Universidad Zaragoza
1542

Trabajo Fin de Grado

# Emparejamiento repetible en SLAM visual para escenas de endoscopia médica.

# Repetible pairing on visual SLAM in medical endoscopy environment.


Autor

Rodrigo Lozano Puñet

Directores

José María Martinez Montiel

José Lamarca Peiro

ESCUELA DE INGENIERÍA Y ARQUITECTURA


2016

# AGRADECIMIENTOS

Agradezco al apoyo financiero del Ministerio de Economía y Competitividad del Estado Español mediante el proyecto "DPI2017-91104-EXP:SLAM Visual Deformable para Endoscopia".

# RESUMEN

Este Trabajo de Fin de grado se encuentra dentro un proyecto de investigación que tiene como objetivo SLAM (Simultaneous Localization and Mapping) dentro del cuerpo, esto es estimación de un modelo 3D (mapa) de la escena observada y también la localización de la cámara respecto de este modelo. El dato de entrada del algoritmo es el flujo de vídeo de un endoscopio explorando el interior del cuerpo humano. Específicamente, se propone un nuevo método de emparejamiento de puntos de interés en escenas deformables, su implementación y su validación experimental sobre endoscopias in vivo, tanto en secuencias animales como humanas. El método propuesto consigue mejorar el *recall* del emparejamiento. El proyecto está implementado como optimización del sistema DefSLAM *(Deformable Simultaneous Localization and Mapping)* [1].

Para evaluar las prestaciones del nuevo algoritmo se medirán tres características, el error geométrico, el recall y la distribución de puntos de interés. El recall es la ratio entre de los puntos de interés emparejados y los puntos de interés de la plantilla que entran dentro de nuestro campo de visión, que denominaremos in frustum.

Debido al bajo recall en el emparejamiento de puntos se propone un nuevo método de emparejamiento que funciona conjuntamente con el anterior. El nuevo método se denomina emparejamiento por correlación. Consiste añadir al algoritmo de emparejamiento una etapa de búsqueda por correlación normalizada.

Se implementa también una nueva lógica para elegir qué puntos del mapa van a ser buscados. Esta nueva lógica será algo más flexible que la anterior para así obtener un mayor recall.

La medida de estas tres características se llevará a cabo utilizando secuencias de Hamlyn Dataset [2][3][4], de donde se obtienen dos secuencias in vivo de los órganos y corazón de un animal, y una secuencia de una endoscopia humana. En el caso de las secuencias animales se dispone cámaras estéreo que proporcionan un ground truth (solución de referencia) con la cual podemos comparar el mapa obtenido para evaluar su precisión geométrica. Por otra parte, en el caso de la endoscopia humana, la secuencia es monocular, lo cual nos impide comparar el error de en nuestro mapa, así que solo evaluaremos la densidad de puntos de interés y el recall.

Tras la evaluación vemos un aumento considerable en el recall, mientras que el error geométrico permanece prácticamente igual. Esto es un resultado positivo ya que conseguimos una representación más completa del mapa sin ocasionar un error en la estimación. En el caso de la endoscopia observamos que la correlación es vital para el correcto funcionamiento del sistema.

# Índice

# Capítulo 1

# Introducción y objetivos

El sistema DefSLAM (Deformable Simultaneous Localization and Mapping) es el primer sistema capaz de llevar a cabo la localización y reconstrucción de escenas deformables. Este sistema que lleva a cabo una reconstrucción secuencial con el fin de realizar exploraciones escenas no rígidas grabadas. El sistema realiza el mapa utilizando secuencias monoculares.

Una de las principales aplicaciones de este sistema es la reconstrucción de escenas médicas. Para realizar la reconstrucción y la localización relativa de la cámara con respecto a el mapa se abre una la línea de investigación. El principal tipo de escenas a tratar van a ser endoscopias, concretamente colonoscopias y endoscopias gastrointestinales.

El método DefSLAM utiliza descriptores ORB para emparejar imágenes que son dependientes del detector FAST utilizado para su extracción. Se ha comprobado la combinación de este detector y descriptor, no es lo suficientemente robusta para imágenes médicas. Este proyecto tiene como objetivo la mejora del método por el cual se emparejan los puntos del mapa creado con las imágenes de la secuencia y la posterior adaptación del sistema a endoscopias humanas reales.

Para ello, se propone la unión de un método de emparejamiento ya existente con el método de búsqueda activa por correlación [5]. Concretamente, al añadir esta etapa se busca mejorar la repetitividad del emparejamiento consiguiendo mayor robustez en imágenes médicas.

El sistema ha sido evaluado anteriormente [1] con secuencias in vivo de animales sacadas del Hamlyn Dataset [2][3][4], estas secuencias han sido grabadas con cámaras estéreo, lo cuál nos permite la obtención de un Ground Truth con el que comparar el mapa creado por DefSLAM. Gracias a esto se puede evaluar la influencia del la optimización del emparejamiento llevado a cabo de dos maneras distintas. Por un lado vamos a medir la diferencia de los puntos emparejados en términos de recall, y por otro lado, vamos a medir la diferencia del error geométrico entre el mapa estimado por el

sistema y el Ground Truth calculado.

En el caso de la adaptación del sistema a endoscopias humanas reales también se utilizará una endoscopia gastrointestinal proporcionada por el Hamlyn Dataset. una vez tenidas en cuenta las modificaciones oportunas para el funcionamiento de DefSLAM a lo largo de la endoscopia se procederá a la evaluación de su funcionamiento. En este caso no se posee imágenes estéreo de la operación, por lo que la evaluación se va a limitar a un análisis del recall y a una percepción cualitativa del funcionamiento del sistema.

Todo esto será implementado en un ordenador que funciona con un Sistema Operativo Ubuntu 16.04. El programa es un algoritmo en C++.

# Capítulo 2

# Descripción del sistema DefSLAM

El sistema DefSLAM tiene como objetivo la realización del mapa 3D de una escena deformable y la estimación de la posición relativa de la cámara con respeto a dicho mapa utilizando secuencias de imágenes tomadas con cámara monocular.

Para la realización del sistema se usa la unión de dos métodos, NRSfM *(Non-Rigid Structure-from-Motion)* isométrico [6][7] y SfT *(Shape-from-Template)*[8]. Estos métodos van embebidos en dos hilos distintos, Deformation Mapping y Deformation Tracking respectivamente. Los dos hilos tienen funciones distintas pero complementarias para el correcto funcionamiento del sistema.

Deformation Mapping se encarga de estimar la posición en reposo de la superficie, con la cual se genera una plantilla, que también llamaremos template. Esta plantilla es utilizada por el Deformation Tracking para estimar la posición relativa de la cámara respecto a la plantilla y calcular la deformación de la superficie. El sistema entenderá esta plantilla como la estructura que posee el entorno. La unión de estas plantillas a lo largo de la secuencia conformará su mapa 3D.

Dado que el método NRSfM conlleva un tiempo de cómputo mayor, no se ejecuta para cada imagen de la secuencia, sino, solo cuando se cumpla una condición estipulada. La imagen seleccionada para hacer este proceso la denominaremos Keyframe. La condición no solo reúne aspectos de peso en el cómputo, si no también espaciales, si estamos explorando una parte nueva del entorno dónde no se tienen puntos, este método nos permite ampliar nuestra escena haciendo una nueva plantilla. Otro aspecto a valorar si una nueva imagen debe ser un Keyframe es funcional, cuando la cantidad de puntos emparejados pertenecientes a la plantilla es reducido se considerará la creación de una nueva plantilla para así poder estimar mejor la deformación de la misma.

Los dos hilos trabajan en paralelo, por lo que cuando se quiera hacer un Keyframe en la imagen $I_t$, el procesamiento de esta nueva plantilla tardará más que lo que tarda en ejecutarse el hilo Deformation Tracking. Esto conlleva que podrán haber pasado unas pocos fotogramas procesados ya por Deformation Tracking para cuando la plantilla

esté lista para ser implementada en el mapa. Esta plantilla será efectiva en el momento en el que esté lista, no en el que ha sido solicitada.

En la unión de las plantillas se emparejan los puntos de interés que sean los mismos para no crear dos plantillas inconexas. Para realizar este emparejamiento, primero se computa un detector FAST, que detecta puntos de interés. Después, se computa el descriptor ORB. Se intenta emparejar los puntos de interés con los del Keyframe anterior por medio de descriptores ORB. Si la diferencia entre los descriptores ORB es reducida, estos puntos se emparejan uniendo así las dos plantillas. La similitud entre los dos descriptores se llevará a cabo mediante la distancia de Hamming entre las listas de números, si es inferior a cierto umbral, el punto se emparejará.

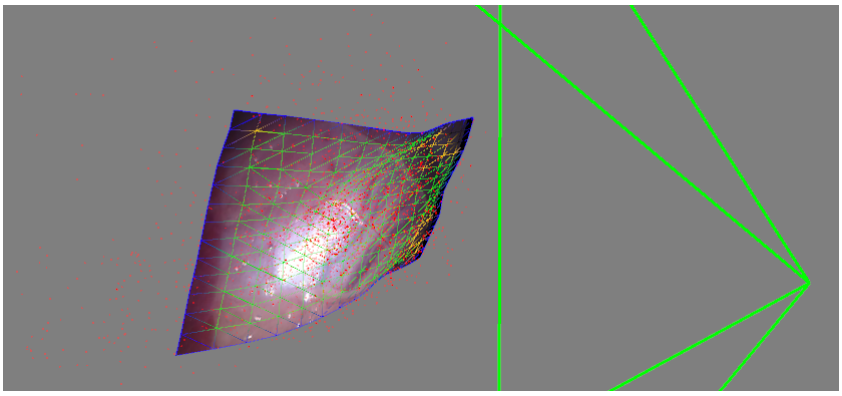

Figura 2.1: Template generado y la posición relativa de la cámara (verde)

La plantilla es una malla triangular que consta de nodos y aristas uniformemente distribuida a lo largo y ancho de la imagen como vemos en la Figura 2.1. Esta textura es la que utilizaremos para determinar nuestro mapa. La posición de cada nodo depende de los puntos de interés cercanos a él. Este funcionamiento remarca la importancia de los puntos de interés para la realización de un mapa correcto. Estos puntos de interés son puntos que presentan entornos característicos que nos permiten hacer un seguimiento a lo largo de los frames del mismo punto con el fin de estimar la posición del mapa en ese punto.

El siguiente paso es estimar la posición de la cámara. Para ello se hace un primer emparejamiento bastante restrictivo con descriptores ORB que nos ayudará a ver la posición actual de la cámara. Con esta posición, en la ecuación 2.1, se determina el modelo de cámara.

$$\pi(\mathbf{X}_j^t, T_{cw}^t) = \begin{pmatrix} f_x \frac{X_j^t}{Z_j^t} + C_x \\ f_y \frac{Y_j^t}{Z_j^t} + C_y \end{pmatrix} \qquad Siendo[X_j^t Y_j^t Z_j^t] = \mathbf{R}_{cw}^t \mathbf{X}_j^t + \mathbf{t}_{cw}^t \tag{2.1}$$

Este modelo de cámara lo usaremos para estimar la posición relativa de los puntos

de interés a buscar. Utilizando esta ecuación podemos llevar a cabo un emparejamiento activo. Este emparejamiento se llevará a cabo utilizando descriptores ORB. Una vez emparejados los puntos se determina la posición de cada uno de los nodos de la plantilla. La posición de estos nodos puede depender de varios puntos a la vez, lo cual nos hace llegar a un compromiso de minimizar los errores que pueda haber en el mapa debido a las variaciones en las estimaciones de la profundidad de los puntos de interés. Esto se hace minimizando el error de reproyección y la energía de deformación 2.2.

$$\arg \min_{T_{cw}^t, L_k^t} \varphi_d(I^t, T_{cw}^t, L_k^t) + \varphi_e(L_k^t, L_k^{t-1}, T_k^t) \tag{2.2}$$

Esta ecuación es divisible en dos términos, por un lado, el término relativo al error de reproyección 2.3, siendo $\mathbf{I}^t$ la imagen en el instante actual t, $\mathbf{T}_{cw}^t$ la posición de la cámara en el instante t, $\pi(\mathbf{X}_j^t, \mathbf{T}_{cw}^t)$ el modelo de cámara calculado en la ecuación (1), y $\mathbf{L}_k^t$ la parte de la plantilla que es vista por la cámara, también llamado mapa local o *in frustum*.

$$\varphi_d(I^t, T_{cw}^t, L_k^t) = \sum_{j \epsilon \mathbf{x}^t} \rho(||\pi(\mathbf{X}_j^t, T_{cw}^t) - \mathbf{x}^t||). \tag{2.3}$$

El segundo término de la ecuación 2.2 es relativo a la energía de deformación, que, a su vez, está dividida en otros tres términos, como vemos en la ecuación 2.4. El primero de estos relacionado con la energía de extensión, el segundo relacionado con la energía de flexión, y por último, el tercer término es un filtro temporal que evita cambios abruptos entre estimaciones consecutivas de las platillas. Cada uno de estos términos va multiplicado por un parámetro ($\lambda$) que da un peso relativo a cada uno de ellos.

$$\varphi_d(I^t, T_{cw}^t, L_k^t) = \lambda_s \varphi_s(L_k^t, T_k^t) + \lambda_b \varphi_b(L_k^t, T_k^t) + \lambda_t \varphi_t(L_k^t, L_k^{t-1}). \tag{2.4}$$

Con esto obtendremos la mejor estimación de la plantilla. Los parámetros $\lambda$ son característicos de cada entorno. Así pues, se tendrá que estimar experimentalmente sus valores para lograr un mapa más preciso.

Como vemos la calidad de la estimación del mapa depende enormemente de la cantidad de puntos emparejados, ya que solo los puntos emparejados consiguen influir en la posición de los nodos cambiando así la estimación de la superficie del sistema.

# Capítulo 3

# Emparejamiento activo por ORB en Deformation Tracking

Una de las etapas clave del Deformation Tracking es determinar, en cada frame, donde han sido observado cada uno de los puntos de la plantilla. Nos referimos a este proceso como emparejamiento (del inglés matching).

En la práctica, no siempre podemos emparejar todos los puntos de la plantilla con puntos de la imagen, este proyecto busca aumentar el número de emparejamientos conseguidos añadiendo una etapa al algoritmo de emparejamiento ya existente. Nos referiremos al nuevo método como búsqueda por correlación y el anterior, como emparejamiento por ORB.

Un índice habitual de calidad del emparejamiento es el *recall*. Se define como la ratio entre el número de puntos que emparejo correctamente (true positives) y los puntos de pertenecientes a la plantilla que podría emparejar idealmente. En general, la plantilla es demasiado grande para entrar dentro del campo vista de la cámara, por lo que idealmente, un emparejador perfecto sólo podría emparejar todos los puntos de la plantilla que caen dentro de la imagen. A este conjunto de puntos nos referiremos como in frustrum. En nuestro caso consideramos que emparejamos todos los puntos correctamente porque tenemos una tasa casi despreciable de falsos positivos. Esto es debido el algoritmo tiene un post-procesado en el que se eliminan los emparejamientos con valores atípicos de innovación. Por ello aproxímanos el recall como la ratio entre puntos emparejados respecto de los puntos in fustrum.

El sistema DefSLAM utiliza inicialmente un emparejamiento activo por descriptores ORB. Este método se basa en la detector de puntos de interés en la imagen por medio de un detector FAST [9] y su posterior emparejamiento con los puntos de la plantilla que tengan un descriptor ORB similar. Este método tiene un recall bajo, entorno al 25-35 %. En cualquier caso, el emparejamiento activo por ORB tiene tres principales ventajas con respecto al emparejamiento por correlación, por lo que primero, se intentarán encontrar

los puntos por ORB, y los puntos que no hayan sido encontrados pasarán a ser buscados por correlación. Las tres principales ventajas de este método son, la invariabilidad ante el giro de la cámara, la leve invariabilidad ante variaciones de escala y tener un peso de cómputo menor.

El detector FAST determina que píxeles de la imagen son un punto de interés, analizando la diferencia de nivel de gris entre un píxel un anillo de píxeles entorno a él. Si esta diferencia supera o no cierto umbral, dará como resultado un 1 o un 0. Si hay una ristra de 1 o 0 que conformar más de la mitad del semicírculo ininterrumpido este punto se considerará que es una esquina, y por ende, un punto de interés para nosotros.

A continuación, en la Figura 3.1, vemos un ejemplo utilizando un anillo de 16 píxeles.

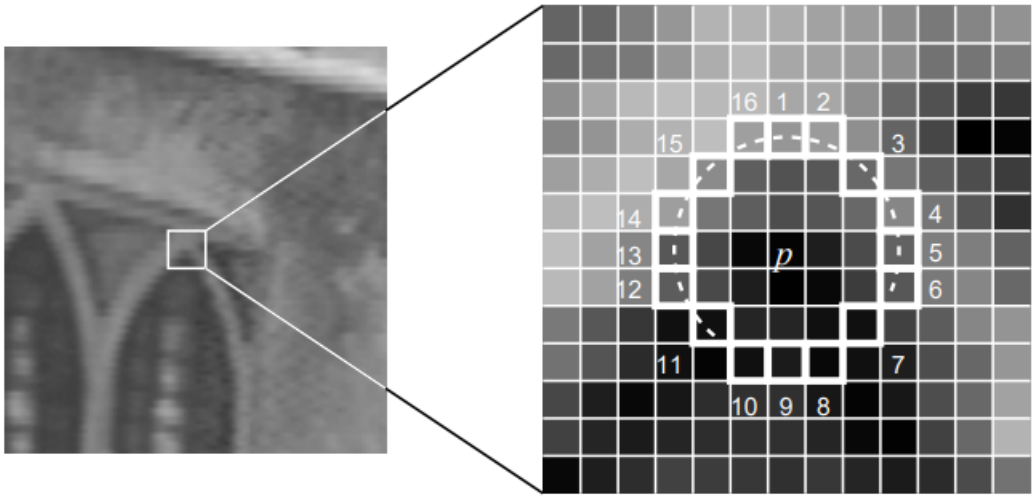


Figura 3.1: Ejemplo FAST con un anillo de 16 píxeles. © 2010 IEEE [9]

Una vez detectado el punto FAST, se calcula un descriptor ORB binario compuesto de una cadena de 256 bits que está asociado al píxel donde ha disparado el FAST en ese fotograma. Cada punto de plantilla tiene también un descriptor ORB, que corresponde al ORB de la primera imagen que lo detectó. En posteriores frames se emparejan los puntos del mapa con los descriptores ORB observados en los nuevos frame.

Este emparejamiento se lleva a cabo con emparejamiento activo. Esto consiste en la predicción de los puntos de la plantilla en la imagen, a partir de la estimación de la posición de la cámara y de la deformación de la plantilla en el frame anterior. Asumiendo que tanto la deformación como el movimiento son pequeños, podemos predecir que el la observación del punto el nuevo frame estará en las proximidades de la predicción. Gracias a esto podemos delimitar una región de búsqueda dentro de la cual suponemos que se encontrará la observación. Utilizando el descriptor ORB del punto del mapa que queremos encontrar calcularemos la distancia de Hamming con

todos los ORBs detectados dentro de la región de búsqueda. Una vez computadas todas las distancias de Hamming seleccionaremos la de menor distancia, esto es el descriptor más similar, siempre que esté por debajo de un umbral.

# Capítulo 4

# Emparejamiento ORB + Correlacion

El principal déficit del emparejamiento por ORB es su baja repetitividad, ya que para emparejar un punto necesita que este haya sido detectado por el detector FAST, y este detector tiene una repetitividad baja, esto es, un punto que en un fotograma ha sido detectado probablemente no sea detectado en el siguiente.

El emparejamiento por correlación se basa en la selección de una subimagen cuadrada en torno al punto de interés para poder reconocer el punto del mapa que queremos emparejar (Fig. 4.1). El tamaño de este cuadrado, $n \times n$ , al que también llamaremos patch o patrón de correlación, $V$, será elegido experimentalmente para adaptarlo a las las secuencias. Este patch se coge alrededor de la observación en el frame anterior del el punto de la plantilla a tratar .

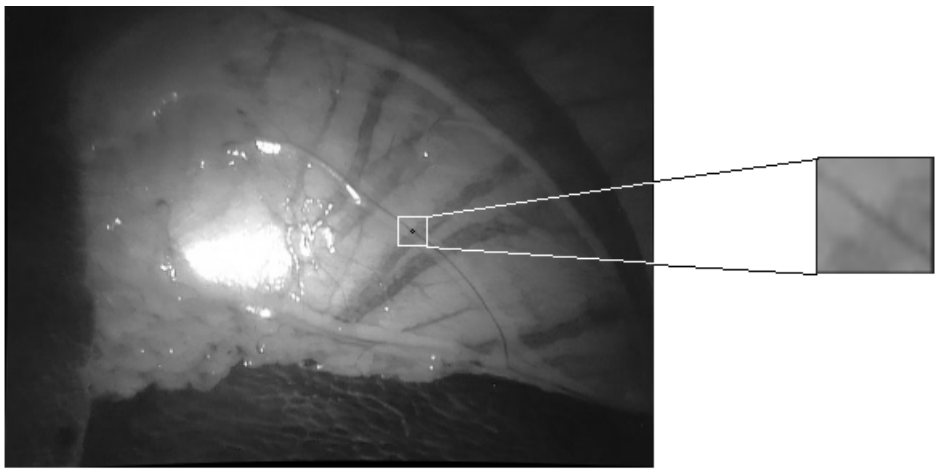

Figura 4.1: Patch patrón seleccionado en torno al punto de interés. El punto en negro es el punto de interés a buscar en el nuevo frame.

El método de correlación normalizada busca encontrar una relación lineal entre el patch patrón y un patch perteneciente al fotograma actual, de forma que si se

consigue encontrar una transformación lineal del patrón haga coincidir los dos patches, estos serán emparejados. La búsqueda de una relación lineal entre los patrones hace al método invariante ante cambios de iluminación. Para tener en cuenta los errores de medida se ha de delimitar un umbral en la similitud entre los patrones. La correlación normalizada (eq. 4.1) nos proporciona una mediada de la calidad del emparejamiento reportando valores en el intervalo $[-1, 1]$.

La búsqueda del punto se lleva a cabo dentro de una región de $(R)$, que estará delimitada a una ventana de $m \times m$ píxeles. Siendo $m$ mayor que $n$. Tanto $m$ como $n$ serán números impares para que exista un píxel en el centro del cuadrado. El resultado es una matriz de $(m - (\frac{n+1}{2}), m - (\frac{n+1}{2}))$.

$$\rho = \frac{\sum_{i=1,j=1}^{i,j \leqslant n}(V_{i,j} - \bar{V})(R_{i+x,j+y} - \bar{R})}{\sqrt{\sum_{i=1,j=1}^{i,j \leqslant n}(V_{i,j} - \bar{V})^2}\sqrt{\sum_{i=1,j=1}^{i,j \leqslant n}(R_{i,j} - \bar{R})^2}}. \tag{4.1}$$

Al ser una comparación pixel a pixel de los patches, la correlación no funciona bien si hay rotación entre los pacthes, también en caso de acercarse o alejarse del mapa, se incurre en una variación de escala, lo cual también afecta significativamente al rendimiento de la correlación. Otro inconveniente es el coste computacional del método al tener que comparar todos los píxeles que componen el patrón de correlación para cada pixel de la ventan de búsqueda, adicionalmente al estar normalizada, también hay que calcular la media y la desviación típica para cada comparación.

A continuación en la Figura 4.2. se muestra un ejemplo numérico utilizando $n = 3$, y $m = 7$, por lo que la matriz de similitud será de $5 \times 5$ (Figura 4.3).

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| 36 | 37 | 38 | 39 | 40 | 41 | 42 |
| 43 | 44 | 45 | 46 | 47 | 48 | 49 |

| | | |
|---|---|---|
| 16 | 19 | 22 |
| 23 | 26 | 28 |
| 30 | 31 | 35 |

Figura 4.2: Izq: Región de búsqueda $R$, en verde los píxeles que conforman la matriz solución Drch: Patrón de correlación $V$

| | | | | |
|---|---|---|---|---|
| 0,311 | 0,328 | 0,347 | 0,368 | 0,392 |
| 0,483 | 0,522 | 0,565 | 0,615 | 0,670 |
| 0,857 | 0,913 | 0,952 | 0,966 | 0,952 |
| 0,794 | 0,730 | 0,670 | 0,615 | 0,565 |
| 0,449 | 0,419 | 0,392 | 0,368 | 0,347 |

Figura 4.3: Matriz de similitud

Aplicando la ecuación 4.1 obtenemos la matriz de similitud. Siempre nos quedaremos con el de mayor similitud.

## 4.1. Búsqueda de emparejamientos por correlación

La implementación se llevará a cabo proyectando cada uno de los puntos de la plantilla a coordenadas píxel de la cámara utilizando la ecuación (2.1). Si al proyectar el punto, este es in frustrum, y además, está alejado una distancia suficiente del borde de la imagen como para poder albergar la ventana de correlación entera, lo consideraremos como punto a encontrar en nuestro nuevo frame.

Al estar trabajando con imágenes de vídeo, las deformaciones y movimientos que pueda haber entre un fotograma y su siguiente son pequeños. Por lo que, para considerar la correlación como método válido, vamos a utilizar en la comparación fotogramas consecutivos. Gracias a esto podemos asumir que la nueva ubicación de un mismo punto de interés no estará alejada de la ubicación del punto de interés en el anterior frame. Así pues, podemos delimitar una región para la búsqueda de cada punto de interés en el nuevo frame.

Una vez seleccionado el patrón de correlación y la región de búsqueda, con la ec. (4.1) obtendremos los porcentajes de similitud. Hay que decidir si es lo suficientemente parecido el como poder emparejarlos. Se establece un umbral mínimo de correlación del 95 %.

Aplicando este nuevo método conseguiremos que algunos puntos que no se han emparejado utilizando un emparejamiento por ORB, sean emparejados, subiendo así el recall.

## 4.2. Tamaño de la región de búsqueda y del patrón de correlación

Como vemos en la realización del método hay un gran costo computacional debido a la cantidad de píxeles a comparar. Este costo computacional crece cuadráticamente con el tamaño de ventana $n$ y la región de búsqueda $m$.

Por otro lado, también hay que considerar que al hacer un tamaño de ventana $n$ mayor, compararemos mayor parte del entorno permitiéndonos así ser más precisos a la hora de calcular la similitud en cada píxel, dando un menos número de falsos positivos, pero por el contrario, ser más vulnerable a deformaciones lejanas al punto que queremos encontrar y esto hacer que no se empareje, dando lugar a un aumento de falsos negativos.

Dadas estás razones se ha de llegar a un compromiso según las imágenes a tratar. En nuestro caso, utilizaremos $n = 19$ y $m = 27$.

# Capítulo 5

# Selección de puntos del mapa a buscar.

Para elegir los puntos de interés a buscar, el sistema usado establece una condición, se buscarán los puntos del template que hayan sido visto en el frame anterior.

Esta condición es bastante restrictiva, ya que limita enormemente la cantidad de puntos que se buscan durante los frames que hay entre dos keyframes. Al ser tan restrictiva, hay implementado un método que consiste en la proyección de puntos pertenecientes al mapa cercanos al template actual $T_k^t$ y su posterior búsqueda por ORB con los puntos no emparejados de la plantilla.

Si hay emparejamiento, ese punto se considerará el mismo de ahí en adelante. De esta manera se garantiza la supervivencia del sistema ya que sin este método el sistema dejaría de encontrar los puntos necesarios para su correcto funcionamiento.

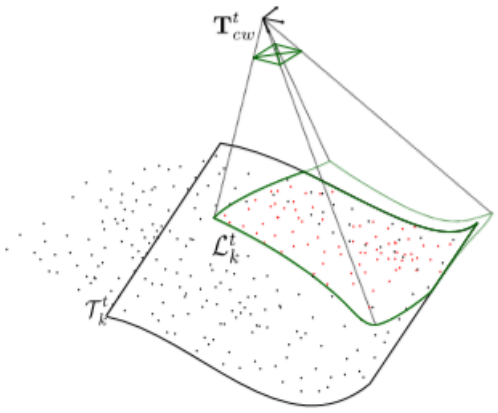


Figura 5.1: Puntos negros, pertenecientes al mapa. Puntos rojos, emparejados. $T_k^t$ plantilla en el instante t. $L_k$, puntos in frustum pertenecientes a la plantilla, $T_{cw}^t$, posición de la cámara. [1]

En este proyecto se propone y se evalúa un nuevo enfoque para decidir que puntos se buscarán. Siguiendo la lógica del sistema, al depender el mapa de la plantilla, y

ésta depender únicamente de los puntos que se encuentran en ella, se decide buscar únicamente los puntos de la plantilla. De esta manera, no se ha de proyectar todos los puntos del mapa, sino solo aquellos que formen parte de la plantilla.

Esto por sí solo haría el que número de emparejamientos disminuyese, lo cual se compensa con la segunda decisión tomada, se buscarán todos los puntos de la plantilla en cada uno de los frames que la utilicen, esto significa, dejaremos de poner la condición que, para buscar un punto, este debe haber sido visto en la anterior imagen.

# Capítulo 6

# Adaptación a endoscopias humanas

Uno de los ámbitos de aplicación del DefSLAM son escenas médicas. Para ello se ha de comprobar una posible necesidad de adaptación del sistema a estos entornos.

En evaluaciones anteriores el sistema ha sido adaptado para trabajar con secuencias in vivo del Hamlyn Dataset[2][3][4]. Principalmente con dos secuencias, la primera se trata de una cámara explorando los órganos de un cerdo y la segunda, una cámara fija enfocada al corazón latiente de un cerdo.

Si observamos alguna endoscopia podemos apreciar diferencias con las secuencias tratadas con anterioridad. La primera, la velocidad con la que se mueve la cámara es mucho mayor, y los cambios de velocidad son mucho más bruscos. Esto constata la necesidad de realizar una estimación de la cámara con un modelo basado en la velocidad que tiene la cámara en los instantes anteriores al frame actual y no por su aceleración.

Otra característica de las endoscopias es su cambio brusco de la iluminación del entorno ya que, si la linterna incorporada en el endoscopio queda tapada por alguna parte del entorno, la imagen sea prácticamente negra, o excesivamente iluminada. Esto afectará en gran medida al método por correlación, ya que si la varianza del entorno del punto de interés es muy baja el punto puede ser emparejado por error.

La última característica notable es la baja calidad de las imágenes que forman la secuencia. Esto también influirá en el método por correlación al bajar la varianza del entorno.

En la figura 6.1 se muestran dos imágenes ejemplo de la secuencia.

Debido a esto se proponen tres cambios para asegurar el correcto funcionamiento del sistema. Dado al aumento de velocidad de la cámara, se aumenta la región de búsqueda en la que intuimos que puede estar el punto, ya que nuestra estimación de la posición de la cámara y, por ende, la estimación de la localización del punto en coordenadas relativas a la cámara puede tener un mayor error.

Por otra parte, al ser un entorno más abrupto y con mayores deformaciones, el parámetro $\lambda_e$ de la sección 2.1 que modela la energía de deformación será disminuido,

dándole así más libertad al mapa para deformarse.

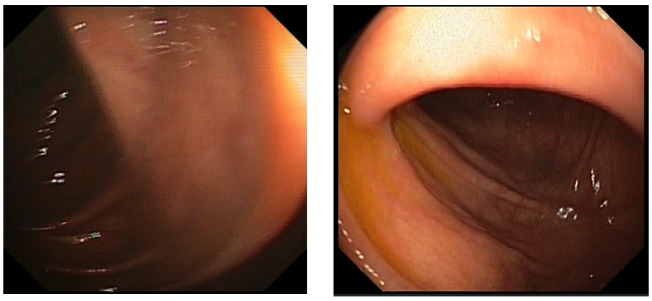

Figura 6.1: Imágenes de una secuencia endoscópica. [4]

Por último, al apreciarse varianzas menores en algunas regiones en algunos momentos de las deformaciones y una menor calidad en las imágenes se aumentará el porcentaje de similitud requerido para la aceptación de un emparejamiento por correlación de un 95 % a un 98 %.

# Capítulo 7

# Ground Truth

El Ground Truth es la estimación del mapa a partir de secuencias grabadas con cámaras estero. Esta estimación es usada como mapa de referencia para calcular el error geométrico que obtenemos utilizando secuencias monoculares. La estimación del Ground Truth ya ha sido implementada en [10].

Esta estimación se lleva a cabo proyectando todos los puntos pertenecientes a la plantilla in frustum en la imagen tomada con la cámara izquierda. A partir de la posición del punto en la cámara izquierda se delimitará una región de búsqueda en la imagen obtenida por la cámara derecha.

La búsqueda de este punto en la cámara derecha se hará por correlación. Una vez encontrado el punto en la cámara de la derecha se triangulará sabiendo la distancia entre los focos de ambas cámaras y así calcular la profundidad de dicho punto como vemos en la Figura 8.

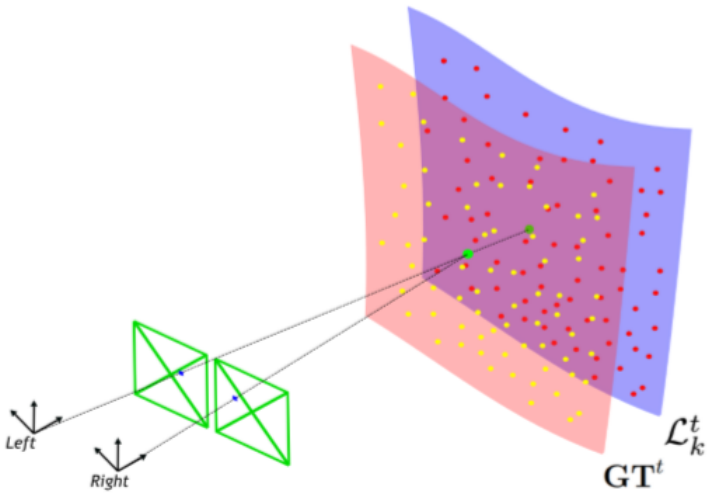


Figura 7.1: Estimación de la profundidad de un punto en el Ground Truth. En salmón vemos la plantilla creada por el Ground Truth que consta de los puntos amarillos. En azul el mapa local creado utilizando el sistema con solo una cámara que consta de los puntos rojos. En verde se destaca un punto de interés utilizado. Javier Morlana [10]

Una vez computada la profundidad para todos los puntos de la plantilla in frustum se obtiene una buena estimación de la escena.

# Capítulo 8

# Evaluación del método de búsqueda por correlación.

Una vez implementado el nuevo método ha sido evaluado. El sistema ha sido evaluado con anterioridad en [1] para compararlo con el funcionamiento del SLAM rígido en escenas deformables. Las secuencias usadas fueros obtenidas del Hamlyn Dataset[2][3][4], donde hay secuencias in vivo de animales. Las secuencias seleccionadas han sido grabadas con cámaras estero, lo cual hace posible conocer el Ground Truth de la escena utilizando el método implementado por [10].

Concretamente se van a utilizar dos secuencias distintas. La primera de ellas tiene carácter exploratorio, donde irán apareciendo nuevas regiones del mapa y se usa una herramienta que llega a ocupar gran parte de la imagen, esta secuencia se llama Organs. La segunda es una imagen estática del corazón de una vaca, también se puede apreciar la aparición de una herramienta, pero más alejada y con menor influencia en la escena que en la anterior secuencia, esta secuencia se llama Heart.

El sistema utiliza concurrencia, por lo que, dependiendo de la velocidad de procesamiento de cada uno de los hilos, el resultado puede variar ligeramente. Debido a esto, para evaluarlo con más precisión, las gráficas y resultados conseguidos serán obtenidos tras hacer la mediana en cada frame de 5 ejecuciones independientes del programa.

Para evaluar la diferencia causada por el nuevo método se tendrán en cuenta dos características esenciales del sistem. El error del mapa y el recall.

## 8.1. Evaluación de los emparejamientos con los puntos del mapa

La principal ventaja de la implementación de la búsqueda por correlación es el aumento en el recall.

La figura 8.1 y la figura 8.2 muestran los parámetros a evaluar. En rojo observamos la cantidad de puntos perteneciente a la plantilla que somos capaces de ver con la cámara (in frustum), en verde la cantidad de puntos emparejados y, en azul los puntos emparejados por ORB.

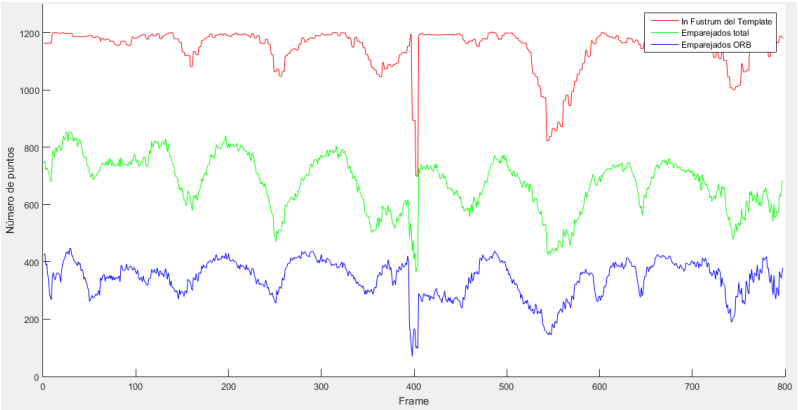


Figura 8.1: Puntos en la Secuencia Organs

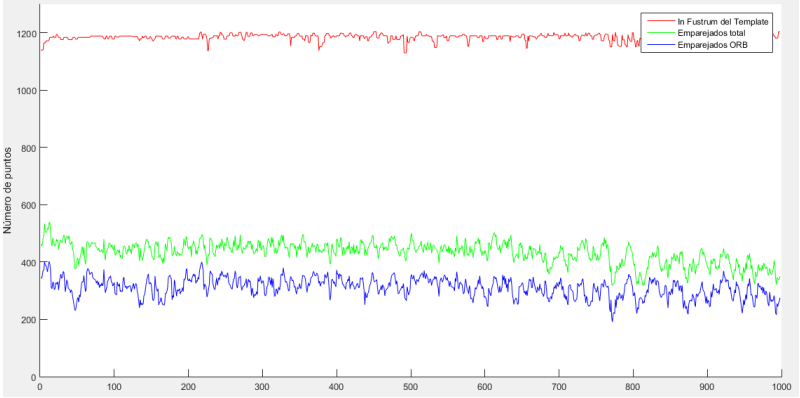


Figura 8.2: Puntos en la Secuencia Heart

La diferencia entre la línea verde y la azul muestra la cantidad de puntos emparejados por correlación, como vemos es significativa.

Utilizando solo ORB obtenemos un recall de 30.76 % y añadiendo correlación, aumentamos hasta un 58.89 % en la secuencia Organs. En la secuencia Heart, obtenemos un 26.3 % solo con ORB y un 36.6 % si añadimos la correlación. El aumento es notable en ambas secuencias.

Cabe destacar que ahora los puntos emparejados pueden describir mejor el mapa, ya que este nuevo método nos permite emparejar y estimar nuevas zonas dónde antes no podíamos comparar el mapa. Como vemos en la Figura 11. Los puntos en azul, son los puntos del template no emparejados, los puntos rojos son los emparejados, las líneas cían muestran que el punto al cual están unidas ha sido emparejado por correlación y

con qué punto ha sido emparejado, y las líneas en verde que han sido emparejados por ORB y con el punto con el que han sido emparejados.

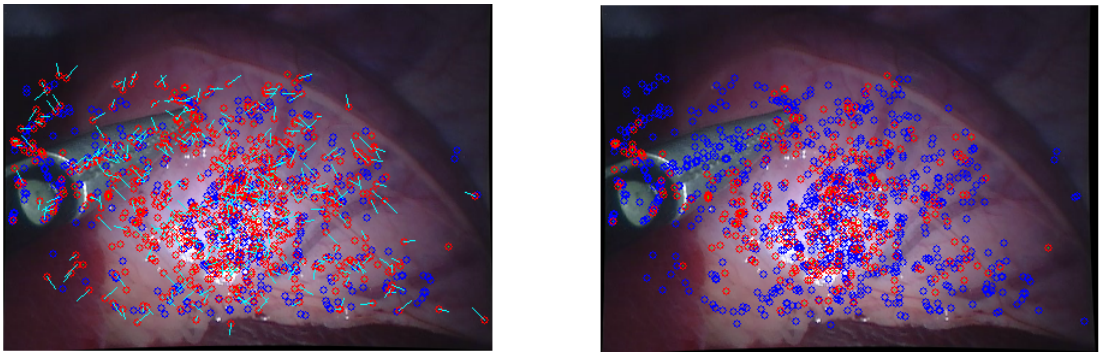

Figura 8.3: Izq: Puntos emparejados usando correlación Drch: Puntos emparejados solo usando ORB

Estas dos imágenes son en el mismo frame de la secuencia Organs, pero en la de la izquierda el sistema usa el método por correlación y en la derecha no. Como vemos, cuando se utiliza la correlación, emparejamos puntos distribuidos por toda la imagen, lo que nos permitirá llevar a cabo una deformación de la plantilla más real y teniendo en cuenta toda la superficie de la misma. Mientras que, cuando se utiliza solo ORB los puntos emparejados se centran en el regiones pequeñas y poco distribuidas de la imagen, esto hace que haya regiones de la plantilla en las que no se estime la deformación, y, por ende, que esa parte el mapa solo sea estimable por en la generación de la plantilla en el hilo Deformation Mapping, toda deformación que sufra la escena en esas regiones antes permanecía desconocida, mientras que ahora somos capaces de estimarla.

## 8.2. Evaluación del error geométrico.

Hemos constatado el aumento en los emparejamientos, pero si esos emparejamientos nos llevan a una solución errónea del mapa, no nos ayudan. Por esta razón, para evaluar el correcto funcionamiento del sistema, vamos a calcular el error del mapa. Esto se puede hacer debido a que las secuencias elegidas han sido grabadas con imágenes estero. Para nuestro sistema hemos cogido solo una de las dos cámaras, simulando que la secuencia es monocular, como ocurre en las endoscopias. Para calcular el mapa real se utilizarán las dos cámaras, de manera que, los puntos pertenecientes a la plantilla actual serán buscados en una región cercana en la imagen tomada por la otra cámara. De esta manera se obtendrán la posición del mismo punto visto desde dos sitios distintos. Conociendo la distancia entre el foco de las dos cámaras y triangulando, podemos calcular la profundidad a la que se encuentra el punto de interés. De esta manera y computando esto para todos los puntos de interés pertenecientes a la plantilla se obtiene

el Ground Truth.

El cálculo del error se lleva a cabo mediante la raíz cuadrada de la media del error (RMSE) (8.1), donde $\hat{\mathbf{X}}_i$ es la estimación de la localización del punto computada por el sistema DefSLAM, $\mathbf{X}_i$ es la posición del mismo punto computada utilizando las dos cámaras estero, N el número de puntos que constan dentro de la plantilla, y, $\lambda$ es un parámetro que nos permite compensar la deriva de escala.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{X}}_i - \lambda \mathbf{X}_i)}. \tag{8.1}$$

Para poder calcular el RMSE primero hay que estimar la deriva de escala, y esto se hará minimizando la ecuación (8.2). Con este método se consigue alinear los dos planos, ya que en un sistema monocular se incurre en un error de deriva de escala. De esta manera se considerarán comparables los dos mapas para utilizar (8).

$$median_i \arg\min_\lambda ||\hat{\mathbf{X}}_i - \lambda \mathbf{X}_i||^2 \tag{8.2}$$

Este cálculo se llevará a cabo cogiendo todos los puntos pertenecientes al template, hayan sido emparejados o no por alguno de los métodos.

Las figuras 6 y 7 muestran la comparación del error en la secuencia Organs y la secuencia Heart respectivamente medido en milímetros a lo largo de la secuencia. En azul se muestra el error obtenido utilizando solo ORB, en naranja se pone a la vista el error obtenido por el sistema utilizando ORB y correlación.

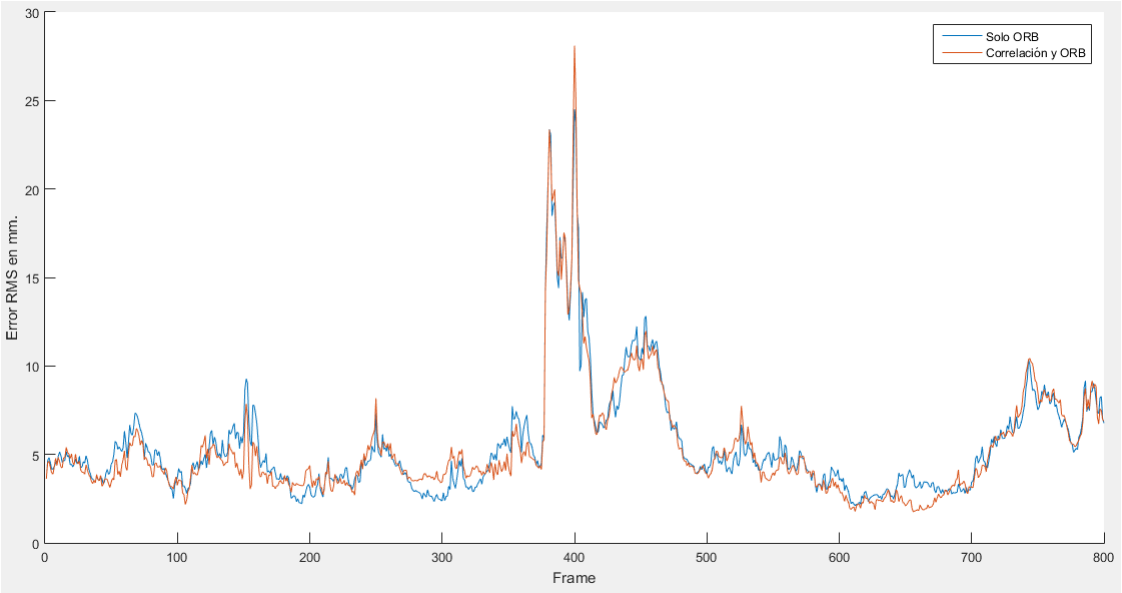


Figura 8.4: Error medido en milímetros en la secuencia Organs

Analizando la secuencia Organs (Figura 8.4) vemos que en los primeros 400

frames, la secuencia tiene un carácter exploratorio, donde el sistema no sufre grandes deformaciones, hasta que en el frame 380 entra una herramienta que llega a taponar gran parte de la visión de la cámara, esto genera grandes perturbaciones en el mapa estimado, con lo que conlleva un aumento del error. En el momento que la herramienta se separa de la cámara, vemos que el error vuelve a bajar. A partir de ese momento la herramienta incide sobre el tejido deformándolo. Esto conlleva aumento del error en el sistema hasta que se consigue adaptar a esta deformación, lo cual tarda en adaptarse 50 frames.

Para evaluar cuantitativamente el error vamos a calcular la media y la mediana del error a lo largo de la secuencia. Cuando el sistema empareja puntos utilizando solo descriptores ORB, se obtiene una media de 5.460 milímetros y una mediana de 4.638 milímetros. En el caso de que el sistema añada el emparejamiento por correlación a su funcionamiento, se obtiene una media de 5.312 y una mediana de 4.359. Como vemos se reduce el error levemente, un 3 % en l media y un 6 % en la mediana.

La evaluación de la segunda secuencia tratada se lleva a cabo en la figura 8.5. En esta secuencia la cámara permanece estática y es el corazón de un cerdo el que genera deformaciones constantes a lo largo de la secuencia. Añadido a esto, entorno al frame 800 aparece una herramienta que ocupa parcialmente la imagen, en esta ocasión, la influencia de la aparición de la herramienta en la escena no influye tanto como en la secuencia Organs.

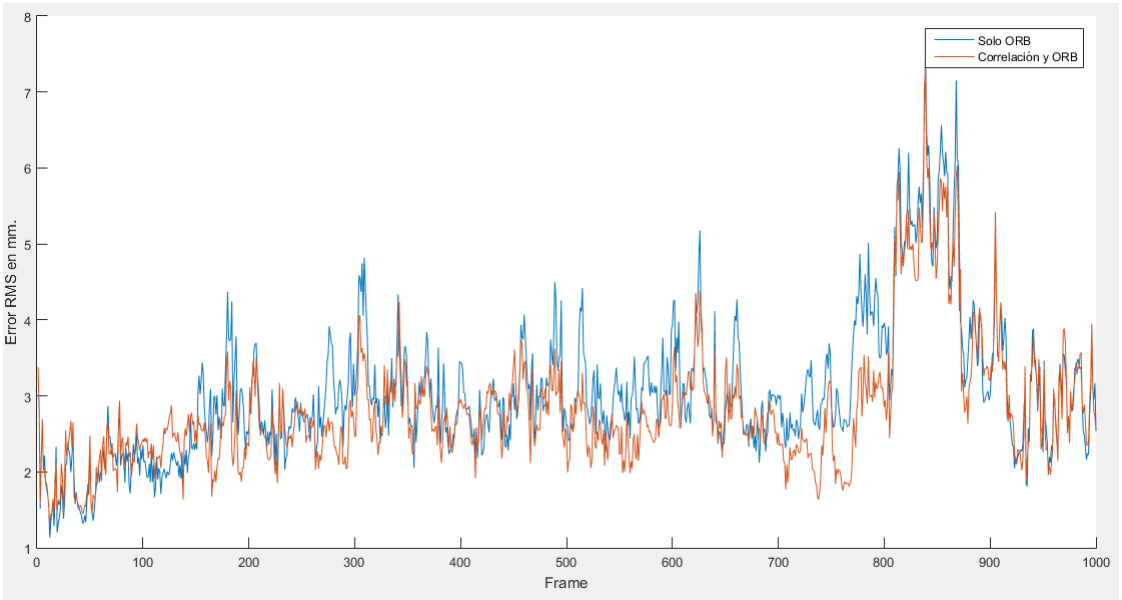


Figura 8.5: Error medido en milímetros en la secuencia Heart

Según vemos en la gráfica, cuando el sistema solo hace emparejamiento por ORB, el error tiene más picos en la estimación del mapa, cosa que cuando se usa correlación no pasa y la influencia de la herramienta es algo menor cuando se usa la correlación.

Pasando a un análisis cuantitativo del error, si el sistema no usa el nuevo método de emparejamiento, se obtiene una media de 3.043 milímetros y una mediana de 2.870 milímetros. Cuando el sistema incluye la correlación, la media del error es 2.810 milímetros y la mediana 2.6428 milímetros. Como vemos, esto supone una reducción del error de un 7.6 % de la media y un 7.9 % en la mediana.

### 8.2.1. Evaluación cualitativa en endoscopias humanas.

En las endoscopias humanas no existe la posibilidad de obtener imágenes estero. Debido a esto no se puede evaluar cuantitativamente el error en el mapa.

Como vemos en la Figura 8.6, donde se nos muestra la cantidad de puntos que hay en el template en cada frame (rojo), la cantidad de puntos emparejados por ORB (azul). El sistema en esta ocasión ha sido formado solo a partir de puntos ORB.

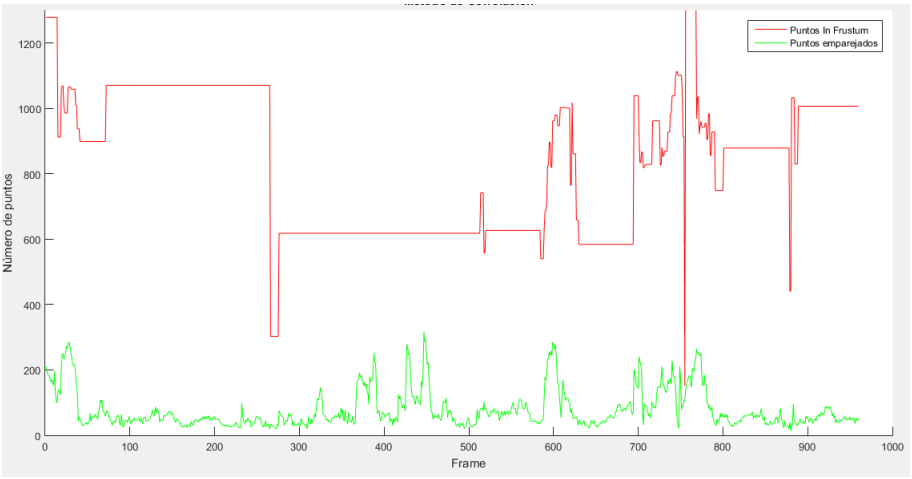


Figura 8.6: Puntos emparejados en endoscopia utilizando solo ORB

El análisis de esta secuencia no solo se ha de llevar a cabo en términos de recall, ya que hay un factor importante a tener en cuenta. Para la correcta formación del mapa y la cámara se considera necesario al menos el emparejamiento de 20 puntos en cada frame de la secuencia. Como vemos en la gráfica, el sistema está por debajo de ese umbral en repetidas ocasiones. Esto genera que el sistema se tenga que reiniciar y borrar todo el mapa generado con anterioridad. El sistema llega a tener que reiniciarse una media de 27 veces durante la secuencia.

Añadido a este problema, debido al bajo número de emparejamientos que obtiene, no es capaz de discernir con precisión cuando es necesario un Keyframe. Esto provoca que haya muchas menos plantillas de las necesarias provocando a su vez menor emparejamiento y errores en la realización del mapa.

Esto hace que la elaboración del mapa intracorpóreo del cuerpo humano sea impracticable utilizando solo este método.

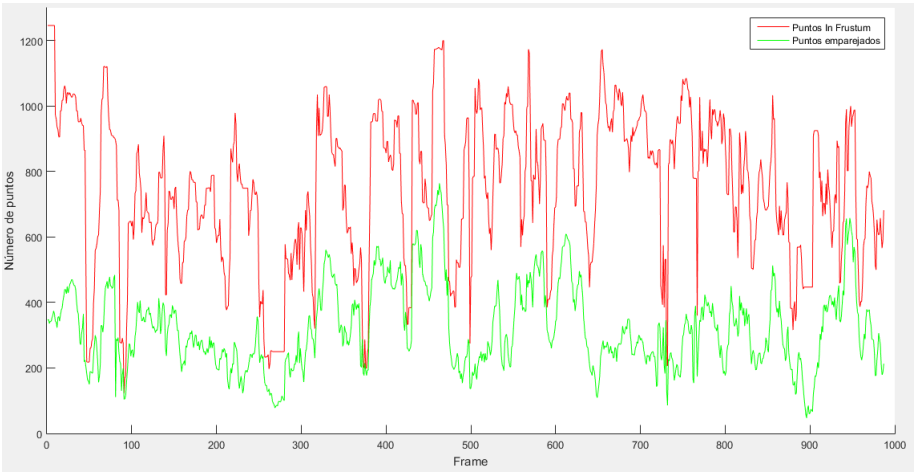


Figura 8.7: Puntos emparejados en endoscopia con los dos métodos

En la Figura 8.7, se puede ver en rojo los puntos pertenecientes al template para cada frame y en azul los puntos emparejados. En esta figura el resultado cuando el sistema utiliza ambos métodos.

Como vemos en la gráfica, el sistema ahora no se pierde, es más, teniendo en cuenta la dificultad de la secuencia, ya que la cámara llega a estar casi por completo taponada en una pared, el resultado es prometedor. Se obtiene un recall de 44.26 %, suficiente para llevar a cabo el mapa y la estimación de la cámara respecto al mismo

# Capítulo 9

# Conclusiones y trabajos futuros

Al inicio se partía de una base de un sistema que obtenía una repetitividad reducida y los puntos que se conseguían emparejar estaban concentrados en regiones pequeñas. Esto hacía inviable el desarrollo y creación de mapas en entornos endoscópicos.

Con la implementación del nuevo método de búsqueda de puntos de interés, se ha logrado duplicar la repetitividad de dichos puntos, y, además, estos nuevos emparejamientos permiten la estimación de la deformación y el mapa en regiones donde el anterior método no era capaz de conseguir resultados.

Viendo los resultados expuestos por el capítulo de la evalución , se concluye que la búsqueda por correlación es beneficiosa para el sistema. No solo esto, sino que, como se aprecia en la evaluación de endoscopias, la unión de los dos métodos es esencial para poder evaluar imágenes médicas endoscópicas.

El sistema se propone trabajar en el Hospital Clínico Lozano Blesa de Zaragoza, para ello, en la sala dónde se realicen las endoscopias se colocará un ordenador capaz de guardar las secuencias. Obtenidos los permisos éticos y legales necesarios, se podrán visualizar endoscopias grabadas reales y realizadas por el mismo endoscopio que se plantea utilizar para finalmente.

Con estas secuencias se podrá decidir si conviene adaptar algún parámetro o incluso, llegar a plantearse la utilización de un filtro paso alto aplicable a cada frame. Con esto se espera que tanto el método que utiliza descriptores ORB como la búsqueda por correlación puedan mejorar. El principal inconveniente será el tiempo de cómputo. El programa tarda en ejecutarse unos 20 minutos, si añadimos esta etapa de filtro a cada imagen el tiempo incrementará. Esto puede generar que el sistema no haya terminado de procesar a tiempo el mapa del entorno antes de que los usuarios quieran comenzar a hacer uno nuevo.

Este entorno médico con el que queremos trabajar, sin duda utilizará herramientas que oculten la visión de la cámara y esto, como hemos visto en el apartado de la evaluación genera grandes errores en la estimación del mapa. Una solución ya propuesta

sería el uso de una red neuronal para evitar que el sistema intente estimar la herramienta entendiendo que esta es parte del mapa [10]. Esto también puede generar un aumento en el tiempo de cómputo, pero al estar la red neuronal ya entrenada, el impacto será leve.

Por último, se propone un cambio en el enfoque a la hora de realizar la plantilla. Dado que el sistema quiere implantarse en endoscopias humanas, se propone un cambio de forma y de uso de coordenadas del sistema, pasando de una malla cuadrada a una coniforme o cilíndrica. Esto puede hacer posible una mejor orientación y avance y retroceso a lo largo de la escena.

# Capítulo 10

# Bibliografía


[1] Jose Lamarca, Shaifali Parashar, Adrien Bartoli, and JMM Montiel. Defslam: Tracking and mapping of deforming scenes from monocular sequences. *arXiv preprint arXiv:1908.08918*, 2019.

[2] Danail Stoyanov, George P Mylonas, Fani Deligianni, Ara Darzi, and Guang Zhong Yang. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 139–146. Springer, 2005.

[3] Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4):14–24, 2010.

[4] Menglong Ye, Stamatia Giannarou, Alexander Meining, and Guang-Zhong Yang. Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. *Medical image analysis*, 30:144–157, 2016.

[5] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[6] Erol Özgür and Adrien Bartoli. Particle-sft: A provably-convergent, fast shape-from-template algorithm. *International Journal of Computer Vision*, 123(2):184–205, 2017.

[7] Ajad Chhatkuli, Daniel Pizarro, and Adrien Bartoli. Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In *BMVC*, 2014.

[8] Adrien Bartoli, Yan Gérard, François Chadebecq, Toby Collins, and Daniel Pizarro. Shape-from-template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2099–2118, 2015.

[9] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):105–119, 2008.

[10] Javier Morlana Ledesma, José María Martínez Montiel, and José Lamarca Peiro. Evaluación y procesamiento de escenas médicas con sistema VSLAM no rígido. 2019.

# Lista de Figuras