM tiene sus raíces en la línea de investigación conocida como *Structure from motion*. Dicha línea está centrada en la reconstrucción automática de estructuras 3D a partir de un conjunto de imágenes. Este proceso se basa en la premisa de que, dadas múltiples vistas de un mismo punto tridimensional a lo largo de distintas imágenes, es posible estimar su posición en el espacio mediante el uso de tiangulación. En la Figura 1.3 se puede observar un ejemplo de reconstrucción 3D a partir de 3 imágenes.

Structure from motion supuso un gran avance en las técnicas de detección de puntos característicos, también llamados puntos de interés, que serían de gran utilidad posteriormente en Visual SLAM. Estos puntos son llamados característicos porque son lo suficientemente únicos (por sus propiedades o entorno que los rodean) que los vuelven lo muy robustos de cara a ser detectados de nuevo ante cambios de iluminación o distintos ángulos de vista. Los extractores de puntos característicos más conocidos son *SURF*, *SIFT* y *ORB*.

Esta técnica serviría como idea básica del funcionamiento de Visual SLAM. De forma general, la estructura a reconstruir es el entorno por donde navega la cámara. Por otra parte, el conjunto de las imágenes es generado a lo largo del tiempo, y de cada una de ellas se extraen puntos de interés para estimar su posición tridimensional, dando lugar a un mapa de puntos. Por último, se hace uso del mapa para estimar la posición de la cámara.

Esta no es la única aproximación al problema, ya que debido al avance de la tecnología, es posible emplear no sólo sistemas formados por una única cámara (modelos que denominaremos *Monoculares*), sino utilizar sistemas más complejos. Por ejemplo se puede emplear un

Figura 1.3: Reconstrucción 3D mediante *Structure from Motion*.

par estéreo, formado por dos cámaras separadas entre sí a una distancia conocida, aprovechando esta configuración para estimar la información de profundidad de toda la escena; o directamente emplear una cámara *RGBD* que ya aporta la imagen de profundidad, sin necesidad de calcularla. Todos estos modelos y sistemas tienen sus ventajas y desventajas como veremos en la sección **??**.

1.2.1. Conceptos

Los algoritmos de localización como SLAM tienen asociados una serie de conceptos [1] que merece la pena detallar, dado que se hará uso de ellos a lo largo del proyecto.

Calidad: la calidad del algoritmo dependerá de la eficiencia temporal, la precisión espacial de la pose y la robustez.

Eficiencia temporal: medida como el tiempo de ejecución de cada iteración. Para considerar que el algoritmo es apto para trabajar en tiempo real deberá ser capaz de procesar al menos 30 fotogramas. (*frames*) por segundo.

Precisión de la posición: diferencia entre la pose estimada y la pose real, expresada como el error lineal y angular.

Robustez: entendida como la capacidad de recuperarse o seguir funcionando ante situaciones inesperadas, como oclusiones, imágenes borrosas, objetos dinámicos en la escena, etc.

Oclusiones: situación donde la cámara del sistema esté tapada total o parcialmente, de modo que no sea posible utilizar la totalidad de la imagen para obtener información.

Relocalización: capacidad para recuperarse de una pérdida por falta de información, siendo capaz de volver a estimar la posición de forma correcta dentro del mapa.

1.3. Robótica

Uno de los principales campos de investigación donde la visión artificial tiene una fuerte presencia, como se ha comentado anteriormente, es la robótica. Los robots son sistemas electromecánicos diseñados y programados con el fin de realizar unas determinadas funciones.

Para lograr este fin hacen uso de tres tipos de componentes *hardware*: actuadores, sensores y unidades de procesamiento. Los actuadores y sensores permiten interactuar y obtener información del mundo real, mientras la unidad de procesamiento actúa como el cerebro que analiza los datos proporcionados por los sensores y, en consecuencia, toma las decisiones que deben realizar en cada momento los actuadores.

Existen gran variedad de robots desarrollados para entornos industriales, ya que tienen la capacidad de realizar el trabajo de una forma muy precisa, siendo un claro ejemplo los brazos robóticos (Figura 1.4a) utilizados en numerosas cadenas de montaje. Sin embargo, también son empleados en otros campos como la educación, defensa, medicina, la ayuda en el hogar e incluso la exploración espacial.

Uno de los sensores más utilizados en robótica son las cámaras, debido a la gran información del entorno que pueden proporcionar. Principalmente, son empleados en robots móviles autónomos, es decir, aquellos especializados en la navegación sobre un terreno conocido, o desconocido, sin control humano.

Algunos ejemplos de estos tipos de robots pueden ser los robots aspiradora y UAVs mencionados anteriormente, o más recientemente el desarrollo de los coches autónomos. En la conducción autónoma, los vehículos son dotados de numerosos sensores entre los cuales se incluyen cámaras. A partir de las imágenes obtenidas es posible detectar los carriles de circulación, otros vehículos (Figura 1.4b) e incluso peatones.

(a) (b)

Figura 1.4: Brazo robótico industrial (a). *Software* de conducción autónoma (b).

1.4. Estructura del documento

Una vez introducidas las disciplinas en las cuales se enmarca el desarrollo del proyecto, en los siguientes capítulos se describe el trabajo realizado.

En primer lugar, se realiza un repaso al estado del arte actual de las técnicas de auto-localización en entornos desconocidos, para posteriormente plantear el objetivo del proyecto. Una vez planteados los objetivos, los capítulos 4 y 5 describirán el desarrollo y los resultados obtenidos, incluyendo la metodología empleada para la obtención de los datos. Por último, se expondrás las conclusiones finales del trabajo tras haber analizado los resultados.

Bibliografía

[1] E. P. García. Técnicas para la localización visual robusta de robots en tiempo real con y sin mapas.

[2] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.