

**MÁSTER UNIVERSITARIO
EN VISION ARTIFICIAL**

Curso Académico 2016/2017

Asignatura de Introducción a la Investigación en Visión Artificial

Trabajo sobre el Estado del Arte de
VisualSLAM

Autor: Elías Barcia Mejias

Tutores: José María Cañas Plaza , Eduardo Perdices García

Resumen

Actualmente la investigación y desarrollo en robótica móvil está en pleno auge. Los robots modernos están equipados con múltiples sensores y uno de los más utilizados son las cámaras, ya que permiten al robot captar en imágenes todo el entorno que le rodea. En contra partida, el procesado de imágenes conlleva una carga notable de CPU debido a la enorme cantidad de información que puede aportar cada imagen.

Una de las funcionalidades más importantes que se persigue, es que los robots móviles puedan desplazarse por su entorno y navegar desde la posición A a la posición B de forma autónoma. Esta tarea no resulta muy complicada en entornos estructurados, donde el robot conoce de antemano el terreno por el que se mueve o sabe de la existencia de alguna baliza que le dé pistas de su posición. Pero en entornos no estructurados, donde el robot desconoce por completo el terreno, carece de mapas y no existe a priori ningún tipo de marca o baliza que pueda guiar al robot, la navegación resulta mucho más compleja.

En exteriores, podríamos guiar al robot mediante GPS, pero la señal GPS no llega con la suficiente potencia a todas partes. Por ejemplo en interiores de edificios o en zonas subterráneas, o mejor aún, imaginemos que enviásemos el robot a explorar la superficie del planeta Marte, donde la señal GPS es inexistente. ¿Cómo se las arreglaría el robot para desplazarse por el terreno de forma autónoma sin perderse?

Hoy en día ya existe una familia de técnicas que permite al robot navegar de manera autónoma por zonas desconocidas para él, esta técnica se llama VisualSLAM.

Visual SLAM (*Simultaneous Localization and Mapping*) es una técnica utilizada principalmente con robots móviles y que aporta al robot la capacidad de autolocalizarse y generar mapas del entorno que le rodea en tiempo real. Gracias a ese mapa y principalmente a esa autolocalización se pueden utilizar las técnicas de navegación autónoma, que requieren inevitablemente de una estimación de posición propia fiable. VisualSLAM básicamente se comporta como una caja negra que procesa las imágenes en secuencia captadas por una o varias cámaras. A partir de esas imágenes el robot es capaz de obtener su posición 3D en

el mundo que le rodea. De esta forma el robot podrá desplazarse en su entorno de forma autónoma sin perderse.

El robot, debe contar con una capacidad de cálculo suficiente que le permita ejecutar el software de procesado de imágenes y al mismo tiempo realizar la generación de mapas. Estas tareas requieren ser ejecutadas en tiempo real, unos 30 fotogramas por segundo. Es posible utilizar la técnica de VisualSLAM hoy en día en pequeños dispositivos gracias al aumento de su potencia de computación.

Dependiendo del tipo de cámaras con las que esté equipado el robot, tendrá mayor o menor capacidad de ejecutar VisualSLAM. Como mínimo el robot debe tener una cámara RGB, muy común en los drones, aunque también puede tener 2 cámaras estéreo que le ayudarán a representar el entorno en 3D con mayor fiabilidad. Otras cámaras, RGBD como las utilizadas en el proyecto Tango se ayudan de un sensor de profundidad que también capacita al robot para representar en tres dimensiones el mundo que les rodea con mayor robustez y precisión.

El presente documento está dividido en 5 secciones y trata de describir el estado del arte de Visual SLAM. Esta primera sección o módulo es una introducción a Visual SLAM. El siguiente modulo será una descripción de las aplicaciones actuales de Visual SLAM en distintos dispositivos, desde robots aspiradora, pasando por drones y Smartphones de última generación. En el módulo 3 hablaremos de la problemática de Visual SLAM, cuales son las principales dificultades que debemos solventar a la hora de implementar un algoritmo de VisualSLAM. El punto 4 es el más extenso , contiene una breve descripción de los módulos principales que componen el algoritmo de Visual SLAM y un resumen de las técnicas actuales más conocidas de Visual SLAM. Y por último, el punto 5, donde se muestran las conclusiones.

Índice general

1.	Introducción	1
1.1.	Visión Artificial	2
1.2.	Visual SLAM	4
1.3.	Aplicaciones en VisualSLAM	5
1.4.	Visual SLAM en Robótica Móvil	10
1.5.	Conceptos	13
2.	Objetivos	1
2.1.	Requisitos	2
3.	Problemas de Visual SLAM	3
3.1.	Inicialización del Mapa:	3
3.2.	Ambigüedad en la escala:	3
3.3.	Distorsión Rolling Shutter:	4
3.4.	Dificultad para operar en entornos con pocas texturas:	4
4.	Estado del Arte	5
4.1.	MonoSLAM	5
4.2.	PTAM	7
4.3.	DTAM	8
4.4.	SVO	9
4.5.	LSD-SLAM	10
4.6.	ORB-SLAM	11
4.7.	DSO	12
4.8.	SDVL	14

4.9.	RGB-D Visual SLAM	15
4.10.	Herramientas para realizar evaluaciones comparativas de algoritmos SLAM	16
4.11.	Comparativa de los algoritmos más representativos	20
5.	Conclusiones	21
	Bibliografía	23

1. Introducción

Actualmente la investigación y desarrollo en robótica móvil está en pleno auge. Los robots modernos están equipados con múltiples sensores y uno de los más utilizados son las cámaras, ya que permiten al robot captar en imágenes todo el entorno que le rodea. En contra partida, el procesado de imágenes conlleva una carga notable de CPU debido a la enorme cantidad de información que puede aportar cada imagen.

Una de las funcionalidades más importantes que se persigue, es que los robots móviles puedan desplazarse por su entorno y navegar desde la posición A a la posición B de forma autónoma. Esta tarea no resulta muy complicada en entornos estructurados, donde el robot conoce de antemano el terreno por el que se mueve o sabe de la existencia de alguna baliza que le dé pistas de su posición. Pero en entornos no estructurados, donde el robot desconoce por completo el terreno, carece de mapas y no existe a priori ningún tipo de marca o baliza que pueda guiar al robot, la navegación resulta mucho más compleja.

En exteriores, podríamos guiar al robot mediante GPS, pero la señal GPS no llega con la suficiente potencia a todas partes. Por ejemplo en interiores de edificios o en zonas subterráneas, o mejor aún, imaginemos que enviásemos el robot a explorar la superficie del planeta Marte, donde la señal GPS es inexistente. ¿Cómo se las arreglaría el robot para desplazarse por el terreno de forma autónoma sin perderse?

Hoy en día ya existe una familia de técnicas que permite al robot navegar de manera autónoma por zonas desconocidas para él, esta técnica se llama VisualSLAM.

Visual SLAM (*Simultaneous Localization and Mapping*) es una técnica utilizada principalmente con robots móviles y que aporta al robot la capacidad de autolocalizarse y generar mapas del entorno que le rodea en tiempo real. Gracias a ese mapa y principalmente a esa autolocalización se pueden utilizar las técnicas de navegación autónoma, que requieren inevitablemente de una estimación de posición propia fiable. VisualSLAM básicamente se comporta como una caja negra que procesa las imágenes en secuencia captadas por una o varias cámaras. A partir de esas imágenes el robot es capaz de obtener su posición 3D en el mundo que le rodea. De esta forma el robot podrá desplazarse en su entorno de forma autónoma sin perderse.

El robot, debe contar con una capacidad de cálculo suficiente que le permita ejecutar el software de procesado de imágenes y al mismo tiempo realizar la generación de mapas. Estas tareas requieren ser ejecutadas en tiempo real, unos 30 fotogramas por segundo. Es posible utilizar la técnica de VisualSLAM hoy en día en pequeños dispositivos gracias al aumento de su potencia de computación.

Dependiendo del tipo de cámaras con las que esté equipado el robot, tendrá mayor o menor capacidad de ejecutar VisualSLAM. Como mínimo el robot debe tener una cámara RGB, muy común en los drones, aunque también puede tener 2 cámaras estéreo que le ayudarán a representar el entorno en 3D con mayor fiabilidad. Otras cámaras, RGBD como las utilizadas en el proyecto Tango se ayudan de un sensor de profundidad que también capacita al robot para representar en tres dimensiones el mundo que les rodea con mayor robustez y precisión.

El presente documento está dividido en 5 secciones y trata de describir el estado del arte de Visual SLAM. Esta primera sección o módulo es una introducción a Visual SLAM. El siguiente modulo será una descripción de las aplicaciones actuales de Visual SLAM en distintos dispositivos, desde robots aspiradora, pasando por drones y Smartphones de última generación. En el módulo 3 hablaremos de la problemática de Visual SLAM, cuales son las principales dificultades que debemos solventar a la hora de implementar un algoritmo de VisualSLAM. El punto 4 es el más extenso , contiene una breve descripción de los módulos principales que componen el algoritmo de Visual SLAM y un resumen de las técnicas actuales más conocidas de Visual SLAM. Y por último, el punto 5, donde se muestran las conclusiones.

1.1. Visión Artificial

La Visión Artificial , es una rama científico técnica creada para extraer y procesar información a partir de imágenes, para ello genera sistemas que intentan emular el sentido de visión de los seres humanos .

Sus inicios, se remontan hacia mediados del siglo pasado, cuando en 1961 Larry Roberts desarrolló en la universidad un programa que era capaz de ver una estructura de bloques, analizar su contenido y reproducirla desde otra perspectiva, para ello tuvo que utilizar los dos componentes principales de un sistema de visión artificial , una cámara y necesariamente un ordenador. Pero debido a la alta complejidad de las tareas de visión artificial y a la primitiva capacidad de proceso de las computadoras de la época, los resultados en la investigación sobre visión artificial fueron muy limitados y podemos decir que la evolución de la visión artificial ha ido ligada a los avances en la computación. Con la aparición de microprocesadores más rápidos, el aumento exponencial de la memoria y la creación y mejora de los algoritmos se han ido consiguiendo mejores resultados hasta poder crear sistemas de visión artificial que son capaces de operar en tiempo real, permitiendo que un automóvil sea capaz de conducir de forma autónoma, o que un robot sea capaz de

agarrar una pelota cuando se la lanzamos.

Actualmente, la visión artificial se utiliza en muchos procesos científicos, industriales y militares, por ejemplo para el reconocimiento de objetos o en el seguimiento de éstos:

Reconocimiento de objetos: Se trata de buscar unas propiedades concretas de un determinado objeto (forma, color o cualquier otro patrón) en una imagen para determinar si un objeto se encuentra o no en ella. Por ejemplo mediante la obtención de píxeles característicos que destaque en la imagen o utilizándose técnicas de Deep Learning con redes neuronales.

Seguimiento de objetos: Tras ser detectado, se pueden realizar tareas de seguimiento de un objeto. Podrá efectuarse dicho seguimiento teniendo en cuenta sus propiedades (texturas, bordes, etc) o analizando su desplazamiento respecto a imágenes anteriores.

La mayoría de sistemas de visión artificial o visión por computador están compuestos por dos elementos principales: El sistema de adquisición de imágenes y el sistema de procesado de imágenes. El primero se compone de la iluminación, captura de imágenes y sistemas de adquisición de señales. El segundo implementa los algoritmos de visión que procesan las imágenes para extraer información de ellas.

- **El sistema de iluminación:** está compuesto por todos los elementos que iluminan los objetos con cualquier tipo de radiación electromagnética. Como ejemplo de estos artefactos podríamos citar la luz natural del sol, o la luz artificial proporcionada por lámparas, lasers o leds.

- **El sistema de captura de imagen:** Transforma en señales eléctricas la luz que se refleja en los objetos. El elemento más utilizado son las cámaras, tanto de espectro visible como de espectro invisible.

- **El sistema de adquisición de señales:** Las imágenes capturadas por las cámaras se utilizan para generar señales de vídeo. Su principal objetivo es enviar la señal de vídeo a la entrada de datos del ordenador.

- **Sistema de procesamiento:** Suele ser un ordenador u otro dispositivo con capacidad de computación que implementa los algoritmos necesarios para procesar la imagen digital y para elaborar la información requerida por el sistema de visión artificial

El sistema de procesamiento de imágenes suele componerse de las siguientes fases:

Preproceso: Durante esta fase la imagen puede ser adaptada para extraer mejor la información requerida por los algoritmos o métodos usados. El principal objetivo de esta

fase es obtener una mejor calidad de la imagen de entrada , usando técnicas como filtrados de ruidos, convolución, resaltado de bordes etc

Segmentación: En esta fase la imagen es dividida en áreas de interés. Por ejemplo diferenciando objetos cuadrados de objetos esféricos o seleccionando las líneas de la carretera obviando el resto de la imagen. Para este propósito se pueden utilizar varias técnicas: umbrales, discontinuidades, crecimiento de regiones, filtros de color, detección de movimiento, etc

Clasificación: Una vez la imagen ha sido dividida por regiones de interés (Regions of Interest), se pueden extraer las características específicas de cada una. Esto puede realizarse por análisis morfológico , por texturas o usando técnicas de clasificación de color.

- Sistemas Periféricos: Se trata de los elementos receptores de información, incluyendo monitores, dispositivos que usan la información para tomar decisiones etc.

Hoy en día , las aplicaciones de la visión por computador están creciendo muy deprisa, debido a la disponibilidad de hardware barato que es capaz de ejecutar complejos algoritmos de visión artificial en un tiempo razonable. Por ejemplo podemos encontrar aplicaciones en robótica para vehículos no tripulados, en medicina la visión artificial ya ayuda en diagnósticos mediante análisis de imágenes de los pacientes (cáncer, enfermedades degenerativas , etc), en astronomía ayuda a generar imágenes de mayor calidad etc. Una de las aplicaciones más populares para la visión artificial es el reconocimiento de caracteres (OCR). Es propósito de estos sistemas son la identificación de caracteres , por ejemplo hay aplicaciones que te permiten sacar una foto a la lista de componentes de un producto envasado y la aplicación gracias al OCR puede revisar todos los ingredientes del alimento y avisar si el producto contiene algún elemento al que pueda ser alérgico el usuario como por ejemplo el gluten.

1.2. Visual SLAM

Visual Simultaneous Localization And Mapping (Localización y Mapeo Visual Simultáneo), no se refiere a un algoritmo en particular o a una aplicación software. Se refiere al proceso de estimar la posición y orientación de una cámara con respecto al mundo que le rodea, mientras que simultáneamente mapea el entorno que rodea a la cámara gracias a un procedimiento que analiza y extrae información de las imágenes capturadas por dicha cámara.

Hay varios tipos de tecnología SLAM, algunos no necesitan ni siquiera cámara. Visual SLAM es un tipo específico de sistema SLAM que se basa en algoritmos de visión 3D para

realizar tareas de autolocalización y mapeo cuando ni la localización de la posición de la cámara ni el entorno son conocidos.

La mayoría de los sistemas SLAM funcionan estimando la posición de un conjunto de puntos en varias imágenes sucesivas y así triangular la posición 3D, mientras que simultáneamente utiliza esta información para dar un posición aproximada de la cámara. Básicamente, el objetivo de estos sistemas es hacer un mapa de sus alrededores de su localización para así poder realizar tareas de navegación por el entorno.

Esto es posible con una sola cámara, al contrario de otros tipos de tecnología SLAM. Mientras existan un número suficiente de puntos que puedan ser seguidos a través de varios fotogramas, tanto la orientación del sensor de orientación como la estructura física del entorno pueden ser rápidamente estimados.

Todos los sistemas Visual SLAM están constantemente trabajando para minimizar el error de reproyección, o la diferencia entre los puntos reales y los puntos proyectados, para ello suele utilizar una solución llamada Bundle Adjustment.

Visual SLAM es todavía una tecnología emergente, pero con muchísimo potencial. Será una parte importante en aplicaciones de realidad aumentada, ya que esta tecnología es capaz de mapear el mundo físico con gran exactitud. También se utilizará en una gran variedad de robots , por ejemplo los robots que se envían a Marte utilizan sistemas de Visual SLAM para navegar de forma autónoma. De la misma manera drones y robots en agricultura pueden utilizar esta misma tecnología para moverse por campos de cultivo, incluso los vehículos autónomos podrían utilizar sistemas Visual SLAM para mapear y entender el mundo a su alrededor. Otro gran potencial del VisualSLAM es que permite reemplazar el GPS en ciertos entornos, ya que el GPS no es muy útil en interiores y en grandes ciudades , donde el GPS tiene una exactitud de metros mientras que con Visual SLAM no existirían estos problemas y además tiene mayor exactitud

1.3. Aplicaciones en VisualSLAM

Hoy en día VisualSLAM ya tiene muchas aplicaciones y aún más que están por llegar en un futuro próximo, a continuación se expondrán varios ejemplos de aplicaciones, desde teléfonos móviles hasta robots aspiradora.

- 1. Proyecto Tango** El proyecto Tango es un proyecto colaborativo que trata de equipar a los smartphones y Tablets con sistema operativo Android la capacidad de medir la profundidad a la que se encuentra cada píxel de las imágenes capturadas por la

cámara. Para ello los dispositivos compatibles con Tango dispondrán de 2 cámaras, una cámara RGB y otra que captura la profundidad, así el smartphone es capaz de construir un mapa en 3D del entorno (Figura 1(c)). Los sensores del smartphone son capaces de tomar más de 250 millones de medidas 3D por segundo y con estos datos pueden construir un modelo 3D de los alrededores del teléfono. Las posibilidades que



Figura 1: El primer smartphone compatible con Tango de Lenovo(a). El primer Smartphone compatible con Tango y DayDream de ASUS (b).Esquema de prototipo de smartphone Tango (c).

ofrecerán este tipo de dispositivos serán muy variadas, desde medir las dimensiones de la habitación, hasta lo más útil como guiar a personas con discapacidades visuales en el interior de edificios. Pero también tendrá utilidades para el entretenimiento como convertir una habitación en el escenario de un juego mediante realidad aumentada.

Al ser una tecnología nueva aún no hay un elevado número de dispositivos que lo soporten. De momento existen 2 móviles compatibles con Tango ¹, el Lenovo Phab 2 pro (Figura 1(a)) y el Asus Zenfone AR (Figura 1(b)). En el caso del Zenfone AR estará equipado con 3 cámaras traseras, una para seguir objetos (motion tracking), otra para detectar profundidad y otra de alta resolución de 23 MP. Con estas 3 cámaras el smartphone podrá crear una modelo tridimensional del entorno y seguir su movimiento. La cámara de localización permitirá al ZenFone conocer su posición 3D en todo momento mientras se mueve por el entorno. La cámara de profundidad está equipada con un proyector de Infrarrojos que le permite medir distancias hasta los objetos en el mundo real.

2. Magic Plan

Magic Plan es una aplicación que permite de forma interactiva obtener

¹<https://get.google.com/tango/>

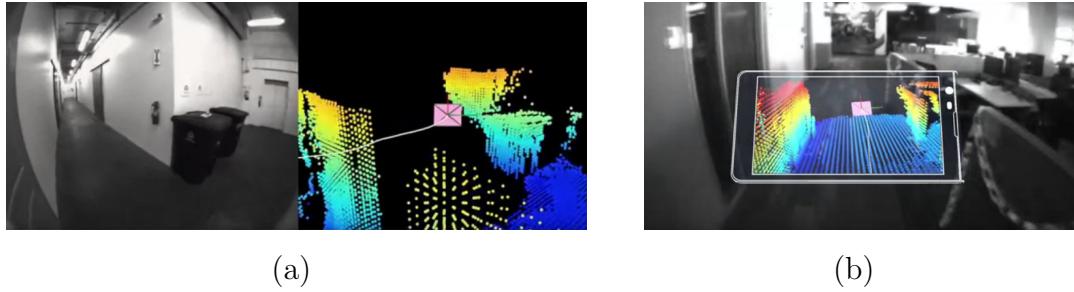


Figura 2: Generación de mapa 1 (d). Generación de mapa 2 (e)

planos de habitaciones o del interior de un edificio, utilizando para ello la cámara de nuestra tablet o smartphone, sólo es necesario sacar fotos. Esta aplicación es gratuita, aunque si se desea obtener el plano en formato digital (pdf, jpg, csv y otros) será necesario pagar una pequeña cantidad de dinero. Es muy sencilla de utilizar y en cuestión de minutos se obtiene un plano fiable (Figura ??) sin necesidad de medir, dibujar, mover muebles y sin necesidad de ser un experto. La aplicación utiliza técnicas de VisualSLAM y se apoya también en la información de los giroscopios de los dispositivos. Es compatible con Android y dispositivos Apple.

En el caso de Android, actualmente la última versión es compatible con el sistema Tango, por tanto el procedimiento de captura es mucho más sencillo, robusto y preciso ya que permite detectar con mayor exactitud todas las paredes de la habitación, visualizarlas en 3D y aplicar realidad aumentada.

3. **Pix4D** Pix4D ² es un software especializado en fotogrametría. Permite la posibilidad de generar mapas 2D y 3D desde fotografías. Las imágenes pueden ser transmitidas vía wireless a Pix4DDim para procesarlas y convertirlas a mapas 2D y 3D. Posteriormente esta información será accesible desde la nube para poder analizarlas y compartirlas. Pix4D permite crear mapas con exactitud a partir de fotografías de interiores, también tiene aplicaciones en minería para medir superficies y volúmenes (Figura 3(a)) de minas a cielo abierto, incluso se utiliza con finalidades forenses para recrear en 3D escenarios de accidentes, que posteriormente pueden ser analizadas con todo detalle. También tiene aplicaciones en la agricultura para obtener mapas de cosechas utilizando la información que proporcionan las cámaras especiales como la Parrot Sequoia (Figura 3(b)). Con la aplicación Pix4DCapture podremos controlar un dron desde nuestro smartphone para que genere un mapa. El dron puede volar de forma autónoma siguiendo algunas de las trayectoria de vuelo que trae por defecto

²<https://pix4d.com/>

el producto (Figura ??) o también puede generar el mapa mientras lo teledirigimos.

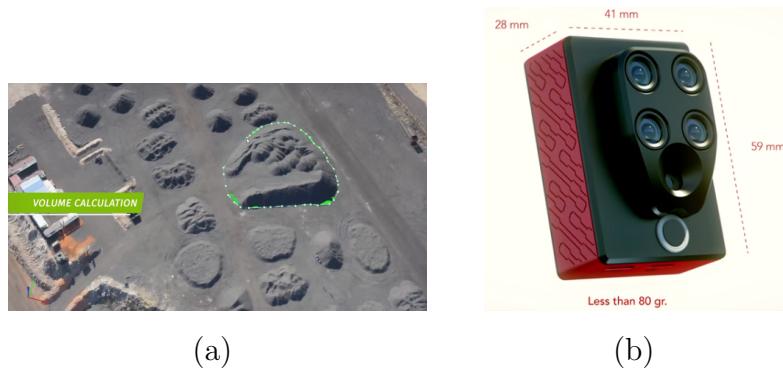
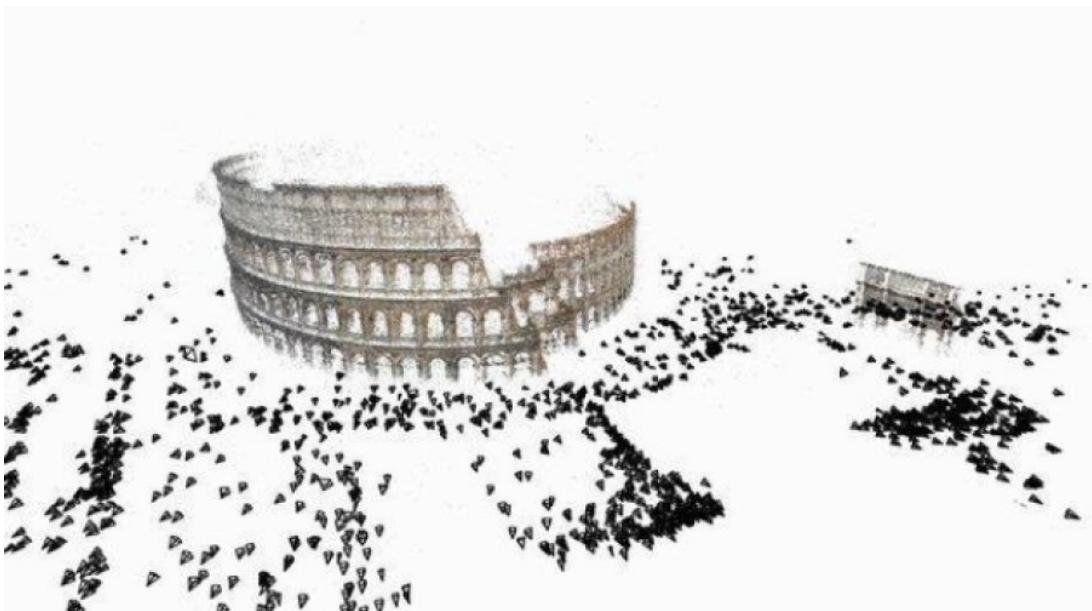


Figura 3: Pix4D cálculo de volumen(a). Cámara multiespectral Parrot Sequoia (b)

4. Photo Tourism PhotoTourism o Photo Synth es un software inicialmente creado por la universidad de Washington en colaboración con Microsoft. Es un sistema que toma grupos de conjuntos de fotografías disponibles online sobre un lugar en concreto, normalmente sobre un monumento turístico mundialmente conocido (como NotreDame, el Coliseo (Figura 4(a)), La Fontana de Trevi) y es capaz de reconstruir puntos 3D de los monumentos y también calcular o estimar la posición de la cámara desde donde se tomaron las fotografías. Proporciona una nueva forma de navegar a través de fotografías de un destino turístico y una nueva forma de hacer visitas virtuales a monumentos. Este sistema utiliza la técnica de *Structure From Motion* SFM. SFM encuentra coincidencias de puntos característicos entre distintas fotografías de un mismo lugar y que han sido tomadas desde distintos puntos de vista y así es capaz de calcular la localización 3D de dichos puntos característicos y también la localización 3D desde donde se tomaron las fotografías. A diferencia de VisualSLAM, el procesamiento de estas fotografía es offline, sin necesidad de tiempo real, por lo que pueden ser ejecutadas desde un PC que por lo general tiene una capacidad de computación mucho mayor que una tablet o teléfono móvil.

5. Canvas y el sensor Structure Canvas³ es una herramienta de escaneo 3D enfocada a profesionales de la construcción o incluso aficionados al bricolaje en casa. La aplicación se ayuda del sensor de profundidad Structure. Este sensor se acopla en la parte trasera de un Ipad. Canvas permite obtener los planos en 3D de cualquier habitación de una manera fácil y sencilla, simplemente tendremos que pasear el Ipad

³<https://canvas.io/>



(a)

Figura 4: Recreación del Coliseo de Roma generado con Photo Tourism.

equipado con el sensor Structure⁴ alrededor de la habitación y podremos ver como el mapa 3D comienza a generarse en tiempo real. El sensor (Figura 5(a)) toma miles de medidas de profundidad que utilizará para generar el plano tridimensional. Los planos son almacenados en el Ipad y pueden ser consultados de manera interactiva posteriormente. Además permite que los planos generados sean convertidos a ficheros CAD.



(a)

Figura 5: El sensor de profundidad Structure para Ipad.

⁴<https://structure.io/>

1.4. Visual SLAM en Robótica Móvil

Visual SLAM tiene aplicaciones directas en robótica. Un ejemplo podría ser el Robot Gita de Piaggio.

1. El robot Gita: Este novedoso robot (Figura 6) tiene incorporadas varios pares de cámaras estéreo, en la parte trasera y delantera. Con las imágenes captadas por estas cámaras se puede realizar VisualSLAM, además es capaz de seguir a su dueño siempre y cuando el humano lleve un cinturón con otras 2 cámaras estéreo, esta funcionalidad se consigue comparando el SLAM del robot con el SLAM captado por el cinturón. El robot dispone de un compartimento interior o maletero y tiene suficiente potencia como para poder transportar hasta 20 Kg. Podría ser de gran utilidad a la hora de ir al supermercado, ya que el robot nos seguirá transportando la compra en su interior, no necesitaremos el típico carrito, incluso nos permitiría ir al centro comercial en bicicleta.

Otra utilidad sería en el interior de un hotel, el robot podría realizar las funciones de camarero y hacer servicio de habitaciones transportando la comida directamente a las habitaciones del hotel. También podría ser un estupendo ayudante para un mecánico, ya que podría transportar la pesadas herramientas o piezas ⁵.



(a)



(b)

Figura 6: El cinturón con cámaras estéreo (a) La capacidad de carga del robot Gita (b)

2. Reconocimiento de Objetos: Otra utilidad de SLAM es que mejoran la capacidad de los robots móviles a la hora de reconocer objetos. Los sistemas de reconocimiento

⁵<http://spectrum.ieee.org/automaton/robotics/home-robots/piaggio-cargo-robot>

de objetos utilizarán la información proporcionada por SLAM para mejorar su capacidad de reconocimiento. La capacidad de reconocimiento será muy útil para aquellos robots que tengan que manipular objetos en su entorno. Con SLAM, los sistemas de reconocimiento pueden tomar como entradas varias imágenes desde distintos puntos de vista, por lo tanto el reconocimiento resulta más sencillo que si tuviesen tan sólo una imagen estática ⁶.

3. **Robot Aspirador:** Recientemente ha entrado en los hogares el uso de VisualSLAM gracias a los últimos modelos de aspiradora equipados con cámaras. Estos aspiradores robotizados disponen de cámaras que le permiten obtener un mapa de la habitación o planta del edificio y gracias a este mapa son capaces de aspirar toda la superficie del suelo de la habitación de manera eficiente, sin dejar ninguna zona de la planta sin limpiar. Además están equipados con sensores de proximidad, que les permiten esquivar obstáculos y aunque tengan que modificar su recorrido momentáneamente son capaces de seguir limpiando ya que pueden utilizar el mapa para continuar su ruta. Entre los distintos aspiradores estarían:

- Aspirador Dyson 360 Eye (Figura 7(a))⁷.
- Aspirador Roomba 966 (Figura 7(b))⁸.
- Aspirador LG-Hombot (Figura ??)⁹.

Tanto el modelo de Dyson como Roomba utilizan una cámara de 360 grados, en cambio el modelo de LG utiliza una doble cámara, y es capaz de aspirar la casa incluso en la oscuridad.

⁶<http://www.roboticsproceedings.org/rss11/p34.pdf>

⁷<http://www.dyson.com>

⁸<http://www.irobot.es/robots-domesticos/aspiracion>

⁹<http://www.lg.com/es/aspiradoras/lg-VR64702LVMT>



Figura 7: Robot Dyson 360 Eye (a) Robot Roomba 966 (b)

4. **Drones:** Por último no podemos olvidar los drones, robots voladores equipados con cámara que también pueden obtener mapas de su entorno con VisualSLAM. Existen también proyectos para equipar a drones con dispositivos compatibles con Tango para que sean capaces de obtener mapas de interiores con mayor precisión, robustez y velocidad ¹⁰.

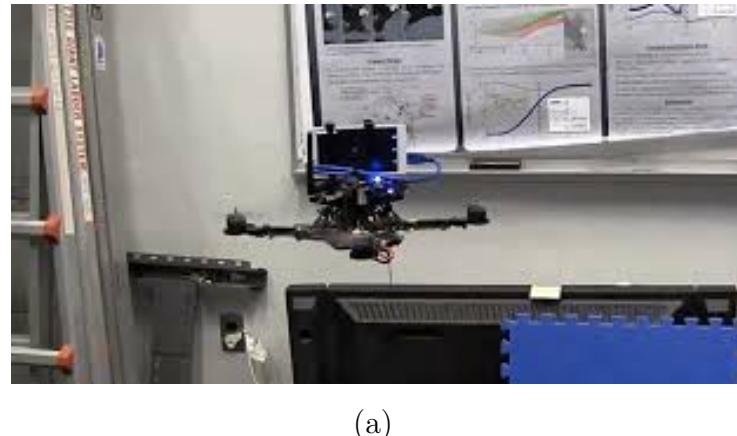


Figura 8: Dron equipado con dispositivo compatible con Tango

¹⁰<http://spectrum.ieee.org/automaton/robotics/drones/autonomous-quadrotor-flight-based-on-google-project-tango>

1.5. Conceptos

Conviene explicar una serie de conceptos relacionados con Visual SLAM ya que aparecerán más adelante cuando expliquemos en profundidad los algoritmos más importantes que existen hoy en día para localización visual.

Calidad: La calidad del algoritmo dependerá de tres factores. La eficiencia temporal, la precisión espacial en la estimación de la posición y la robustez del algoritmo.

Eficiencia: Mediremos la eficiencia como el tiempo de ejecución de cada iteración del algoritmo. Los algoritmos deberán ser capaces de procesar al menos 30 fotogramas por segundo

Precisión: El error lineal y error angular entre la posición estimada y la posición lineal determinará la precisión. Mediremos el error lineal como la distancia euclídea entre las dos posiciones, mientras que el error angular vendrá determinado por la diferencia entre las 2 orientaciones.

Robustez: Diremos que el algoritmo de localización es robusto siempre que pueda seguir funcionando con normalidad tras encontrarse con una situación imprevista (como una oclusión, secuestros, movimiento de objetos en la escena, mala calidad de imagen, etc)

Hipótesis múltiple: Los algoritmos pueden manejar al mismo tiempo múltiples hipótesis candidatas como solución a la localización. Esto se producirá con frecuencia en aquellos entornos donde aparezcan simetrías o zonas que parezcan visualmente similares para el algoritmo.

Oclusiones: Cuando la cámara del robot esté tapada parcial o totalmente se producirá una oclusión, por tanto no se podrá extraer información en la región de la imagen ocluida. Los algoritmos deben estar preparados para cuando existan oclusiones.

Secuestros: Se producirá un secuestro cuando la cámara o robot sea movida deliberadamente por un tercero, de tal forma que los cálculos de la posición anterior ya no sean válidos. Los algoritmos deberían detectar secuestros e intentar localizarse desde su nueva posición.

Localización Absoluta: se produce cuando tenemos un mapa conocido, si estimamos la posición del robot dentro del mapa en coordenadas respecto del origen de coordenadas de ese mapa.

Localización Incremental: En entornos con mapa desconocido, la localización del robot se establecerá de forma incremental respecto de posiciones previas pasadas (por

ejemplo en el instante anterior), lo que dará lugar a un error en la localización incremental que aumentará con el tiempo.

Dinamismo de la escena: El movimiento de los objetos en la escena suele interferir con las estimaciones de autolocalización visual. No todos los desplazamientos en las imágenes se deben entonces al movimiento propio del robot.

Cierre de bucle: Se produce cuando el robot vuelva a pasar por una zona del mundo que ya haya visitado anteriormente. Si se vuelve a pasar por el punto de origen se puede determinar el error que se ha producido comparando la posición real en el origen con la estimada por el algoritmo de localización.

Relocalización: Si tras un secuestro, el robot consigue recuperarse y estima correctamente la posición absoluta del robot dentro del mapa.

2. Objetivos

En el siguiente apartado se detallaran los objetivos que se pretenden alcanzar en este Trabajo Fin de Máster.

Desde la existencia de Visual SLAM , se están desarrollando algoritmos que permitan a los robots estimar su posición , para poder generar un mapa en 3D del entorno y así poder navegar por el espacio que le rodea. Estos algoritmos son complejos y están compuestos de múltiples etapas, a menudo un ligero cambio en alguna de estas etapas puede hacer que los resultados del algoritmo mejoren significativamente o por el contrario empeoren. Sería muy útil y conveniente contar con una herramienta que permitiese analizar o estimar la bondad de los nuevos algoritmos visual SLAM, por tanto el objetivo de este proyecto es presentar un conjunto de herramientas que permitan comparar el rendimiento de los algoritmos visual SLAM.

Esta herramienta permitirá comparar el rendimiento de los nuevos algoritmos de visual SLAM así como estudiar adaptaciones y mejoras en los algoritmos ya existentes.

En el marco de este TFM nos hemos centrado en una plataforma que permita medir la exactitud de las estimaciones de posición sin tener en cuenta la generación de mapas, es decir se ha puesto toda la atención en comparar las mediciones de tracking dejando el mapping para futuros proyectos.

Entenderemos que un algoritmo A es mejor que otro B cuando el algoritmo A sea capaz de estimar la posición con mayor exactitud que otro algoritmo B en el menor tiempo posible,

por lo tanto en los algoritmos de VisualSLAM se medirá la precisión de las estimaciones de posición y la agilidad entendiéndose esta como el tiempo de proceso dedicado a realizar dichas estimaciones de posición.

La herramienta comparará los ficheros de datos generados en entornos 3D

2.1. Requisitos

Los requisitos principales de este TFM han sido:

R1. La herramienta debe restringirse a la comparación de algoritmos VisualSLAM que utilizan como sensor de visión una cámara. Así quedan descartados los algoritmos que emplean cámaras RGBD

R2. Se restringirá solo a los algoritmos VisualSLAM que empleen una sola cámara.

R3. La herramienta debe ser extensible a los resultados de nuevos algoritmos de VisualSLAM.

R4. Facilidad de uso.

R7. Debe ofrecer una única métrica cuantitativa para la comparación entre varios algoritmos de VisualSLAM.

R8. Debe ofrecer resultados fácilmente legibles y reconocibles para el usuario.

2.3. Metodología y plan de trabajo Como metodología no se ha seguido al pie de la letra ninguna concreta, aunque se han seguido las indicaciones del modelo de ciclo de vida en espiral. Esto es una de las primeras lecciones que se aprenden de las metodologías ágiles, la metodología debe adaptarse al proyecto y no al revés. Dicha premisa cobra más importancia en un proyecto realizado por una sola persona.

Por consiguiente, y enmarcándolo dentro de la metodología en espiral, primero se ha realizado un estudio previo muy amplio que sirve para obtener una visión completa del problema y detectar puntos críticos, y luego se ha ido acotando hasta llegar al desarrollo principal, el cual ha sido revisado y validado en cada iteración. Esta amplitud inicial es importante ya que no sólo nos permite avanzar en todas las vías en paralelo, sino porque ofrece una prueba de concepto para las ramificaciones que se han paralizado en favor del desarrollo troncal.

El proceso de desarrollo ha sido supervisado por los tutores mediante tres herramientas de trabajo: reuniones semanales, definición de hitos y diario de trabajo. Durante las reuniones se debían definir varios hitos de corto o medio plazo en los que se trabajaría

esa semana. Este progreso se puede ver en la página web habilitada para tal uso: <https://jderobot.org/Elias-tfm> Así mismo, el código fuente desarrollado puede encontrarse en:

<https://github.com/FALTAURL>

El plan de trabajo sería dividido en cinco hitos. El primero implicaría el aprendizaje del

3. Problemas de Visual SLAM

Actualmente las técnicas de Visual SLAM presentan algunos problemas o inconvenientes que todavía son difíciles de sortear en la práctica. En esta sección presentaremos algunos de ellos:

3.1. Inicialización del Mapa:

Si en Visual SLAM queremos conseguir una estimación lo más exacta posible de la posición de la cámara es necesario contar con una buena inicialización del Mapa. Se debe contar con un sistema de coordenadas globales definido, y se tomarán puntos de referencia del entorno como puntos iniciales del mapa en el sistema global de coordenadas. Utilizando este método podemos inicializar VisualSLAM en un sistema de coordenadas global en la tierra. La transformación de estos puntos iniciales al sistema de referencia global se realizará mediante homografía.

Objetos de referencia como objetos 3D también se han utilizado para obtener un sistema global de coordenadas, posiciones iniciales de la cámara son estimadas gracias al seguimiento de objetos de referencia. En MonoSLAM por ejemplo se utilizan al menos 4 puntos 3D como objetos de referencia, y la forma del objeto se usa para mejorar el mapa.

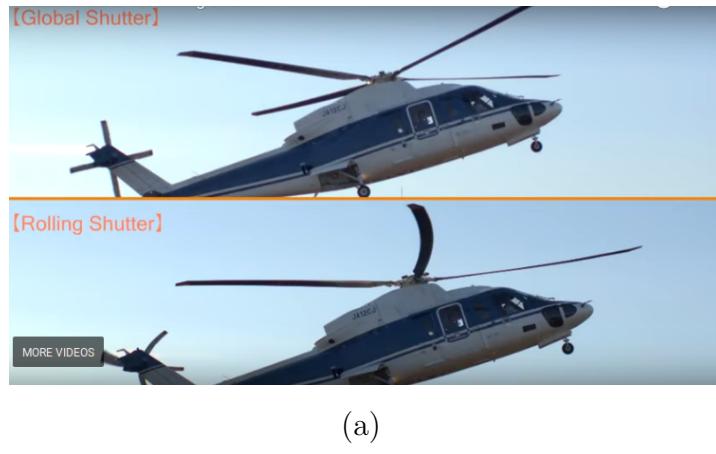
3.2. Ambigüedad en la escala:

En algunas aplicaciones con Visual SLAM se necesita información de escala absoluta. Para obtener una referencia de escala absoluta se pueden utilizar zonas de la anatomía del usuario, como la cara, su mano o el propio cuerpo. En todos estos métodos se asume que entre personas la diferencia de tamaño es mínima para dichas partes del cuerpo. Se han dado otras aproximaciones como utilizar algunos de los sensores con los que ya están equipados

la mayoría smartphones tales como acelerómetros, giroscopios y sensores magnéticos. Para eliminar el ruido de estos sensores se utiliza una técnica de filtro de dominio de frecuencia.

3.3. Distorsión Rolling Shutter:

Para conseguir una estimación de la posición de la cámara exacta, es importante considerar el tipo de shutter en la captura de la imagen. La mayoría de las técnicas de VisualSLAM asumen algoritmos *global shutter*, y estos algoritmos estiman una posición de cámara para cada frame. Sin embargo existen en el mercado multitud de cámaras incluidas las RGB-D que emplean *Rolling Shutter* debido a su coste. En las cámaras que utilizan el modo *Rolling Shutter*, cada fila de una imagen capturada es tomada por una posición diferente de la cámara. Es obvio que es bastante complicado estimar directamente la posición de la cámara para cada fila. Se utilizará una aproximación basada en interpolación para estimar el *Rolling shutter*. En algunas ocasiones se ha utilizado con éxito una función spline para interpolar la trayectoria de la cámara ¹¹.



(a)

Figura 9: Diferencias entre Global Shutter y Rolling Shutter (a)

3.4. Dificultad para operar en entornos con pocas texturas:

Visual SLAM utiliza el emparejamiento de píxeles o puntos característicos entre varios frames consecutivos. El emparejamiento suele fijarse en esquinas, bordes o puntos distintivos que fácilmente podrán localizarse entre frames. Pero cuando en el entorno hay pocas texturas o presenta una alta monotonía de texturas, el emparejamiento es difícil de

¹¹<https://www.premiumbeat.com/blog/know-the-basics-of-global-shutter-vs-rolling-shutter/>

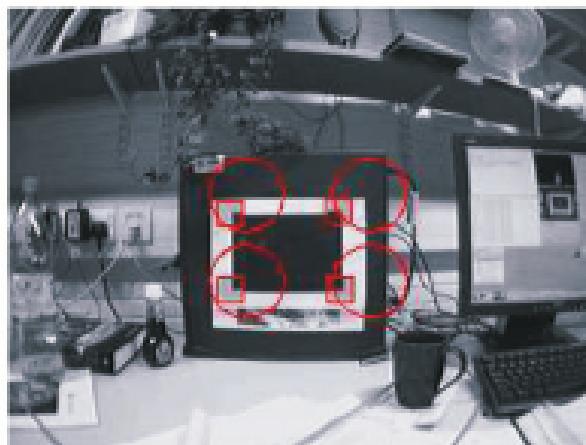
realizar ya que un punto en un fotograma podría corresponder con N puntos en el siguiente fotograma y por tanto se dispararía el error de posición. Quizá este sea uno de los problemas más difíciles de solucionar con VisualSLAM [Takafumi Taketomi, 2017].

4. Estado del Arte

En esta sección explicaremos varios de los algoritmos SLAM más significativos , como MonoSLAM, PTAM, DTAM, SVO,LSD-SLAM,ORB-SLAM,DSO,SDVL y RGBD SLAM. Tambien se describirán brevemente las distintas herramientas que existen en la actualidad para comparar las estimaciones de los algoritmos SLAM , comenzaremos por TUM, seguido de rgb trajectory evaluation , también describiremos la herramienta SLAMBENCH y por último daremos algunos detalles sobre The Kitti Vision Benchmark Suite.

4.1. MonoSLAM

El algoritmo de MonoSLAM (*Monocular SLAM*) [Davison *et al.*, 2007] utiliza solamente una cámara RGB para la localización y mapeo de entornos desconocidos. Fue desarrollado en el año 2002 por Andrew Robinson. Para estimar la posición de la cámara utiliza un filtro extendido de Kalman (EKF) y la posición de una serie de puntos 3D. Este método requiere de una inicialización con al menos 4 puntos 3D conocidos que utilizará para calcular la posición de la cámara y la generación de nuevos puntos para el mapa.



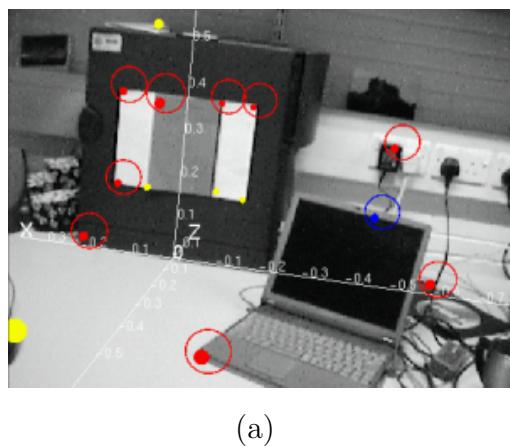
(a)

Figura 10: Inicialización de MonoSLAM con 4 puntos conocidos.

El EKF, tiene un vector de estado compuesto de posición, orientación y velocidad de la cámara y además las coordenadas 3D de los puntos conocidos en un cierto momento, esto implica que el vector de estado irá aumentando de tamaño a medida que vayamos descubriendo nuevos puntos 3D. El modelo de observación estará compuesto de las proyecciones de cada uno de los puntos 3D en el plano imagen.

El uso de un EKF es apropiado, ya que se realizan iteraciones cada pocos milisegundos, y por tanto en intervalos de tiempo tan pequeños, el sistema puede aproximarse a un sistema lineal. Cuantas más iteraciones o frecuencia de muestreo la estimación mejora. En cada iteración se hace una detección de puntos de interés (esquinas con FAST) en la imagen actual de entrada, y obtendremos una serie de puntos que serán candidatos a ser el vector observación de los puntos que queremos seguir. Estos candidatos deberán ser filtrados, pues alguno puede ser un falsa esquina. Se utilizará una función de divergencia ZMSSD (*Zero Mean Sum of Squared Differences*) entre parches para determinar si el candidato es aceptable o no. Al utilizar sólo parches de unos pocos píxeles alrededor del candidato, estamos optimizando el computo ya que no requiere procesar toda la imagen.

Aún así es posible que se acepten puntos candidatos que no sean apropiados. Para tratar de eliminar estos falsos positivos, [Civera *et al.*, 2010] propuso una alternativa conocida como 1-Point RANSAC. MonoSLAM es recomendable para mapas con pocos puntos. Es muy sensible a movimientos bruscos y por tanto difícilmente podrá recuperarse de un secuestro. Si la hipótesis de partida no es correcta el filtro podría desestabilizarse y no llegar nunca a aproximar razonablemente el vector de estado.



(a)

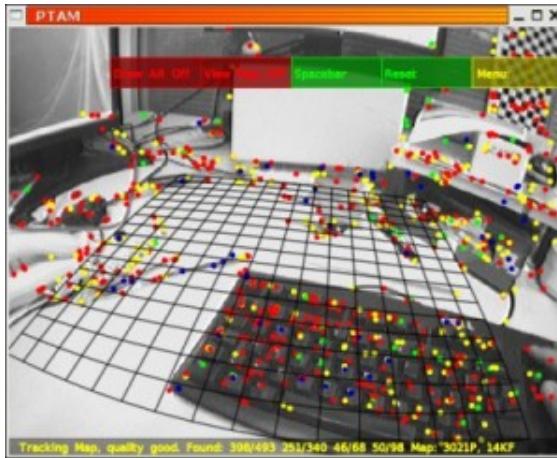
Figura 11: Ejemplo de puntos característicos tomados con MonoSLAM.

4.2. PTAM

Parallel Tracking and Mapping. Es un algoritmo creado en 2007 por George Klein [Klein and Murray, 2007] que también calcula el Tracking y el *Mapping* como en MonoSLAM pero para ello utiliza 2 *Threads*, uno para calcular el posicionamiento de la cámara (Tracking) y el segundo para la generación del mapa (*Mapping*). Esta separación en dos hilos de ejecución se debe a que el Tracking necesita ser calculado en tiempo real para obtener una localización precisa, mientras que el *Mapping* puede demorarse más tiempo sin perjudicar a la localización de la cámara.

Con las imágenes captadas en secuencia se van generando *Keyframes* o fotogramas clave. Se genera un nuevo *Keyframe* a medida que la cámara se va desplazando. Los *Keyframes* se utilizan para la localización y para ir generando el mapa de puntos. Este algoritmo es recomendable para mapas con elevado número de puntos, es capaz de recuperarse fácilmente de un secuestro, extrae los puntos de interés mediante extracción de características como en MonoSLAM y trata de emparejarlos con los puntos extraídos de las imágenes anteriores. Como extractor de características utiliza el detector FAST. Se realizará una subdivisión de la imagen a distintas resoluciones, normalmente 4 niveles, lo que se conoce como pirámide de la imagen y se pasará un filtro FAST sobre esta pirámide para detectar los puntos más característicos de la imagen. Cada *Keyframe* que se genera, contiene la imagen captada junto su pirámide y sus puntos de interés detectados. Cuando añadimos un *Keyframe*, se intenta localizar en este *Keyframe* los puntos que ya se encuentran en el mapa, en caso de no localizarlos se añaden nuevos puntos al mapa. Mientras no se añadan *Keyframes*, se intentará mejorar el mapa con los *Keyframes* disponibles optimizando con Bundle Adjustment.

Se suele utilizar en entornos cerrados y pequeños y utiliza técnicas SFM. Muy utilizado también para realidad aumentada.



(a)

Figura 12: Nube de puntos característicos tomados con PTAM.

4.3. DTAM

Dense Tracking and Mapping. Es un método de reconstrucción de tipo denso que emplea el error fotométrico para poder trabajar en el dominio de la imagen [Newcombe *et al.*, 2011].

La fase de localización (Tracking) se resuelve por una formulación alternativa a EKF utilizando ESM (Minimización Eficiente de segundo orden), de esta forma se puede ejecutar en paralelo. Para la reconstrucción del mapa emplea una metodología basada en la transformada de Radón. Cada píxel 3D será representado como un cubo, que se proyectará a cada una de las imágenes esclavas. Cuando la recta entre estos píxeles sea cero, el error fotométrico será nulo y entonces se podrá considerar que la proyección es correcta. El algoritmo de DTAM está compuesto de los siguientes 3 componentes:

1. Inicialización del mapa, se realiza con medidas estéreo.
2. La posición de la cámara es estimada comparando la imagen de entrada con las imágenes sintéticas generadas por la reconstrucción del mapa.
3. La información de profundidad es estimada para cada píxel usando multi base-line estéreo y es optimizado considerando el espacio continuo.

Para conseguir un buen rendimiento en tiempo real DTAM necesitará el uso GPU.



(a)

Figura 13: Ejemplo de mapa generado con DTAM. Todos los puntos forman parte del mapa.

4.4. SVO

FAST Semi-Direct Monocular Visual Odometry. Permite ser utilizado en ordenadores con poca potencia de computo debido a la rapidez del algoritmo. Es un método híbrido entre los métodos de extracción de características y métodos directos [Forster *et al.*, 2017]. Se asemeja a PTAM en que también utiliza dos *Threads* independientes, el primero para Tracking y el segundo para *Mapping*. En el proceso de Tracking, el algoritmo trata de minimizar el error fotométrico, pero para acelerar el proceso sólo tiene en cuenta ciertas partes de la imagen, unos parches de 4x4 alrededor de los píxeles que se han identificado como candidatos.

Toma los puntos 3D visibles del fotograma anterior, los proyecta sobre la imagen actual, obtiene parches de dimensiones 4x4 alrededor de los píxeles y trata de hallar el mínimo error fotométrico de esos puntos que servirá para hallar el emparejamiento entre las características de dos frames. Calcula el desplazamiento entre imágenes de forma muy eficiente. Para la estimación del movimiento, trataremos de minimizar el error fotométrico de los parches de 4x4 anteriores.

Como último paso haremos la minimización del error de reproyección clásica de los métodos basados en características para corregir los residuos que genere el paso anterior, los cuales podrían provocar la pérdida de ortogonalidad. En cuanto al *Mapping* utilizaremos un modelo gaussiano en torno al valor de profundidad real, cuando la incertidumbre de un

parche decae, este es añadido al mapa. Un nuevo frame tiene posibilidad de convertirse en *Keyframe* si diverge lo suficiente del resto.

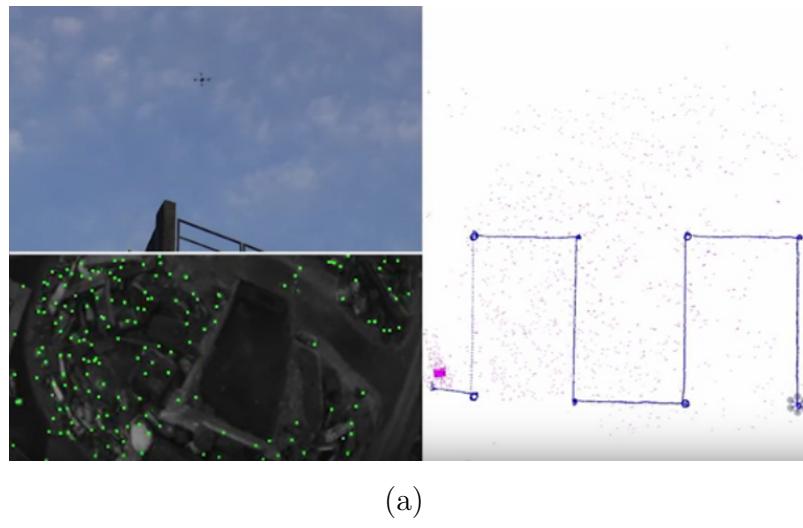
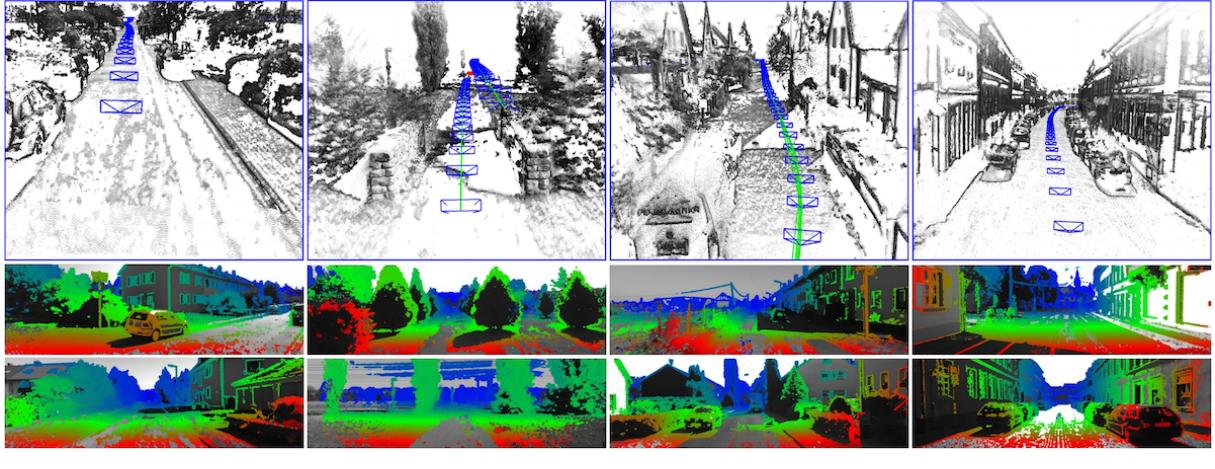


Figura 14: Mapa generado con un dron utilizando la técnica SVO.

4.5. LSD-SLAM

Large-Scale Direct Monocular SLAM La principal característica de este modelo es que trata de generar mapas del entorno a gran escala y consistentes. Utiliza para ello métodos directos. A demás de tener 2 hilos como PTAM uno para *Tracking* y otro para *Mapping*, existe un tercer componente encargado de estimar la profundidad del mapa.[Engel *et al.*, 2014] El *Thread* de *Tracking*, parte de un *Keyframe* para calcular el desplazamiento, minimizando el error fotométrico que estará normalizado por la varianza. Utiliza una optimización ponderada de Gauss-Newton para medir la alineación entre frames. El *Thread* estimador de profundidad, inicializa el mapa de profundidad proyectando los puntos del *Keyframe* anterior. Las imágenes que no son *Keyframes* se usarán para refinar el *Keyframe* actual. Se añadirán nuevos píxeles al mapa de profundidad cuando se encuentren zonas de la imagen con suficiente separación estéreo. En cuanto al *Thread* dedicado al proceso de *Mapping*, cuenta con un mecanismo de cierre de bucle que se ejecutará cada vez que llegue un nuevo *Keyframe*. En cuanto a su inicialización, solo utiliza una única imagen para generar un mapa inicial de profundidad que irá convergiendo hacia unos valores de profundidad correctos a medida que la cámara se vaya desplazando. Este método es capaz de funcionar en tiempo real en un PC, pero no funciona muy bien en dispositivos con limitada potencia de CPU.



(a)

Figura 15: Mapa generado con LSD-SLAM y cámara estéreo

4.6. ORB-SLAM

Es un algoritmo basado en extracción y emparejamiento de píxeles característicos mediante descriptores ORB, estos descriptores son más fiables que los parches tradicionales y por tanto permiten obtener mapas robustos y precisos tanto en escenarios de grandes dimensiones como en zonas pequeñas, sin embargo para su funcionamiento en tiempo real requiere la utilización de ordenadores con alta capacidad de proceso [Mur-Artal *et al.*, 2015]. Puede ser utilizado con una cámara o dos e incluso con cámaras de profundidad RGBD. Para cierres de bucle y relocalización utiliza un modelo de bolsa de palabras [Gálvez-López and Tardos, 2012]. Utiliza 3 *Threads*, el primero para Tracking, el segundo para *Mapping* y un tercero para detectar cierres de bucle.

En el proceso de Tracking, se trata de calcular la posición actual a partir de los emparejamientos encontrados de los puntos 3D en el fotograma anterior, para ello utilizará los descriptores ORB. En caso de perdida, el robot podrá relocalizarse gracias a un modelo de bolsa de palabras que le permitirá encontrar *Keyframes* candidatos que concuerden con la observación actual (Figura 16(a)).

En el proceso de *Mapping*, se inicializarán 2 mapas, uno por homografía y el segundo mediante una matriz fundamental. Los 2 mapas recibirán una puntuación y se elegirá como candidato para inicializar el mapa aquel que obtenga mayor puntuación. Cuando ya se dispone del mapa inicial, se procesan los *Keyframes* creando nuevos puntos 3D y se optimiza localmente el mapa mediante Bundle Adjustment. A su vez se genera un grafo donde cada *Keyframe* se corresponde con un vértice y un vértice estará unido a otro

siempre y cuando los *Keyframe* tengan varios puntos 3D en común. Este grafo permite la eliminación de *Keyframe* redundantes (Figura 16(b)).

En el proceso de Looping, se comprobará si se ha producido un cierre de bucle. Utilizando el grafo de *Keyframe* conectados y el modelo de bolsa de palabras se intenta encontrar *Keyframe* candidatos que tengan una apariencia similar a la imagen actual.

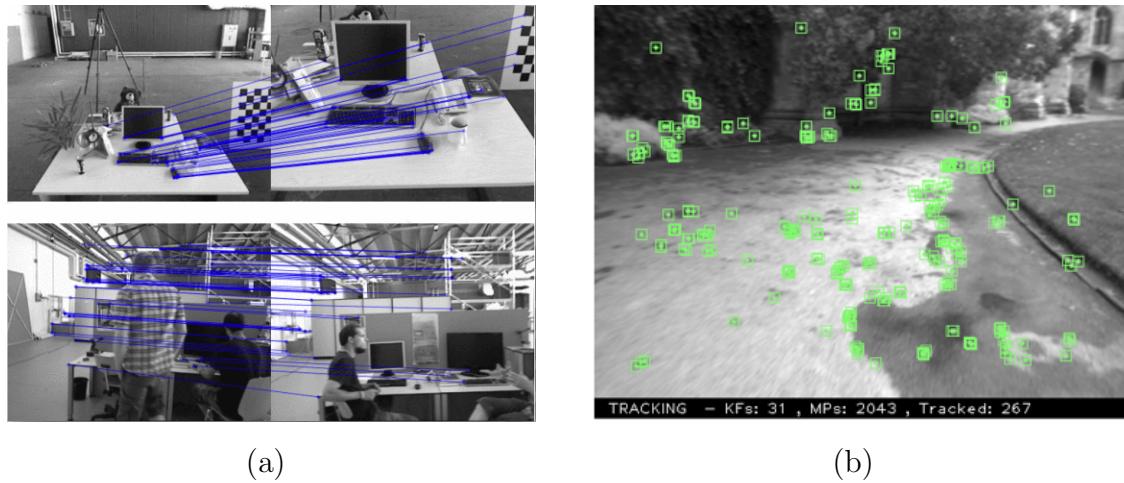


Figura 16: Localización de puntos característicos en 2 imágenes con ORB

4.7. DSO

Direct Sparse Model. Está basado en optimizaciones continuas del error fotométrico sobre una ventana de frames recientes[Engel *et al.*, 2016]. El inicio del Tracking, cuando se crea un nuevo *Keyframe*, todos los puntos activos son proyectados en el y ligeramente dilatados, creando así un mapa de profundidad semi denso. Nuevos frames son creados con respecto a este frame utilizando alineamiento directo de 2 frames, una pirámide multi escala y un modelo de movimiento constante a inicializar. Para la relocalización, se podrán trazar hasta 27 rotaciones pequeñas en diferentes direcciones. Esta recuperación de posición se consigue en el nivel más pequeño de la pirámide de la imagen. La creación de *Keyframes* es similar a ORB-SLAM, existen 3 criterios para determinar cuando se necesita un nuevo *Keyframe*.

1. Se creará un nuevo *Keyframe* (Figura 17(a)) cuando la imagen de entrada cambie notablemente con respecto al último *Keyframe*, esto se medirá con la diferencias de medias al cuadrado entre los píxeles.

2. La translación de la cámara causa occlusiones y des-occlusiones, lo cual indica que se deben generar nuevos *Keyframes*
3. Si el tiempo de exposición de la cámara cambia significativamente, se deberá tomar un nuevo *Keyframe*. Esto se mide por el factor de brillo relativo entre 2 frames.

En cuanto al rechazo de *Keyframes*, sigue la siguiente estrategia. Sean $I_1 \dots I_n$ un conjunto de *Keyframes* activos, siendo I_1 el más nuevo y I_n el más antiguo

1. Siempre se mantendrán los dos últimos *Keyframes* (I_1 e I_2)
2. Frames con menos del 5 % de sus puntos visibles en I_1 son descartados.
3. Si mas de N frames están activos, se descartan (exceptuando I_1 e I_2) aquel que maximice un marcador de distancia $d(i,j)$ donde $d(i,j)$ es la distancia Euclídea entre *Keyframes* I_1 e I_j

Sobre el tratamiento de los puntos, siempre se tratará de mantener un numero fijo de puntos activos repartidos de forma uniforme entre el espacio y los frames activos. En un primer paso, se identifican N_p puntos candidatos en cada nuevo *Keyframe*. Los puntos candidatos no son inmediatamente sumados a la optimización, sino que son localizados individualmente en sucesivos frames generando una primera estimación del valor de profundidad que servirá como inicialización.

En cuanto a la selección de puntos candidatos, se intentará seleccionar aquellos puntos que están bien distribuidos en la imagen y tienen un valor elevado de gradiente con respecto a sus alrededores. Para obtener una distribución uniforme de puntos sobre la imagen, esta se divide en bloques de $d \times d$, de cada bloque se elegirá el píxel con el mayor gradiente siempre y cuando supere un umbral, de lo contrario no se selecciona el píxel de ese bloque. Los puntos candidatos son localizados en siguientes frames utilizando una búsqueda sobre la línea epipolar minimizando el error fotométrico. Una vez hallamos encontrado las coincidencias preparamos un valor de profundidad y la varianza asociada que se utilizará para restringir el intervalo de búsqueda en frames siguientes. Esta estrategia de localización está inspirada en LSD-SLAM.

Por último, la activación de puntos candidatos, cuando un conjunto de puntos antiguos son marginados, nuevos puntos candidatos son activados para remplazarlos, siempre intentando mantener una distribución uniforme de puntos por toda la imagen ¹².

¹²<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7898369>



(a)



(b)

Figura 17: Mapa generado con DSO (a) Ligero error en la posición al volver al punto de partida (b).

4.8. SDVL

Semi-Direct Visual Localization. Al igual que PTAM este método tiene 2 *Threads*. El primer *Thread* se utiliza para Tracking. Calcula el desplazamiento de la cámara entre dos observaciones utilizando el mapa 3D y las imágenes capturadas. Al ser un método híbrido, calculará los desplazamientos basándose en técnicas de métodos directos y píxeles característicos. El sistema cuenta también con un sistema de rechazo de espurios que ayudará a identificar y eliminar puntos 3D mal posicionados [Perdices García, 2017]. El segundo *Thread* se utiliza para el *Mapping*. Se encarga de crear y actualizar el mapa del entorno del robot. Almacenará P puntos 3D en varias imágenes desde las que se pueden

observar los P puntos 3D. Sólo pasarán a ser parte del mapa aquellos fotogramas que cumplan ciertos requisitos, aquellos que sean considerados fotogramas clave o *Keyframes*. El mapa inicial se obtiene por Homografía. Es un método capaz de manejar miles de puntos en el mapa, aunque no es considerado un mapa denso, por lo que el mapa generado no contiene muchos detalles y no es capaz tampoco de generar mapas grandes. En caso de perdida de la posición de la cámara, es capaz de realizar relocalización aunque no es capaz de detectar bucle cerrado.

4.9. RGB-D Visual SLAM

Las cámaras RGB-D, utilizadas en dispositivos como Kinect o smartphones en el Proyecto Tango, son capaces de proporcionar información 3D del entorno en tiempo real y por tanto estas cámaras también son utilizadas en Visual SLAM.

A diferencia con los algoritmos del tipo monocular VisualSLAM, la escala del sistema de coordenadas es conocida para las cámaras RGB-D ya que son capaces de obtener las medidas y dimensiones de los objetos en 3D del entorno que le rodea. Para estimar el movimiento de la cámara se utiliza el algoritmo ICP *Iterative Closest Point*. La mayoría de cámaras que son capaces de medir la profundidad de los píxeles están creadas para entornos cerrados o de pequeñas dimensiones. Esto es debido a que estas cámaras proyectan un patrón de infrarrojos para medir la profundidad del entorno y es difícil detectar el patrón de infrarrojos emitidos en el exterior ya que la propia luz solar generaría ruido o perturbaciones en el rango infrarrojo. Además el rango de profundidad de pueden captar los sensores RGB-D está limitado de 7 a 9 metros. Para la localización, el movimiento relativo de la cámara es estimado identificando la localización de varios puntos característicos entre frames sucesivos. Utilizando estos puntos característicos se hace la estimación de los valores de una matriz de traslación. Con el algoritmo ICP y mapas de profundidad podemos optimizar esta matriz de traslación. También se utilizan métodos de localización basados en consistencia fotométrica similar a las técnicas utilizada en los métodos densos de Visual SLAM [Takafumi Taketomi, 2017].

Para obtener un mapa geométricamente consistente, se utilizan varios algoritmos de optimización como pose-graph y deformation-graph. Pose-graph se utiliza para reducir el error acumulativo. Pose-graph es muy similar al bucle cerrado en los algoritmos de monocular VisualSLAM. En contraste con otros algoritmos, la estimación de mapa también está refinada. La optimización por Deformation Graph es muy utilizada para ciertos frames y la localización de la cámara es estimada con emparejamientos entre las imágenes RGB-D y

el modelo reconstruido. Las APIs para RGB-D SLAM vienen incorporadas en dispositivos como Google Tango y Structure Sensor. Especialmente, Google Tango proporciona una estimación de resultado estable combinando también la información proporcionada por otros sensores internos del dispositivo.

4.10. Herramientas para realizar evaluaciones comparativas de algoritmos SLAM

En este apartado trataremos sobre varias herramientas que podemos utilizar para hacer benchmarking sobre los resultados de algoritmos SLAM y evaluar y comparar dichos resultados

1. Computer Vision Group TUM

Proporciona varias herramientas para evaluar el rendimiento de algoritmos VSLAM, permite evaluar trayectorias y compararlas con la trayectoria ground truth.[Sturm *et al.*, 2012] Utiliza principalmente 2 métodos, el error absoluto de la trayectoria *absolute trajectory error (ATE)* y el error relativo a la posición *relative pose error (RPE)*. Podemos encontrar en la web HERE scripts descargables para ambas métricas. Las trayectorias que vayan a ser evaluadas deben ser almacenadas en el siguiente formato('timestamp tx ty tz qx qy qz qw')

Absolute Trajectory Error (ATE): El error absoluto de trayectoria mide la diferencia entre puntos entre la trayectoria estimada y la trayectoria real. Como paso de preproceso se realiza una asociación entre la posición estimada con la posición ground truth utilizando el emparejamiento por timestamps o marcas de tiempo. Tras esta asociación, se alinea la trayectoria real con la estimada usando SVD (Singular Value Decomposition). Por último, el ordenador calcula la diferencia entre cada par de posiciones y devuelve valores estadísticos como la media, mediana y desviación estandar de estas diferencias. También es posible obtener gráficos con las 2 trayectorias.

Relative Pose Error (RPE): Con el script python evaluaterpe.py es posible calcular el error relativo a la posición . Este script obtiene el error entre el movimiento relativo entre pares de timestamps. Por defecto el script calcula el error entre todos los pares de timestamps del fichero de trayectoria. Como el numero de pares de timestamps en la trayectoria estimada es cuadrático se pueden poner cotas con un número máximo de pares de timestamps. Opcionalmente, tambien se puede elegir usar

un tamaño de ventana fijo (-fixed delta). En este caso cada pose en la trayectoria estimada es asociado con posteriores posiciones dependiendo del tamaño de ventana (-delta) y unidad (-delta unit). Esta técnica de evaluación es útil para calcular el desvío o deriva.

2. Trajectory Evaluation Toolbox for Visual(-inertial) Odometry

Esta herramienta está desarrollada en python e incluye : Métodos de alineamiento de trayectorias para diferentes modalidades de sensores Métricas de error tales como ATE y Relative Odometry Error. [Zhang and Scaramuzza, 2018]

El software ha sido diseñado para uso fácil. Dados 2 ficheros de texto donde especificaremos la trayectoria estimada y el groundtruth, la herramienta establece automáticamente el emparejamiento de tiempos, realiza alineamiento de trayectoria y calcula distintos errores métricos con una línea de comandos. También se puede utilizar para comparar diferentes algoritmos con múltiples datasets. Para que la aplicación pudiera ser utilizada con diferentes formatos, también se proporcionan varios scripts para convertir otros formatos conocidos (e.g., EuRoC, rosbag) al formato utilizado por la herramienta. El formato utilizado para los ficheros de datos es el siguiente: timestamp tx ty tz qx qy qz qw

3. SLAMBENCH

SLAMBench es una herramienta de la Universidad de Edimburgo, esta herramienta ha sido creada para evaluar sistemas SLAM ya sean sistemas de código abierto o sistemas propietarios sobre un conjunto extensible de datasets y métricas. [Bodin *et al.*, 2018] Actualmente soporta 8 tipos distintos de algoritmos (densos, semi-densos y escasos) y 3 datasets. Es una herramienta que permite la reproductibilidad de los resultados para los sistemas SLAM actuales y posibilita la integración y evaluación de los nuevos resultados SLAM.

SLAMBench soporta varios algoritmos diferentes. Entre los algoritmos escasos o poco densos soportaría los siguientes MonoSLAM, PTAM ,OKVIS y ORB-SLAM2. Los 2 primeros algoritmos soportan solo sistemas monoculares, OKVIS soporta sólo estereoscópico y RGB-D y ORBSLAM2 soporta ambos tipos. Estos algoritmos son algoritmos indirectos. Entre los métodos densos se soportan 3 tipos de algoritmos, KinectFusion, InfiniTAM, ElasticFusion que son 2 recientes modelos densos. Por último LSD-SLAM es un sistema SLAM semi-denso de una sola cámara (monocular).

Para medir el rendimiento de algoritmos SLAM se puede utilizar un framework para cuantificar la calidad de los resultados teniendo en cuenta exactitud, tiempo

de ejecución, uso de memoria y consumo de energía. Esta información puede ser visualizada gracias a su interfaz gráfico. Además SLAMBench ofrece una plataforma con grandes posibilidades para investigaciones futuras ya sea para diseño de algoritmos como optimizaciones a nivel de implementación. Es una herramienta multiplataforma y se puede utilizar en PCs de sobremesa, portátiles y móviles. Algunos benchmarks se han obtenido con Ubuntu, OS y Android. También puede ser utilizado con CUDA. SLAMBench proporciona, entre otras métricas, medidas de exactitud del algoritmo utilizado. Las medidas de exactitud son determinadas comparando datos estimados con los datos ground-truth. Absolute Trajectory Error (ATE) y Relative Pose Error RPE son utilizadas para medir la exactitud de la trayectoria. Estas métricas de trayectoria junto con una métrica de mapeo, Reconstrucción de Error , proporcionan comparaciones cuantitativas para varios algoritmos.

ATE y RPE son calculadas en tiempo de ejecución, con un alineamiento mínimo entre la primera posición más cercana de ground-truth y la posición estimada (en términos de timestamp). Ya off-line, técnicas más complejas de alineamiento pueden ser usadas para comparar técnicas densas y semidensas cuando el mapeo de escalas no funciona. RER se calcula mediante la ejecución del algoritmo Iterative Closest Point (ICP) de los modelos de la nube de puntos de la reconstrucción y del ground truth. Como este proceso consume mucha CPU esta evaluación es también ejecutada off-line.

Otras métricas que proporciona SLAMBench es el consumo de energía ,utilización de memoria, velocidad de proceso por frame.

Interface de usuario modular: SLAMBench también permite elegir entre diferentes sistemas de interfaz de usuario. Métricas de evaluación pueden ser cambiadas y personalizadas, así como el interfaz de usuario gráfico (GUI), mientras mantiene independencia de los datasets y algoritmos. Por ejemplo, el visualizador nativo está basado en la librería Pangolin , puede ser reemplazado por un visualizador ROS.

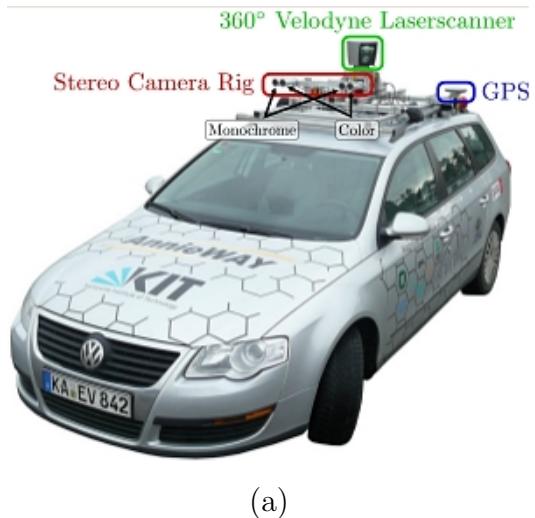
SLAMBench está siendo un componente muy importante en robótica y Sistemas de Realidad Aumentada (AR). Aunque un gran número de algoritmos SLAM han sido presentados, no se ha investigado lo suficiente para tratar de unificar el interface de estos algoritmos, o realizar comparaciones de todas sus capacidades en conjunto. Esto presenta un problema ya que diferentes aplicaciones SLAM pueden tener diferentes requisitos funcionales y no funcionales. Por ejemplo, una solución para Realidad Aumentada desarrollada para móviles tendría que optimizar el consumo de energía,

mientras que otra solución diseñada para vehículos de navegación autónoma estaría enfocada a funcionar con la mayor exactitud posible. SLAMBench2 es un framework de evaluación que compararía sistemas SLAM actuales y futuros, utilizando una lista extensible de datasets, mientras utiliza una lista comparable de métricas de rendimiento. Se podrían utilizar una gran variedad de algoritmos de SLAM y datasets como ElasticFusion, ORB-SLAM2, OKVIS y tambien se podría integrar con nuevos algoritmos y datasets. SLAMBench2 es un software que está disponible de manera pública.

4. The Kitti Vision Benchmark Suite

Es un conjunto de aplicativos que utilizan mapas y secuencias grabados desde la plataforma de coches autónomos Annieway para crear nuevos desafíos o retos en las tareas comparativas o benchmarking de visualslam.[Geiger *et al.*, 2012] Están investigando en varios campos como: vision stereo, flujo optico, odometría, detección de objetos en 3D y seguimiento de objetos 3D. La exactitud de los conjuntos de datos ground truth es medida gracias al scanner laser Velodyne y a los sistemas de localización GPS con los que van equipados sus coches autónomos. Los datasets han sido grabados en la ciudad de Karlsruhe. Además de proporcionar todos los datos en formato raw, para cada uno de sus benchmarks, también proporcionan una metrica de evaluación y una web de evaluación de métricas. En experimentos preliminares se ha comprobado que métodos que obtienen una puntuación alta en algunos benchmarks, cuando son aplicados al mundo real obtienen unos resultados por debajo de la media. El objetivo es reducir esta tendencia y completar los benchmarks existentes proporcionando benchmarks en el mundo real con dificultades novedosas para la comunidad.

Un ejemplo podría ser el benchmark de odometría, que consiste en 22 secuencias stereo, grabadas en formato png. En el dataset se proporcionan 11 secuencias con trayectorias ground truth para entrenar y 11 secuencias sin ground truth para evaluar. Para este benchmark se pueden proporcionar resultados usando una cámara o un sistema de cámaras estereo. La única restricción que se impone es que el metodo deber ser totalmente automático (no se permite el etiquetado manual de cierre de bucle) y que el mismo conjunto de parámetros es usado para todas las secuencias. Para todas las secuencias de test, su evaluador estima los errores de traslación y rotación. En una tabla de evaluación se establece un ranking de métodos de acuerdo con la media de esos valores, donde los errores son medidos en porcentaje para la traslación y en grados por metro para la rotación.



(a)

Figura 18: Coche autónomo Annieway utlizado con Kitti.

4.11. Comparativa de los algoritmos más representativos

A continuación se presenta una tabla que muestra las características principales de cada algoritmo. Esta tabla es similar a la que aparece en [Perdices García, 2017] pero en este caso se ha añadido el algoritmo DSO.

Funcionalidad	Mono-SLAM	PTAM	SVO	LSD-SLAM	ORB-SLAM	SDVL	DSO
Probabilístico	Sí	No	No	No	No	No	No
Hilos de ejecución	1	2	2	3	3	2	2
Emparejamiento	Parches	Parches	Híbrido	Métodos directos	ORB	Híbrido	Métodos directos
Puntos 3D con incertidumbre	No	No	Sí	Sí	No	Sí	Sí
Mapa inicial	Dado	Homograf.	Homograf.	Incertidumbre	Homog. /Matriz F.	Homograf.	Incertidumbre
<i>Keyframes</i>	No	Sí	Sí	Sí	Sí	Sí	Sí
Puntos en mapa	Cientos	Miles	Miles	Miles	Miles	Miles	Miles
Mapa denso	No	No	No	Sí	No	No	Semi-denso
Gestión de mapas grandes	No	No	No	Sí	Sí	No	Sí
Relocalización	No	Sí	Sí	Sí	Sí	Sí	No
Rechazos de espurios	No	No	No	Sí	Sí	Sí	No
Cierre de bucle	No	No	No	Sí	Sí	No	Sí

5. Conclusiones

La robótica móvil es ya una realidad gracias a los algoritmos Visual SLAM que permiten estimar con mínimo error la localización y generación de mapas en entornos desconocidos. En este documento se han descrito algunos de estos algoritmos que ya están funcionando, pero se sigue investigando en la generación de nuevos métodos de navegación autónoma para conseguir mayor fiabilidad, robustez y exactitud de los cálculos.

Dependiendo de las características del entorno o de los requisitos del problema que estemos tratando será más conveniente utilizar un algoritmo u otro. Por ejemplo si necesitásemos generar un mapa de gran exactitud, lo más conveniente sería utilizar DTAM, si por el contrario el mapa no fuese muy importante y la potencia del hardware fuese muy limitada podríamos utilizar SVO.

Por ahora las limitaciones hardware hacen que en robótica móvil se opte por utilizar aquellas técnicas que requieren poca capacidad de cómputo (PTAM,SVO) ya que son fácilmente procesables por los microprocesadores de los actuales robots móviles. En el futuro y a medida que los robots tengan más capacidad de proceso, probablemente se impongan los métodos más robustos que realicen una localización más exacta y cuyos mapas sean muy fiables como podría ser el método ORB-SLAM o DSO.

No obstante todavía queda un camino largo que avanzar en Visual SLAM, ya que algunos algoritmos no son del todo robustos en grandes espacios o entornos donde exista excesivo movimiento alrededor de la cámara, por ejemplo si nuestro robot se encontrase en un jardín frondoso, donde sopla una cierta brisa, le sería difícil al robot mapear el entorno ya que el movimiento de hojas y ramas podría generar inestabilidad en la estimación de la posición 3D de la cámara.

Aunque la gran revolución se producirá cuando la mayoría de smartphones y cámaras estén equipadas con dispositivos que puedan medir la profundidad de las imágenes, como el proyecto Tango. Sin duda los cálculos de mapeo y posición se acelerarán y mejorará notablemente la exactitud de las estimaciones de posición.

No sería de extrañar que próximamente apareciesen nuevos dispositivos periféricos que pudiesen ser controlados por el Smartphone, por ejemplo un nuevo tipo de aspiradora que no tuviese capacidad para realizar Visual SLAM por si sola, solo un par de motores que le permitan avanzar y girar. Si quisiésemos que esta aspiradora comenzase a aspirar de forma autónoma sólo tendríamos que colocar nuestro Smartphone en posición vertical sobre ella. El smartphone comenzaría a mapear la habitación y a dirigir la navegación de la aspiradora hasta que todo el suelo de la habitación quedase limpio. De esta forma todo el proceso de Visual SLAM de la aspiradora quedaría relegada al Smartphone. Y quien sabe, quizás el futuro de la conducción autónoma dependa de la capacidad con la que estén equipados para realizar Visual SLAM los cada vez más potentes Smartphones.

Bibliografía

- [Arribas, 2016] Victor Arribas. Análisis de algoritmos de visualslam: un entorno integral para su evaluación. *Trabajo Fin de Máster. Universidad Rey Juan Carlos*, 2016.
- [Bodin *et al.*, 2018] Bruno Bodin, Harry Wagstaff, Sajad Saeedi, Luigi Nardi, Emanuele Vespa, John H Mayer, Andy Nisbet, Mikel Luján, Steve Furber, Andrew J Davison, Paul H.J. Kelly, and Michael O’Boyle. Slambench2: Multi-objective head-to-head benchmarking for visual slam. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2018.
- [Civera *et al.*, 2010] Javier Civera, Oscar G Grasa, Andrew J Davison, and JMM Montiel. 1-point ransac for extended kalman filtering: Application to real-time structure from motion and visual odometry. *Journal of Field Robotics*, 27(5):609–631, 2010.
- [Davison *et al.*, 2007] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6), 2007.
- [Engel *et al.*, 2014] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.
- [Engel *et al.*, 2016] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *arXiv preprint arXiv:1607.02565*, 2016.
- [Forster *et al.*, 2017] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2017.
- [Gálvez-López and Tardos, 2012] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

- [Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [Hernández, 2014] Alejandro Hernández. Autolocalización visual aplicada a la realidad aumentada. *Trabajo Fin de Máster. Universidad Rey Juan Carlos*, 2014.
- [Jakob *et al.*, 2016] Engel Jakob, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Septiembre 2016.
- [Klein and Murray, 2007] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007.
- [Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.
- [Newcombe *et al.*, 2011] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011.
- [Perdices García, 2017] Eduardo Perdices García. Técnicas para la localización visual robusta de robots en tiempo real con y sin mapas. *Tesis Doctoral. Universidad Rey Juan Carlos*, 2017.
- [Sturm *et al.*, 2012] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.
- [Takafumi Taketomi, 2017] Sei Ikeda Takafumi Taketomi, Hideaki Uchiyama. Visual slam algorithms: a survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, Junio 2017.
- [Zhang and Scaramuzza, 2018] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2018.