



**MÁSTER UNIVERSITARIO
EN VISION ARTIFICIAL**

Curso Académico 2016/2017

Trabajo de Fin de Máster

Estado del Arte de VisualSLAM

Autor: Elías Barcia Mejias

Tutores: José María Cañas Plaza , Eduardo Perdices García

Índice general

1. Introducción	1
2. Aplicaciones en VisualSLAM	3
2.1. Proyecto Tango	3
2.2. Magic Plan	6
2.3. Pix4D	7
2.4. Photo Tourism	8
2.5. Canvas y el sensor Structure	9
2.6. Aplicaciones en Robótica Móvil	10
3. Problemas de Visual Slam	14
4. Técnicas de Visual SLAM	15
4.1. Métodos Densos y Métodos Escasos	15
4.2. Métodos Directos y Métodos Indirectos	16
4.3. MonoSLAM	18
4.4. PTAM	20
4.5. DTAM	21
4.6. SVO	21
4.7. LSD-SLAM	22
4.8. ORB-SLAM	24
4.9. DSO	26

ÍNDICE GENERAL

III

5. Conclusiones

29

Bibliografía

31

Índice de figuras

2.1. El primer smartphone compatible con Tango de Lenovo(a). El primer Smartphone compatilbe con Tango y DayDream de ASUS (b).Esquema de prototipo de smartphone Tango (c).Generación de mapa 1 (d). Generación de mapa 2 (e)	5
2.2. La pantalla de un smartphone utilizando Magic Plan.	6
2.3. Pix4D cálculo de volumen(a). Flujo de datos de Pix4D(b). Trayectorias que pueden seguir los drones con Pix4DCapture(c). Cámara multiespectral Parrot Sequoia (d)	7
2.4. Recreación del Coliseo de Roma.	8
2.5. El sensor de profundidad Structure para Ipad.	9
2.6. El Robot Gita siguiendo a su dueño (a) El cinturón con cámaras estéreo (b) La capacidad de carga del robot Gita (c)	11
2.7. Robot Dyson 360 Eye (a) Robot Roomba 966 (b) Robot Hombot de LG (c).	13
2.8. Dron equipado con dispositivo compatible con Tango	13
4.1. Mapa de clasificación de los principales algoritmos de Visual SLAM.	17
4.2. Ejemplo de puntos característicos tomados con MonoSlam.	19
4.3. Nube de puntos característicos tomados con PTAM.	20
4.4. Ejemplo de mapa generado con DTAM. Todos los puntos forman parte del mapa.	21
4.5. Mapa generado con SVO.El color azul indica proximidad y el rojo lejanía	22
4.6. Mapa generado con LSD-SLAM y cámara estéreo	23
4.7. Localización de puntos característicos en 2 imágenes con ORB	25

4.8. Mapa generado con DSO (a) Ligero error en la posición al volver al punto de partida (b)	28
--	----

Capítulo 1

Introducción

Actualmente la investigación y desarrollo en robótica móvil está en pleno auge. Los robots modernos están equipados con multiples sensores y uno de los sensores 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 aportan cada imagen. Uno 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 enviasemos a nuestro robot a explorar la superficie del planeta Marte donde la señal GPS es inexistente. ¿ Como se las arreglaría el robot para desplazarse por el terreno de forma autónoma sin perderse ? Hoy en día ya existe una técnica 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 basicamente 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 además debe contar con una capacidad de cálculo suficientemente potente que le permita ejecutar un software de visión artificial para procesado de imágenes y al mismo tiempo realizar la generación de mapas. Estas tareas requieren ser ejecutadas con cierta velocidad, unos 30 fotogramas por segundo. 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 estereo que le ayudarán a representar el entorno en 3D con mayor fiabilidad. Otras cámaras 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 robusted y precisión. Es posible utilizar la técnica de visualSLAM hoy en día en pequeños dispositivos gracias al aumento de su potencia de computación.

Capítulo 2

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.

2.1. 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 pixel de las imágenes capturadas por la cámara . Para ello los dispositivos compatibles con Tango dispondrán de 2 cámaras, 1 RGB y otra que captura la profundidad , así el smartphone es capaz de construir un mapa en 3D del entorno .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 ofrecerán este tipo de dispositivos serán muy variadas, desde medir las dimensiones de la habitación , hasta lo más util como guiar a personas don discapacidades visuales en el interior de edificios. Pero tambien 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 y el Asus Zenfone AR. 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

¹<https://get.google.com/tango/>

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.

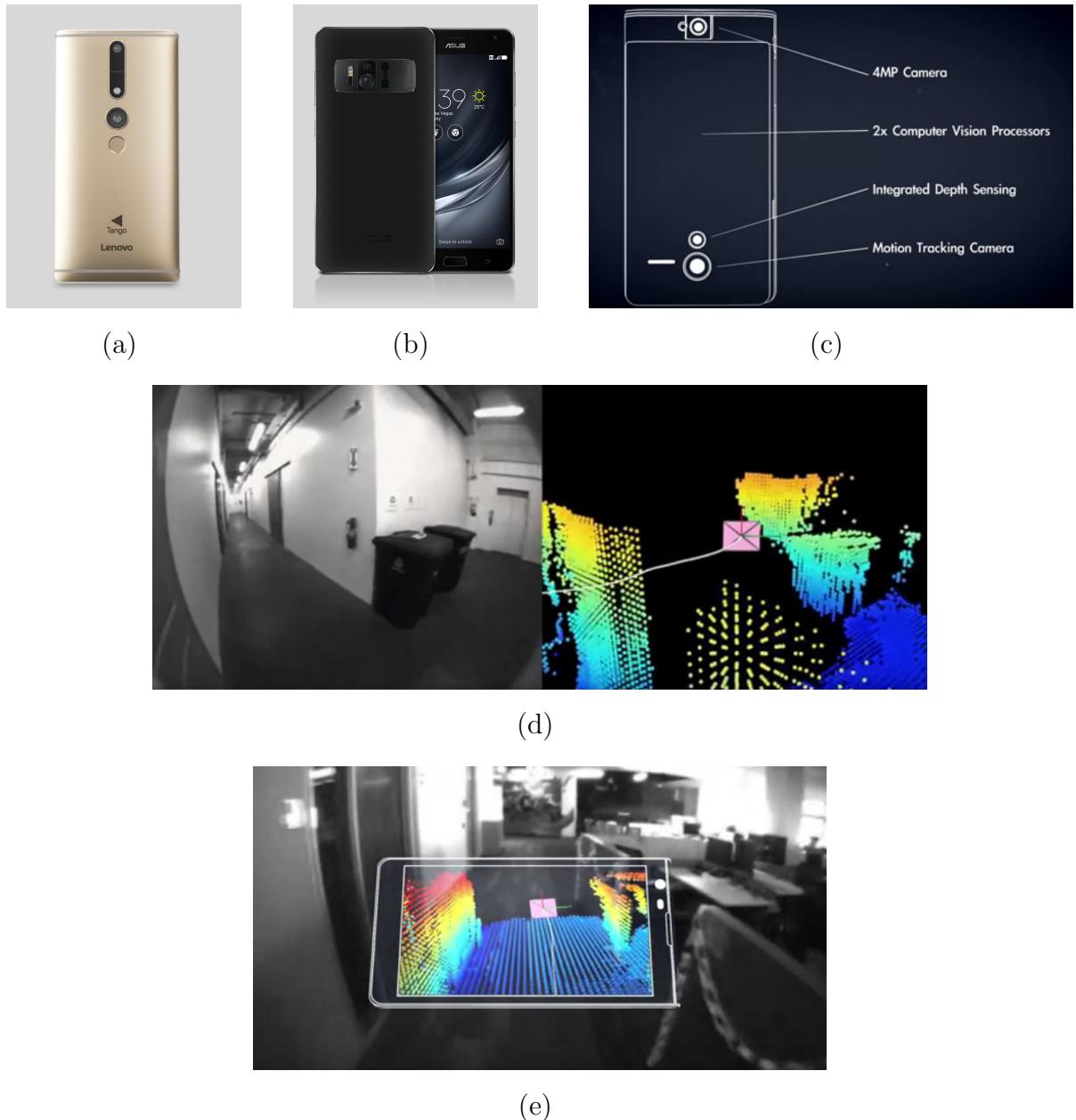
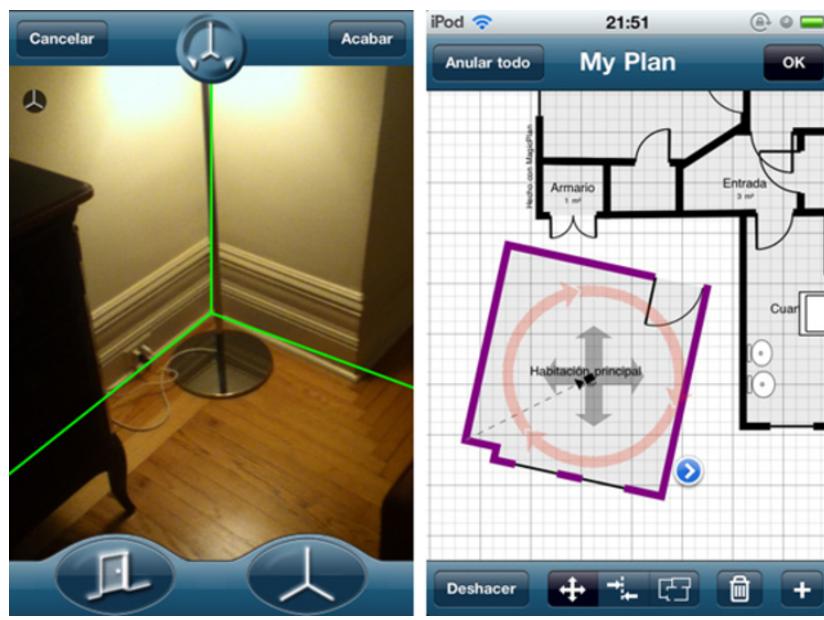


Figura 2.1: El primer smartphone compatible con Tango de Lenovo(a). El primer Smartphone compatible con Tango y DayDream de ASUS (b). Esquema de prototipo de smartphone Tango (c). Generación de mapa 1 (d). Generación de mapa 2 (e)

2.2. Magic Plan

Magic Plan es una aplicación que permite de forma interactiva obtener 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 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.



(a)

Figura 2.2: La pantalla de un smartphone utilizando Magic Plan.

2.3. Pix4D

Pix4D² 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 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. Con la aplicación Pix4DCapture podemos controlar un dron desde nuestro smartphone para que genere un mapa. El dron puede volar de forma autónoma siguiendo algunas de las trayectorias de vuelo que trae por defecto el producto o también puede generar el mapa mientras lo teledirigimos.

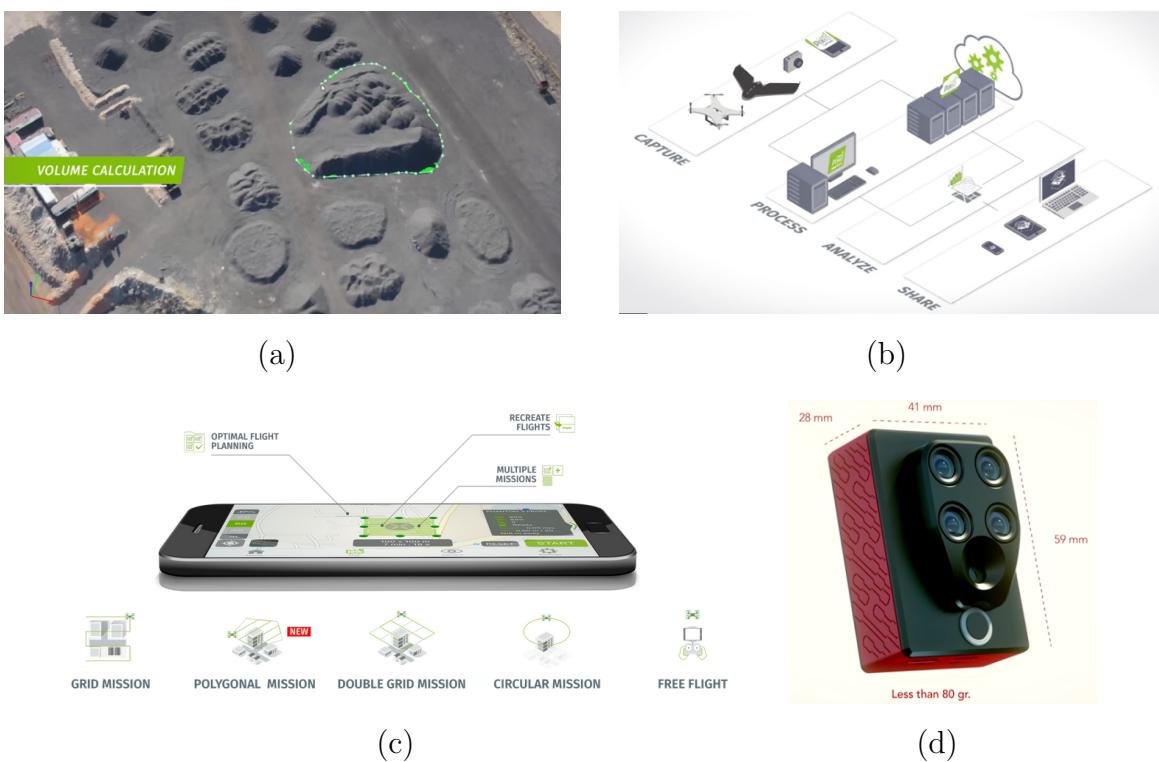
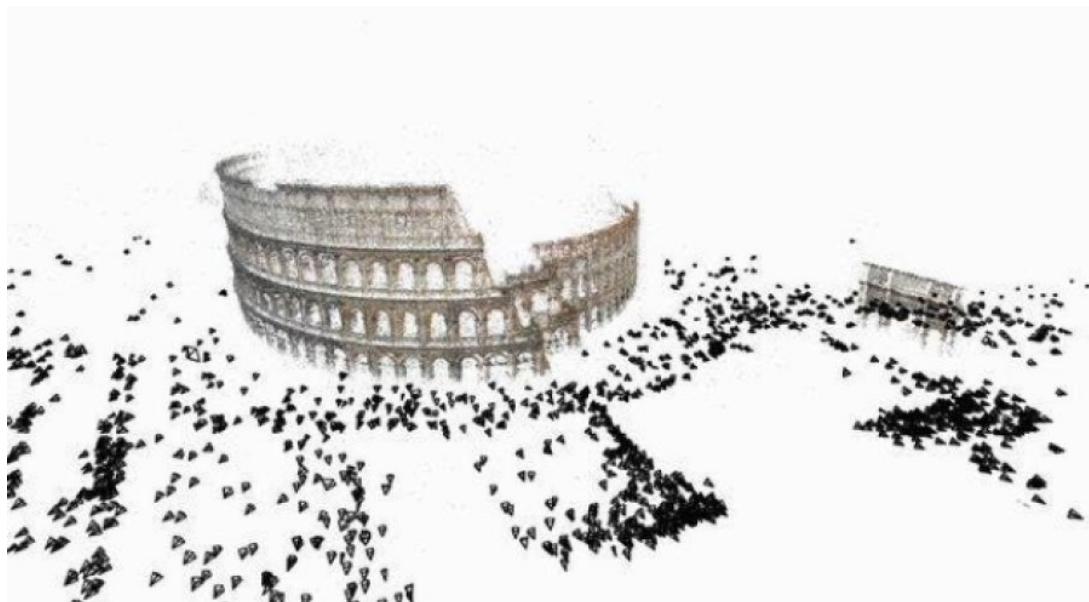


Figura 2.3: Pix4D cálculo de volumen(a). Flujo de datos de Pix4D(b). Trayectorias que pueden seguir los drones con Pix4DCapture(c). Cámara multiespectral Parrot Sequoia (d)

²<https://pix4d.com/>

2.4. 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, La Fontana de Trevi) y es capaz de reconstruir puntos 3D de los monumentos y tambien calcular o estimar la posición de la cámara desde donde se tomaron las fotografías. Proporciona una nueva forma de navegar a traves 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 tambien la localización 3D desde donde se tomaron las fotografías. A diferencia de VisualSLAM, el procesamiento de estas fotografía es offline , 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.



(a)

Figura 2.4: Recreación del Coliseo de Roma.

2.5. 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 equipado con el sensor Structure⁴ alrededor de la habitación y podremos ver como el mapa 3D comienza a generarse en tiempo real. El sensor 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 2.5: El sensor de profundidad Structure para Ipad.

³<https://canvas.io/>

⁴<https://structure.io/>

2.6. Aplicaciones 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 tiene incorporadas varios pares de cámaras estereo, 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 estereo, 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. De esta forma el robot podría sustituir al conocido carrito de la compra Muy útil como carrito de la compra, ya que el robot nos seguirá transportando la compra en su interior, ya no necesitaremos el típico carrito , incluso nos permitiría ir a hacer la compra en bici . Otra utilidad sería en el interior de un hotel podría ser 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
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 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 tuvieran 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 robótizados 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

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

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



Figura 2.6: El Robot Gita siguiendo a su dueño (a) El cinturón con cámaras estéreo (b)
La capacidad de carga del robot Gita (c)

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.⁷
- Aspirador Roomba 966.⁸
- Aspirador LG-Hombot.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.

4. **Drones:** Por último no podemos olvidar los drones, robots voladores equipados con cámara que también pueden obtener mapas de su entorno con VisualSLAM. Existe 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

⁷<http://www.dyson.com>

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

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

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

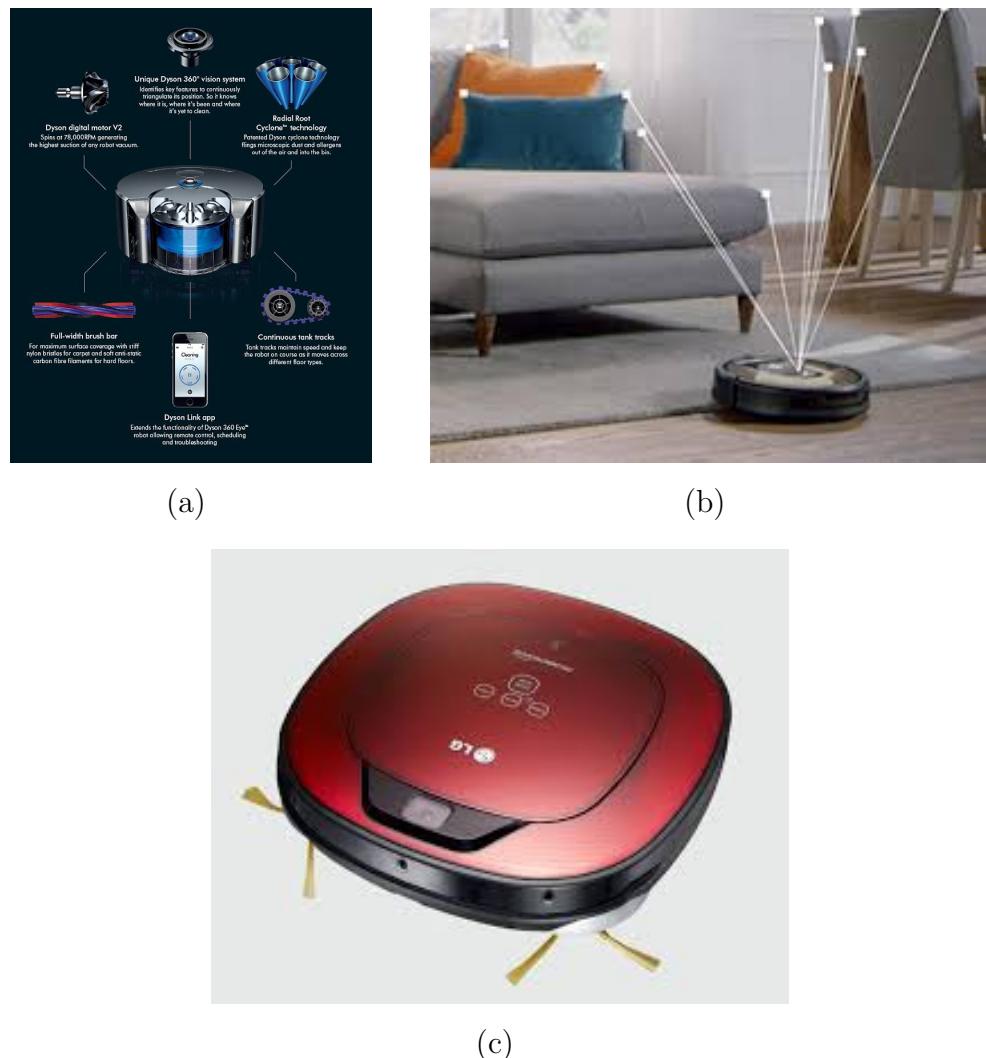


Figura 2.7: Robot Dyson 360 Eye (a) Robot Roomba 966 (b) Robot Hombot de LG (c).

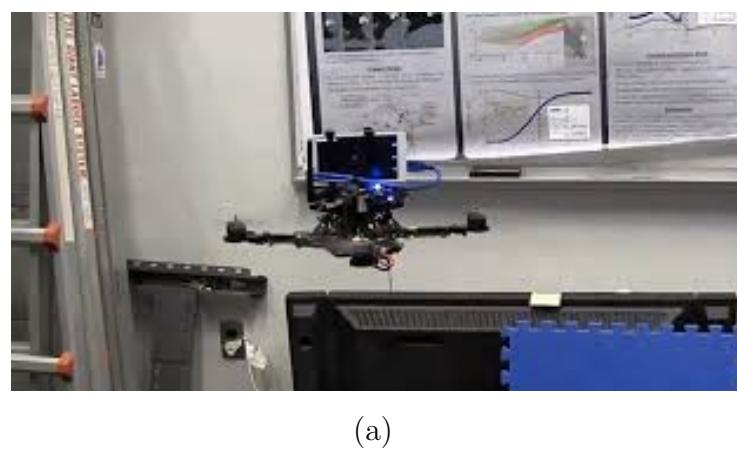


Figura 2.8: Dron equipado con dispositivo compatible con Tango

Capítulo 3

Problemas de Visual Slam

Capítulo 4

Técnicas de Visual SLAM

En esta sección explicaremos varios de los algoritmos conocidos hasta ahora de Visual SLAM , en todos ellos se caracteriza por intentar estimar la posición de la cámara con el menor error posible junto con la generación de un mapeo de la zona.

El conjunto de algoritmos de Visual SLAM se podrían dividir en base al número de regiones que se utiliza de cada frame o imagen recibida para calcular la localización y el mapa . Por un lado estarían el grupo de los algoritmos denso / escaso (*sparse/dense*)¹, por otro lado estarían los Métodos Directos o Métodos basados en características que se caracterizan por el modo en el que se son procesadas las imágenes de entrada (*direct/indirect*)

4.1. Métodos Densos y Métodos Escasos

Los métodos escasos utilizan sólo un pequeño subconjunto de pixeles de ciertas regiones de la imagen, mientras que los Métodos Densos utilizan la mayoría de los pixeles de la imagen captada. Por tanto los mapas generados por los Métodos Densos proporcionan muchos más detalles de la escena al utilizar muchos más puntos, pero también necesitan de una capacidad de computo muy elevada, de hecho la mayoría de los métodos densos requieren la utilización de GPU. Los Métodos Escasos , al tratar menos puntos, obtienen unos mapas con muy pocos detalles , más parecido a una nube de puntos, donde principalmente se representan la trayectoria y las diferentes localizaciones que ha ido ocupando la cámara en el mapa 3D.

¹<https://www.kudan.eu/kudan-news/different-types-visual-slam-systems/>

4.2. Métodos Directos y Métodos Indirectos

Los Métodos Indirectos en Visual SLAM intentan extraer puntos característicos de la imagen y a partir de estos puntos trata de calcular la posición de la cámara y de generar el mapa. Los punto característicos pueden ser desde esquinas y bordes hasta otros descriptores de imagen más sofisticados como SIFT, ORB, FAST . Sin embargo , los Métodos Directos en Visual Slam utilizan directamente los valores de intensidad de los pixeles para construir el mapa y calcular la posición de la cámara. Estos métodos tratan de recuperar la profundidad y estructura del entorno y la posición de la cámara a traves de una optimización del mapa y los parámetros de la cámara al mismo tiempo. Como la extracción de caracteristicas puede llevar mucho tiempo de proceso, los metodos directos pueden llegar a permitir más tiempo para otras computaciones mientras mantienen la misma velocidad de procesamiento de imágenes que los métodos indirectos. Los Métodos indirectos basados en características tienen mayor tolerancia a los cambios de luminosidad ya que no utilizan los valores de intensidad de los pixeles directamente como los métodos Directos.

En el siguiente gráfico se muestran los principales algoritmos de Visual SLAM y que posición ocuparian al clasificarlos entre Métodos Directos e Indirectos y Métodos Densos y Escasos. Cuatro métodos (MonoSLAM, PTAM, ORB-SLAM y SVO) podrían clasificarse dentro del grupo Metodos Indirectos y Escasos. Los métodos (SVO, DSO y LSD-SLAM) se podrían clasificar como métodos Escasos y Directos Los métodos SLAM Y LSD-SLAM podrían clasificarse dentro de la zona de métodos Directos y Densos El método SVO podría clasificarse entre método Directo y método Indirecto Entre Metodo Escaso y Metodos densos nos encontramos con LSD-SLAM Se observa que la zona de Métodos Indirectos y Metodos Densos (cuadrante superior izquierdo), está vacia se entiende que es por la gran necesidad de potencia de CPU que requeririan estos métodos, aunque no se descarta que en el futuro aparezaca algún método en dicha zona.

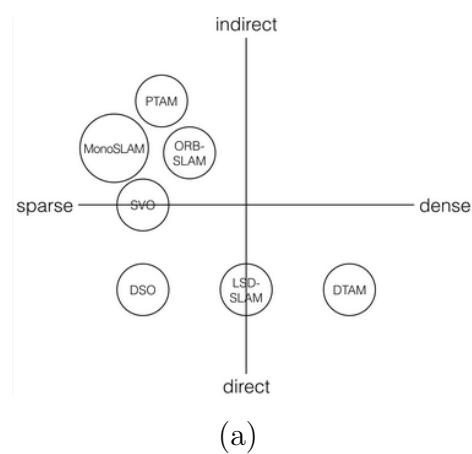


Figura 4.1: Mapa de clasificación de los principales algoritmos de Visual SLAM.

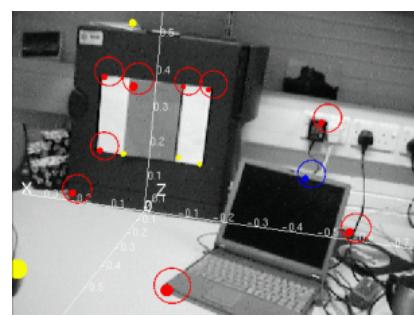
4.3. MonoSLAM

El algoritmo de MonoSLAM (*Monocular SLAM*) 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 3 puntos 3D conocidos que utilizará para calcular la posición de la cámara y la generación de nuevos puntos para el mapa.

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 EKF es apropiado, ya que se realizan iteracciones 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 iteracción se hace una detección de puntos de interes (bordes o esquinas) en la imagen actual de entrada , y obtendremos una serie de puntos que serán candidatos a ser el vector obserbación de los puntos que queremos seguir. Estos candidatos deberán ser filtrados, pues alguno puede ser un falso borde o falsa esquina. Se utilizará una función de divergencia entre parches para determinar si el candidato es aceptable o no. Al utilizar sólo parches de unos pocos pixeles 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 candidatos se utiliza 1-poin 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 inestabilizarse y no llegar nunca a aproximar razonablemente el vector de estado



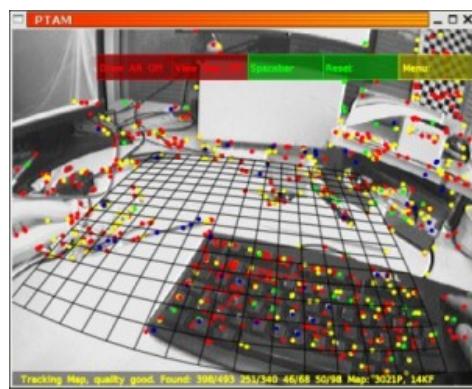
(a)

Figura 4.2: Ejemplo de puntos característicos tomados con MonoSlam.

4.4. PTAM

Parallel Tracking and Mapping. Es un nuevo algoritmo creado en 2007 por George Klein 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 y el segundo para la generación del mapa. Con las imágenes captadas en secuencia se van generando keyFrames o fotogramas clave. A medida que va recibiendo imágenes va generando keyFrames o fotograma clave. Se genera un nuevo KeyFrame a medida que la cámara se va desplazando. Los keyFrames se utilizan para localizarse y como para ir generando el mapa de puntos. Al utilizar 2 threads, el tracking se puede calcular en tiempo real, mientras que el mapping se podría calcular en intervalos ociosos de CPU. 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 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 localizarlos se añaden nuevos puntos al mapa. Mientras no se añadan keyFrames se intentará mejorar el mapa con los keyFrames disponibles.

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

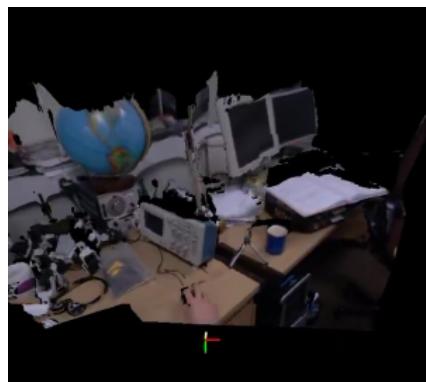


(a)

Figura 4.3: Nube de puntos característicos tomados con PTAM.

4.5. 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. 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 pixel 3D será representado como un cubo, que se proyectará a cada una de las imágenes esclavas. Cuando la recta entre estos pixeles sea cero , el error fotométrico será nulo y entonces se podrá considerar que la proyección es correcta.



(a)

Figura 4.4: Ejemplo de mapa generado con DTAM. Todos los puntos forman parte del mapa.

4.6. 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. Se asemeja a PTAM en que tambien 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, obtiene parches de dimensiones 4x4 alrededor de los pixeles y tarta de hallar el mínimo error fotométrico 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, primero se tratarámos de hallar el error fotométrico mínimo para las zonas de la imagen con profundidad conocida, posteriormente para cada punto 3D visible en la imagen, se escogera el keyframe que mejor se adapta y se optimizará individualmente la transformación afín que minimiza el error fotométrico. Como último paso haremos la minimización del error de reprojeción clásica de los métodos basados en características para corregir los residuos que genera 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.

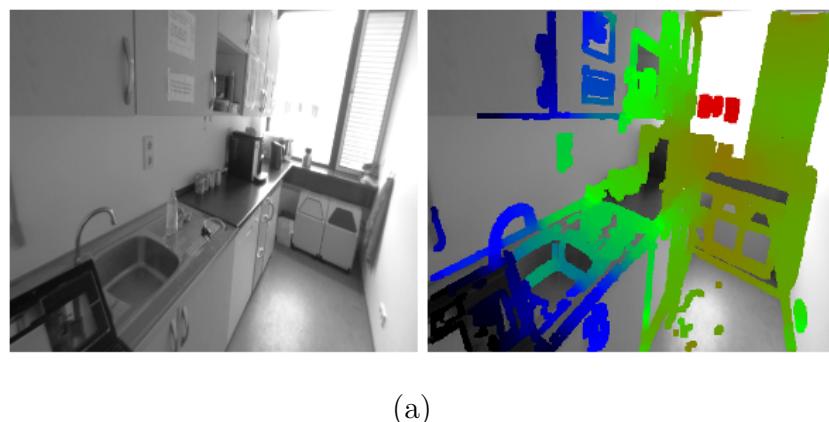
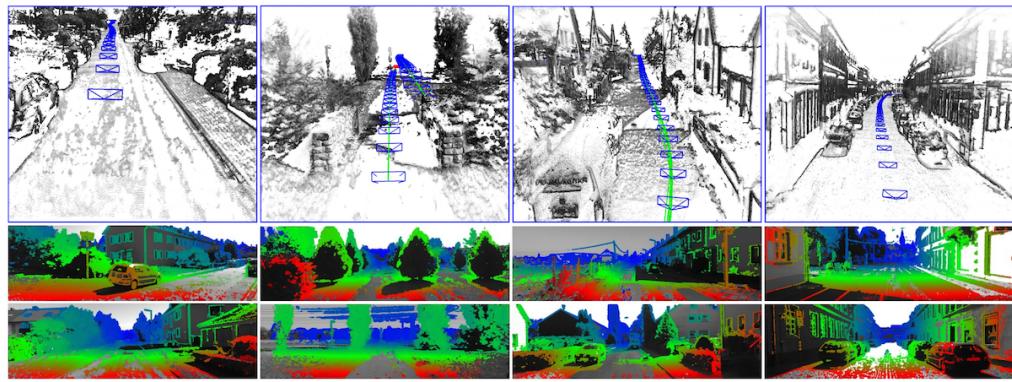


Figura 4.5: Mapa generado con SVO. El color azul indica proximidad y el rojo lejanía

4.7. 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. 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

añadiran 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 sola imagen para generar un mapa inicial de profundidad que ira convergiendo hacia unos valores de profundidad correcta 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 4.6: Mapa generado con LSD-SLAM y cámara estéreo

4.8. ORB-SLAM

Es un algoritmo basado en extracción y emparejamiento de pixeles 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. Puede ser utilizado, con una, dos cámaras e incluso con cámaras de profundidad RGBD. Para cierres de bucle y relocalización utiliza un modelo de bolsa de palabras. 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.

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 Keyframes tengan varios puntos 3D en común. Este grafo permite la eliminación de Keyframes redundantes.

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

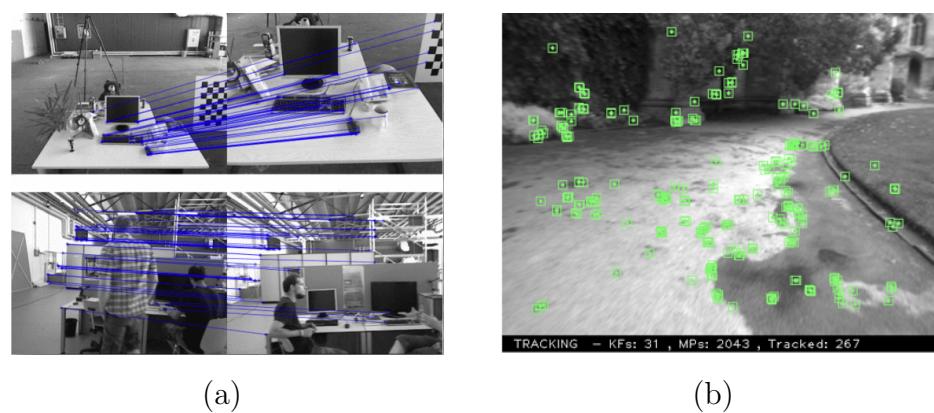


Figura 4.7: Localización de puntos característicos en 2 imágenes con ORB

4.9. DSO

Direct Sparse Model. Está basado en optimizaciones continuas del error fotométrico sobre una ventana de frames recientes. 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 solo 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 cuando la imagen de entrada cambie notablemente con respecto al último Keyframe, esto se medirá con las diferencias de medias al cuadrado entre los pixeles.
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 maximiza un marcador de distancia $d(i,j)$ donde $d(i,j)$ es la distancia Euclídea entre keyframes I_1 y I_j

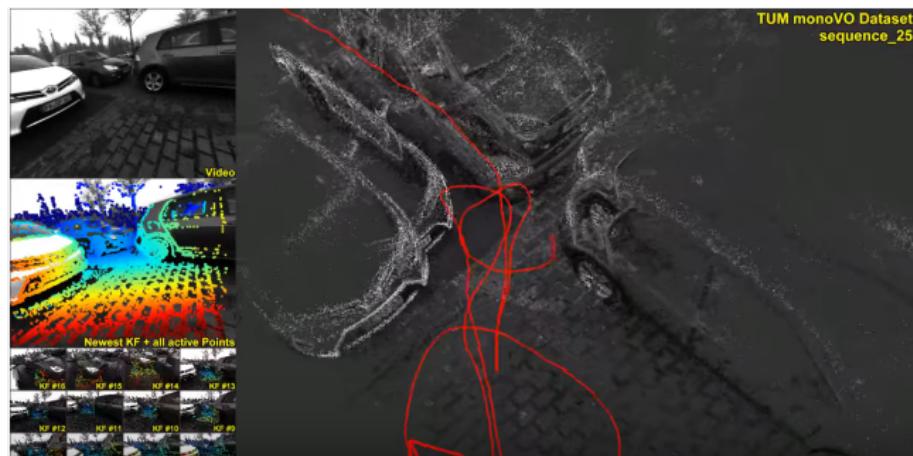
Sobre el tratamiento de los puntos, siempre se tratará de mantener un número fijo de puntos activos repartidos de forma uniforme entre el espacio y los frames activos. En un primer paso, se identifican N_p puntos candidatos en cada nuevo keyframe. Los puntos candidatos no son inmediatamente sumados a 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 pixel con el mayor gradiente siempre y cuando supere un umbral, de lo contrario no se selecciona el pixel 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 intentan mantener una distribución uniforme de puntos por toda la imagen.²

²<http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7898369>



(a)



(b)

Figura 4.8: Mapa generado con DSO (a) Ligero error en la posición al volver al punto de partida (b).

Capítulo 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 necesitasemos 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,MonoSLAM) ya que son fácilmente procesables por los microporcesadores 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.

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 soprase 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 apareciesen nuevos dispositivos periféricos que podrían ser controlados por el smartphone, por ejemplo un nuevo tipo de aspiradora , sin ningun tipo de capacidad para realizar Visual SLAM, solo un par de motores que le permitan avanzar y girar. Si quisiesemos que esta aspiradora comenzase a aspirar de forma autónoma sólo tendríamos que colocar nuestro smartphone de forma 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, quizá el futuro de la conducción autónoma dependa del software de Visual SLAM con el que estén equipados los cada vez más potentes SmartPhones.

Bibliografía

[Arribas, 2016] Victor Arribas. Análisis de algoritmos de visual slam: un entorno integral para su evaluación. *Trabajo Fin de Máster. Universidad Rey Juan Carlos*, 2016.

[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.

[Perdices García, 20017] Eduardo Perdices García. Técnicas para la localización visual robusta de robots en tiempo real con y sin mapas. *Tesis Doctoral. Universidad Rey Juan Carlos*, 20017.