Universidad Zaragoza
1542

Trabajo Fin de Grado

Reconocimiento de lugares en SLAM visual con imágenes de endoscopio

Place recognition in visual SLAM with endoscope images

Autor

Alejandro Paricio García

Directores

Juan Domingo Tardós Solano

Juan José Gómez Rodríguez

ESCUELA DE INGENIERÍA Y ARQUITECTURA
2021

AGRADECIMIENTOS

Me gustaría agradecer a mis directores de trabajo de fin de grado, Juan Domingo Tardós y Juan José Gómez, su guía a través de todo el proyecto, mostrándose siempre dispuestos a orientarme, resolver mis dudas y a ayudarme en mi introducción al mundo de la investigación.

También querría agradecer la concesión de la Beca de Colaboración del Ministerio de Educación y Formación Profesional, que me ha permitido formar parte de este proyecto y colaborar con el Departamento de Informática e Ingeniería de Sistemas de la Universidad de Zaragoza.

Por último, agradecer también a mi familia y mis compañeros de promoción todo su apoyo durante estos años de carrera.

RESUMEN

El SLAM Visual consiste en la reconstrucción de un mapa del entorno de un agente móvil mediante cámaras, a la vez que se estima la posición y orientación del agente en el entorno. Este trabajo de fin de grado se centra en un componente esencial del SLAM visual, el reconocimiento de lugares. Su objetivo es determinar si dos imágenes reflejan una misma escena, pudiendo existir cambios en la apariencia de la misma, diferencia entre los instantes en las que son tomadas y variaciones en el punto de vista. Las técnicas actuales resuelven este problema mediante la obtención y el emparejamiento de características visuales. Actualmente, estas pueden dividirse en dos grupos, aquellas basadas en redes neuronales convolucionales y las llamadas características artesanales. No obstante, las técnicas de emparejamiento están pensadas para escenas rígidas, no existiendo técnicas enfocadas a escenas deformables, que sufren cambios en su estructura a lo largo del tiempo, y en las cuales se centra este trabajo.

El objetivo de este trabajo es emplear características visuales para desarrollar un método capaz de lograr buenos resultados en el reconocimiento de lugares en el interior del colon. Para ello, se propone un procedimiento en el que, tras un preprocesamiento inicial de las imágenes, se extraen, filtran con una máscara y emparejan los puntos de interés, y se impone una consistencia geométrica entre los emparejamientos. Se han comparado las características artesanales SIFT, ORB y AKAZE, emparejadas mediante sus descriptores y las extraídas por la red convolucional SuperPoint, reentrenada para el colon, que son emparejadas mediante la red neuronal SuperGlue.

Ambos tipos de características han sido evaluadas en secuencias de endoscopia médica del proyecto europeo EndoMapper. Los resultados demuestran que las características basadas en redes profundas, tras aplicar un reentrenamiento apropiado, proporcionan mejores resultados que las características artesanales, alcanzándose un 70 % de exhaustividad manteniendo la precisión al 100 % en nuestros datos de test.

Índice

Capítulo 1

Introducción y objetivos

1.1. Motivación

El SLAM, o *Simultaneous Localization and Mapping* es el problema que consiste en la reconstrucción de un mapa del entorno de un agente móvil mediante sensores, a la vez que se estima la pose (posición y orientación) del agente en el entorno. Se denomina SLAM Visual al caso concreto en el que los sensores son cámaras. Este trabajo de fin de grado se centra en un componente esencial de los sistemas de localización y construcción de mapas, el reconocimiento de lugares, problema que trata de emparejar imágenes de la misma escena tridimensional tomadas desde posiciones y orientaciones distintas e instantes de tiempo diferentes.

El trabajo se enmarca dentro del proyecto europeo EndoMapper[1], que trata de lograr la reconstrucción de mapas tridimensionales del interior del cuerpo humano a partir de imágenes obtenidas en procedimientos de endoscopia médica monocular. Se plantea un contexto de entornos deformables, como es el colon, que sufre continuas deformaciones, siendo modificado su aspecto y estructura. Esto entra en conflicto con las técnicas de reconocimiento de lugares desarrolladas hasta ahora, que están pensadas para entornos rígidos, como son el interior de los edificios o las calles de una ciudad. Es por tanto un área inexplorada en la que habrá que investigar que métodos son más apropiados, así como si las técnicas actuales pueden llegar a funcionar en este tipo de entornos, abarcando tanto técnicas de extracción de características artesanales, como de características basadas en redes neuronales convolucionales.

Por enfocarse en intervenciones dentro del colon, aparecen dificultades adicionales, como la falta de textura en las paredes del colon o la aparición de herramientas clínicas entre otras, que habrán que tenerse en cuenta al proponer una solución.

[1] `https://sites.google.com/unizar.es/endomapper`

Lograr obtener un buen sistema para reconocimiento de lugares en secuencias de endoscopia permitiría recuperar la localización de la cámara tras oclusiones, como las provocadas por fluidos y mucosas internas. Además, podría llegar a permitir emparejar zonas a la entrada y salida del procedimiento para guiar al cirujano, así como a cerrar otro tipo de bucles durante el recorrido del endoscopio, al mismo tiempo que contribuye a la reconstrucción del mapa del entorno. Por ende, podría suponer un gran paso en la inclusión de este tipo de sistemas en procedimientos dentro el cuerpo humano, contribuyendo al aumento de la calidad de este tipo de intervenciones.

Por último, es necesario remarcar que este trabajo se basa en la investigación realizada con una Beca de Colaboración del Ministerio de Educación y Formación Profesional, dentro del Departamento de Informática e Ingeniería de Sistemas de la Universidad de Zaragoza.

1.2. Objetivos

El objetivo de este trabajo es desarrollar un método capaz de lograr buenos resultados en el reconocimiento de lugares en entornos deformables, en este caso, en el interior del colon. Para ello, se van a estudiar, adaptar y evaluar distintas técnicas de reconocimiento de lugares en imágenes de endoscopia, lidiando con los problemas asociados al interior del cuerpo humano y los retos concretos de este tipo de secuencias. Para ello se plantean los siguientes subobjetivos:

- Estudio y puesta a punto de los principales métodos actuales de reconocimiento de lugares basados en extracción de características, incluyendo tanto características SIFT [1], ORB [2] y AKAZE [3], como las basadas en redes neuronales convolucionales, como es SuperPoint [4].

- Refinamiento y adaptación de los distintos métodos al dominio del problema, incluyendo un reentrenamiento enfocado a este tipo de secuencias de la red neuronal SuperPoint.

- Evaluación de la capacidad de resolver el problema de reconocimiento de lugares de cada uno de los métodos anteriores, midiendo la capacidad de emparejamiento de imágenes pertenecientes a la misma escena.

1.3. Metodología y herramientas

El núcleo del proyecto consiste en el desarrollo y evaluación de dos tipos de métodos, expuestos en el capítulo 3. Para obtener los resultados finales, se han seguido los siguientes pasos:

- Desarrollo e implementación en C++ del método basado en características artesanales, empleando la implementación de los puntos de interés SIFT y AKAZE de OpenCV [5], y la de ORB de ORB-SLAM [6].
- Desarrollo e implementación en Python del método que utiliza características basadas en redes neuronales convolucionales, partiendo del código de las redes SuperPoint [4] y SuperGlue [7].
- Selección de secuencias de endoscopia médica proporcionadas por el proyecto EndoMapper para ser utilizadas como datos de test.
- Evaluación de los métodos anteriores utilizando las secuencias de endoscopia elegidas para este fin.
- Reentrenamiento de la red neuronal SuperPoint. Se emplean secuencias de endoscopia seleccionadas específicamente para el entrenamiento, distintas a las elegidas como secuencias de test. Desarrollo del código y pruebas en Python, empleando librerías como Pytorch [8] u OpenCV.
- Integración de una red neuronal de segmentación de herramientas quirúrgicas [9] en ambos métodos mediante Pytorch.
- Evaluación de la red reentrenada utilizando las secuencias de endoscopia de test.

Otras herramientas utilizadas a lo largo de todo el proyecto son Git para la gestión de versiones del código y LaTeX para la redacción de este documento.

1.4. Descripción del documento

Tras una breve introducción el documento explica, en el capítulo 2, las bases del conocimiento y técnicas empleadas en entorno rígidos, necesarias para comprender los métodos propuestos para entornos deformables en el capítulo 3. Es en este capítulo en el que se exponen las dos principales líneas de métodos a comparar, tanto la basada en características artesanales en el apartado 3.1, como la basada en características obtenidas mediante redes neuronales convolucionales en el apartado 3.2. Descritos ambos, se documenta el reentrenamiento realizado a la red neuronal SuperPoint en el capítulo 4. A continuación, en el capítulo 5, se exponen los resultados obtenidos, qué método ha funcionado mejor, el análisis del efecto de las distintas fases del entrenamiento y el tiempo requerido por cada método. Por último se adjuntan las conclusiones extraídas durante la realización del proyecto y el posible futuro trabajo, junto a los anexos del mismo.

Capítulo 2

Fundamentos previos

2.1. Reconocimiento de lugares

El problema del reconocimiento de lugares consiste en determinar si dos imágenes reflejan una misma escena. Esto abarca posibilidades como que los fotogramas sean tomados en distinto instante temporal, desde distinto punto de vista, o que la escena sufra cambios en su apariencia. Este es un problema que aparece en muchos campos, como la conducción autónoma o la navegación robótica, y que tiene muchas aplicaciones, como por ejemplo, su uso en la creación de panoramas. No obstante, hasta el momento, únicamente se ha estudiado su resolución en escenarios rígidos, en los que se centra este capítulo, en oposición al contexto de escenario deformable en el que se desarrollarán los métodos presentados más adelante. En la figura 2.1 puede observarse una representación del problema de reconocimiento de lugares.

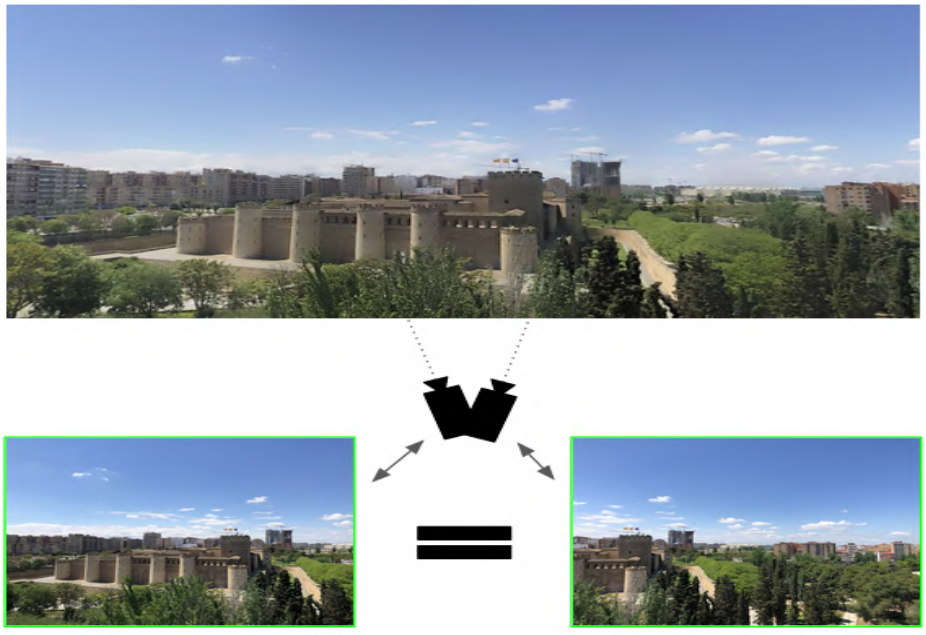


Figura 2.1: Representación del problema de reconocimiento de lugares.

Una de las aproximaciones más comunes es intentar encontrar un conjunto de puntos en una imagen que puedan ser identificados en otra. Estos puntos son los denominados puntos de interés o características, patrones locales bien localizados en posición y, según el tipo, en escala. Es por tanto imprescindible que los patrones sean reconocibles ante variaciones en la localización y orientación de la cámara, o de los objetos de la escena. Idealmente, se pretenden lograr también otros tipos de detección robusta, como ante cambios en la iluminación. Sumado a la detección de estas zonas, en esta aproximación aparece la extracción de los descriptores de los puntos de interés, que almacenan información relacionada con su aspecto visual. Estos son los utilizados para encontrar las correspondencias entre las características.

Con el desarrollo de la disciplina de visión por computador han surgido múltiples tipos de puntos de interés. Para el presente trabajo se va a recurrir a una división entre aquellos métodos que utilizan características artesanales y aquellos que emplean características basadas en redes neuronales convolucionales (CNN).

2.2. Reconocimiento de lugares con características artesanales

Este primer grupo de características es el que primero surgió, estableciéndose como elemento habitual en la resolución de este problema. Existe una amplia variedad de tipos de puntos de interés en este grupo, pero hay algunos que han destacado por encima del resto, como por ejemplo SIFT [1], que es invariante ante rotación y escala, y que se ha instaurado como estándar de calidad en múltiples tipos de escenarios rígidos. Otras alternativas que aparecen en este conjunto son los que poseen descriptores binarios de rápido cálculo, como AKAZE [3] y ORB [2], que también han cobrado fuerza en sistemas que requieren realizar tareas de reconocimiento de lugares a gran velocidad.

Un método muy extendido al utilizarlos para reconocimiento de lugares consiste en, una vez llevada a cabo su detección y la obtención de sus descriptores en ambas imágenes, buscar, para cada uno, y con una medida de distancia apropiada, el descriptor más cercano de entre los de la imagen contraria. Por ejemplo, en el caso de SIFT, se emplea la distancia euclídea como medida de distancia, mientras que para descriptores binarios se utiliza la distancia Hamming. Una vez completada esta búsqueda inicial de emparejamientos, es habitual efectuar un filtrado de emparejamientos válidos mediante heurísticas, tratando de descartar aquellos que son espurios. Ejemplos de estas son el ratio al segundo vecino [1] o la obtención de emparejamientos simétricos. La figura 2.2

muestra un ejemplo de emparejamientos de puntos SIFT aplicando el ratio al segundo vecino.

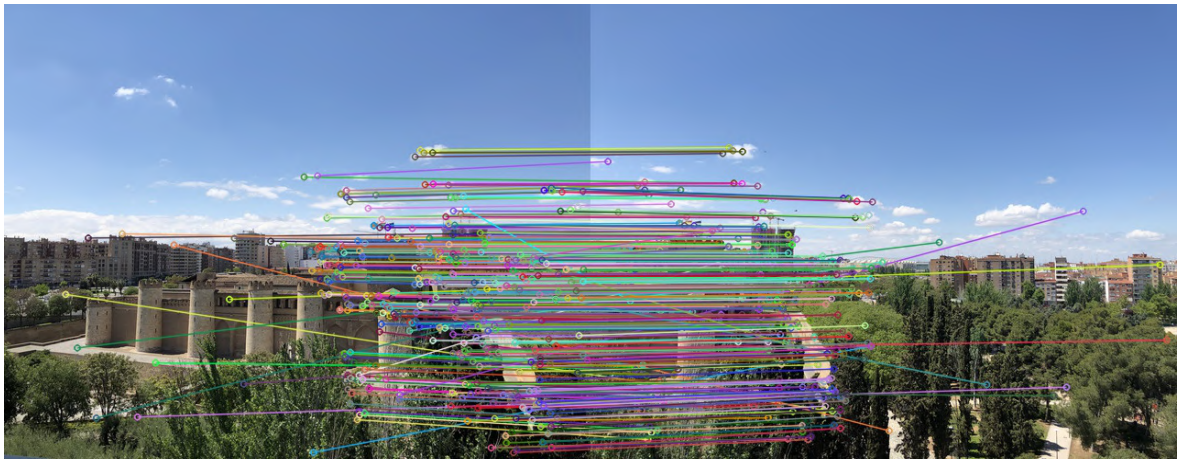


Figura 2.2: Ejemplo de emparejamientos de puntos SIFT aplicando ratio al segundo vecino, escenario rígido.

2.3. Reconocimiento de lugares con características basadas en redes neuronales convolucionales (CNN)

Con la expansión de las redes neuronales a todos los campos, estas han empezado a utilizarse también para tareas específicas relacionadas con el reconocimiento de lugares. Actualmente, tanto la extracción de características como el emparejamiento de las mismas puede ser llevado a cabo por este tipo de sistemas. Combinaciones directas de estas conforman un ejemplo de este tipo de métodos. A la hora de realizar la primera tarea destaca la CNN SuperPoint [4], que extrae puntos de interés que pueden servir de entrada a redes que abordan la segunda, como es SuperGlue [7].

2.3.1. Red neuronal SuperPoint

SuperPoint es una red neuronal convolucional que, partiendo de una imagen inicial, obtiene con una única pasada por la red un valor que representa la probabilidad de que haya punto de interés para cada píxel, además de un descriptor de longitud fija para aquellos píxeles en los que se haya determinado que hay punto de interés. La arquitectura de la red SuperPoint puede observarse en la figura 2.3.

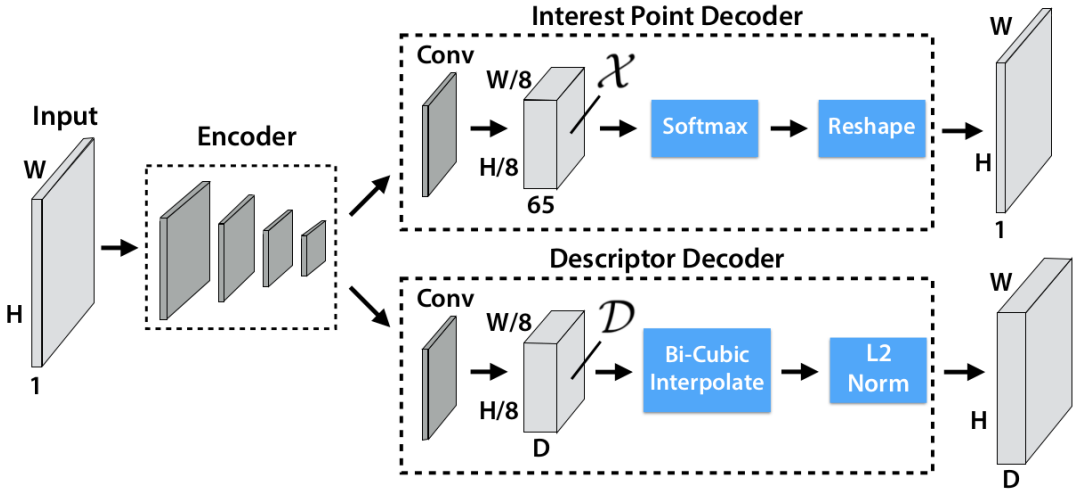


Figura 2.3: Arquitectura de SuperPoint [4].

La red presenta una arquitectura codificador-decodificador con un codificador único que obtiene una representación intermedia para sus dos decodificadores, uno encargado de la detección de características y otro encargado de la extracción de los descriptores.

La salida del decodificador de puntos de interés es tratada para obtener la probabilidad de que haya una característica en cada píxel. Acto seguido se aplica una supresión de no máximos y se seleccionan los píxeles por encima de un umbral, determinando así si lo hay, o no. Por otro lado, la salida del decodificador de descriptores proporciona uno por cada grupo de 8x8 píxeles. Estos son interpolados para obtener el descriptor asociado a la posición de cada característica.

2.3.2. Red neuronal SuperGlue

SuperGlue es una sistema compuesto por dos módulos que, conjuntamente, son capaces de, partiendo de dos conjuntos de características, encontrar una asignación parcial entre ellos, imponiendo que cada una pueda tener como mucho una única correspondencia en el conjunto de la imagen contraria, pero permitiendo que puedan existir algunas sin emparejar. La arquitectura de la red SuperGlue se muestra en la figura 2.4.

El primer módulo es una red neuronal en grafo que, a partir de las posiciones y los descriptores visuales de las características de ambas imágenes obtiene un nuevo descriptor para cada una. Para la obtención del descriptor la red considera las relaciones espaciales y de apariencia visual tanto entre los puntos de interés de la misma imagen como con los de la otra. Con ello consigue desambiguar los emparejamientos dudosos

aprovechando las correspondencias de características cercanas, además de imponer una cierta relación geométrica entre los emparejamientos.

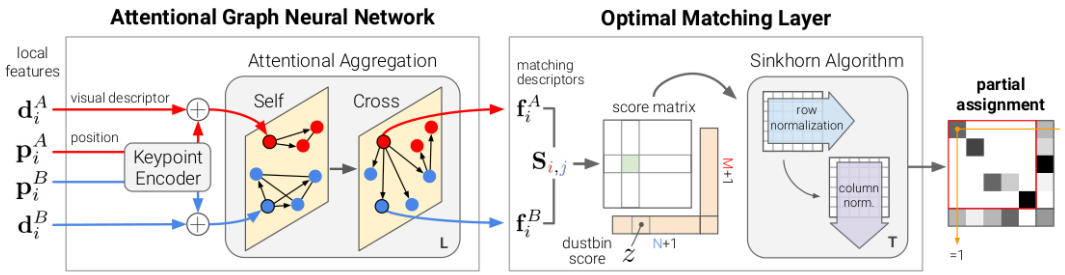


Figura 2.4: Arquitectura de SuperGlue [7].

La segunda capa es la encargada de obtener las asignaciones entre los descriptores calculados en la anterior. Esto se consigue resolviendo un problema de asignación lineal en el que a cada punto de interés solo se le puede asignar uno de la imagen contraria y viceversa, incluyendo la posibilidad de la no existencia de emparejamiento.

La combinación entre las redes SuperPoint y SuperGlue puede ofrecer resultados como los representados en la figura 2.5.

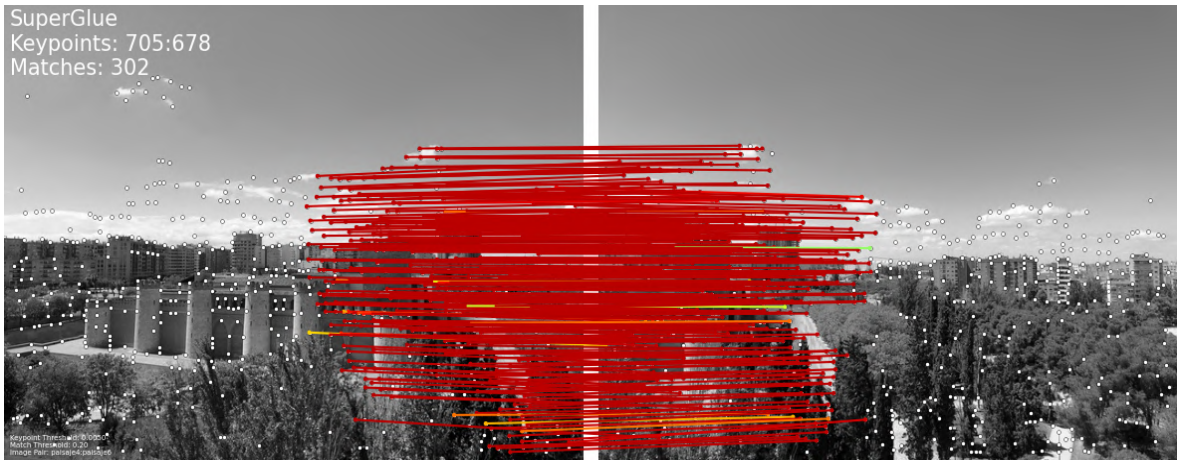


Figura 2.5: Ejemplos de emparejamientos obtenidos con SuperPoint y SuperGlue en un escenario rígido. Los colores de los emparejamientos representan la confianza predicha por SuperGlue. Rojo indica mayor confianza, mientras que azul indica una confianza menor. Se destaca la no aparición de emparejamientos espurios en la imagen.

Capítulo 3

Reconocimiento de lugares en endoscopias

Aprovechando la división entre características establecida en el apartado anterior, se ha diseñado para cada grupo un procedimiento adaptado a las secuencias de endoscopia médica. Ejemplos de resultados obtenidos con cada uno de ellos pueden observarse en el Anexo A.

3.1. Método con características artesanales

El proceso diseñado fundamentado en las características artesanales, como son AKAZE, SIFT y ORB, es el siguiente:

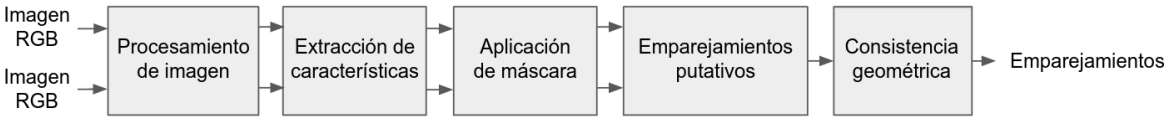
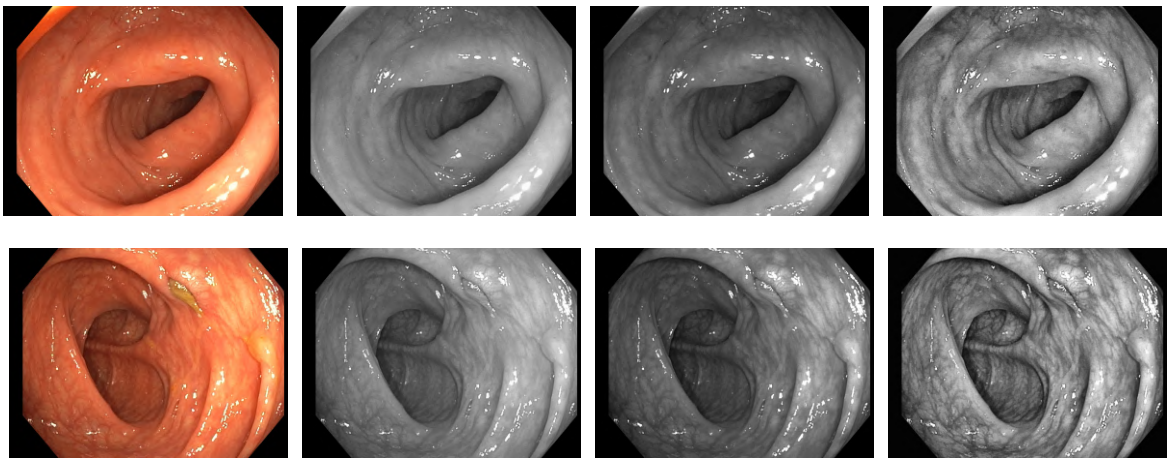


Figura 3.1: Método que emplea características artesanales.

El primer paso es el tratamiento de la imagen. Se consideraron varios tipos de procesamiento, como el uso del canal verde o la conversión de la imagen a escala de grises, probando a combinarlos con redimensiones de la imagen a la mitad de filas y columnas. Se observa en la figura 3.2 cómo algunos procesamientos, como el uso del canal verde, destacan elementos interesantes como son las venas del interior del intestino sin modificar la apariencia visual del resto de los elementos de la escena.



(a) Sin tratamiento (b) Escala de grises (c) Canal verde (d) Ecualizado CLAHE

Figura 3.2: Comparación de imágenes de endoscopia con distintos preprocesados. Tras probar el método con todos ellos, destacan frente al resto el uso del canal verde y de la escala de grises.

A continuación se extraen los puntos de interés correspondientes. El aspecto visual de las paredes del intestino varía mucho entre las secciones del mismo, con el consiguiente cambio de textura. En determinadas zonas estas presentan un aspecto muy uniforme, liso y casi carente de textura visible, lo que dificulta la obtención de características. Así mismo, es imprescindible extraer suficientes como para que, al ser emparejadas y atravesar el resto de las fases del método, permitan tomar la decisión de si ambas imágenes enfocan a la misma sección del colon. Buscando obtener un mínimo de características por imagen se itera la extracción modificando los umbrales de detección, buscando así al menos 1000 puntos de interés a emparejar y evitando establecer un umbral único para todas las imágenes.

Las características obtenidas son filtrados con una máscara, descartando aquellas detectadas en regiones que no son de interés. Su cómputo queda expuesto en la sección 3.3. Se aprovechan las capacidades de OpenCV [5] para llevar a cabo la extracción de características en la zona delimitada por la máscara con una única llamada a función.

El siguiente paso del proceso es obtener los emparejamientos putativos a partir de los puntos de interés. Estos se obtienen por fuerza bruta, empleando una medida de distancia entre descriptores adecuada al punto de interés (Hamming para aquellos con descriptor binario y norma L2 para SIFT). A este primer conjunto se le aplica un filtrado mediante la heurística del ratio al segundo vecino [1], que se ha mostrado superior que otras alternativas, como la obtención de emparejamientos simétricos.

Por último, se lleva a cabo un paso adicional de validación de la consistencia geométrica. Este se basa en el uso de la matriz esencial, y filtra los emparejamientos encontrados que no encajan con la transformación descrita por la anterior, procedimiento detallado en la sección 3.4. La figura 3.3 muestra un ejemplo de resultado obtenido aplicando este método.

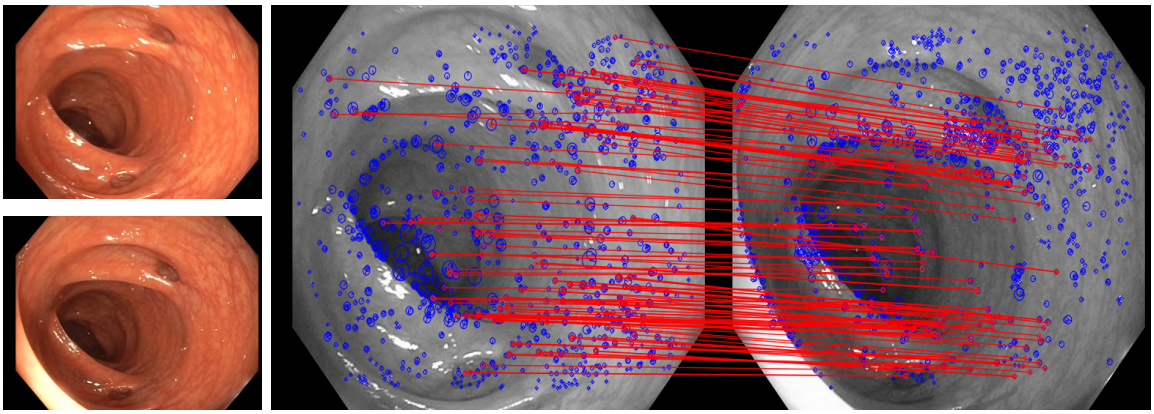


Figura 3.3: Representación del resultado de la ejecución del método con características artesanales. A la izquierda imágenes originales, a la derecha, representación de los emparejamientos obtenidos a partir de características AKAZE empleando la imagen en escala de grises redimensionada a la mitad de filas y columnas.

3.2. Método con características basadas en redes neuronales convolucionales (CNN)

Siguiendo con la idea de desarrollar un método para cada uno de los grupos de puntos de interés, se ha creado el siguiente procedimiento que gira en torno a la obtención de las características y sus emparejamientos mediante redes neuronales:

Figura 3.4: Método con características basadas en redes neuronales convolucionales.

Al igual que con el método anterior, el proceso comienza con un tratamiento a la imagen. Antes del reentrenamiento de la red SuperPoint, descrito en el capítulo 4, se mantenía el uso del canal verde, pasando a instaurarse el uso de la imagen en escala de grises una vez reentrenada la red. Las imágenes son directamente introducidas a la red de SuperPoint, que extrae los puntos de interés y sus descriptores, que son filtrados

con una máscara. Los puntos de interés restantes son pasados a la red SuperGlue, la cual devuelve los emparejamientos putativos.

Los emparejamientos devueltos por la red SuperGlue ya mantienen una cierta coherencia geométrica. No obstante, se impone una restricción adicional sobre estos, llevando a cabo la fase de consistencia geométrica descrita en el apartado 3.4. La figura 3.5 muestra un ejemplo de resultado de la ejecución de este método.

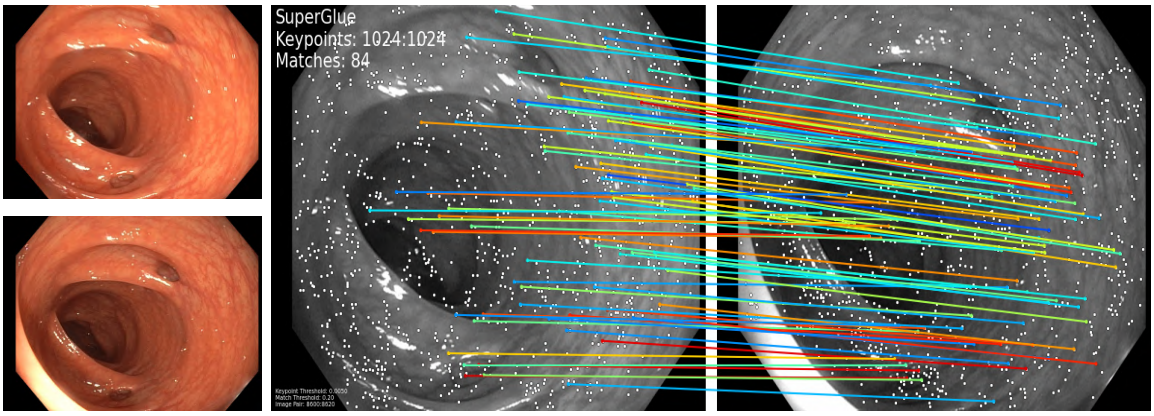


Figura 3.5: Representación del resultado del método basado en redes neuronales. A la izquierda imágenes originales, a la derecha, representación de los emparejamientos obtenidos empleando el canal verde de la imagen. Los colores de los emparejamientos representan la confianza predicha por SuperGlue. Rojo indica mayor confianza, mientras que azul indica una confianza menor.

3.3. Enmascaramiento

Con el objetivo de no utilizar características de baja calidad, o que puedan suponer emparejamientos erróneos, se obtiene una máscara que determina la región de no interés en las imágenes. Esta se centra en los elementos externos a la órganos del ser humano que resultan problemáticos a la hora de llevar a cabo los emparejamientos en secuencias de endoscopia.

El elemento más problemático son los brillos generados en las paredes del intestino, que son provocados por la luz incorporada en el endoscopio. Los anteriores son de tamaño muy variable, y no repetibles ante pequeñas variaciones en la posición de la cámara, incluso entre fotogramas cercanos. Es muy frecuente la detección de puntos de interés en ellos, especialmente en las zonas de menor textura, donde no hay otros elementos tan fácilmente distinguibles para los detectores, y pueden llegar a suponer el emparejamiento erróneo de múltiples características, como puede verse en la figura 3.6. Para calcularla, se computa una máscara inicial seleccionando los píxeles con un valor

superior a 225. Acto seguido se le aplica una dilatación y se invierte el resultado, obteniendo la máscara de brillos final.

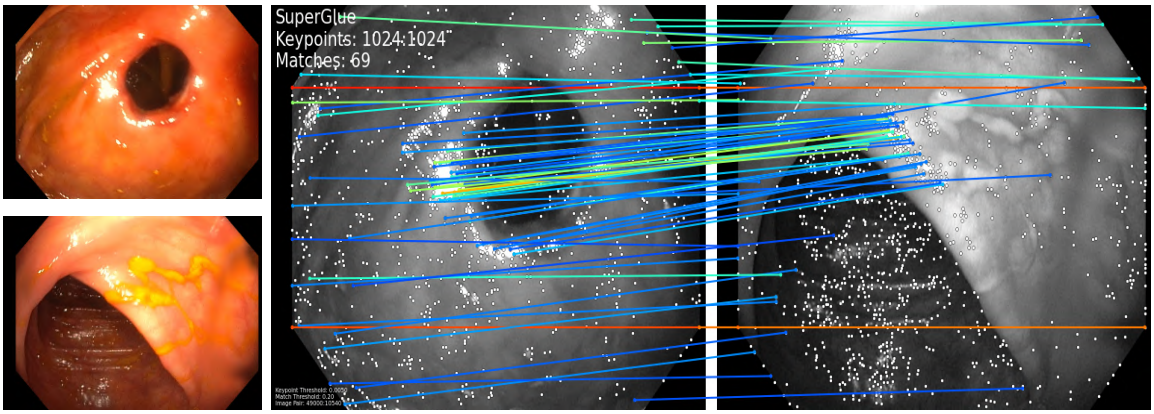


Figura 3.6: Ejemplo de emparejamientos erróneos provocados por brillos utilizando el canal verde de la imagen y sin aplicar máscara ni fase de consistencia geométrica.

El siguiente aspecto destacable es la aparición de las herramientas de las intervenciones que se llevan a cabo en el colon. Pese a ser un elemento no siempre presente, el uso de una de ellas en dos zonas distintas del colon podría llevar a falsos emparejamientos derivados de las características extraídas en estas, por lo que se ha decidido enmascararlas. Con este fin, se ha recurrido a una red neuronal de segmentación de herramientas quirúrgicas [9], que permite obtener a partir de la imagen original una máscara binaria adecuada para los fines del enmascaramiento.

En conjunto a las anteriores, se añade una máscara fija común a todas las imágenes para eliminar las detecciones en los bordes de la imagen. Uniendo las tres puede crearse un procedimiento para generar máscaras como el representado en la figura 3.8, permitiendo filtrar las detecciones en las zonas mencionadas, lográndose resultados como los siguientes:

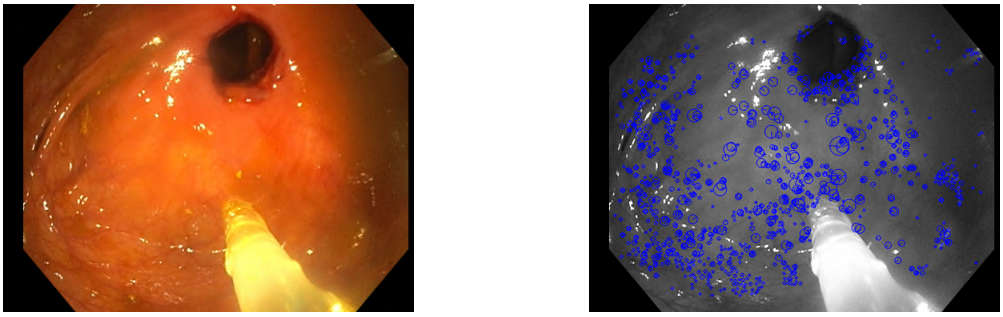


Imagen original Puntos extraídos

Figura 3.7: Ejemplo de detección de características AKAZE con filtrado con máscara.

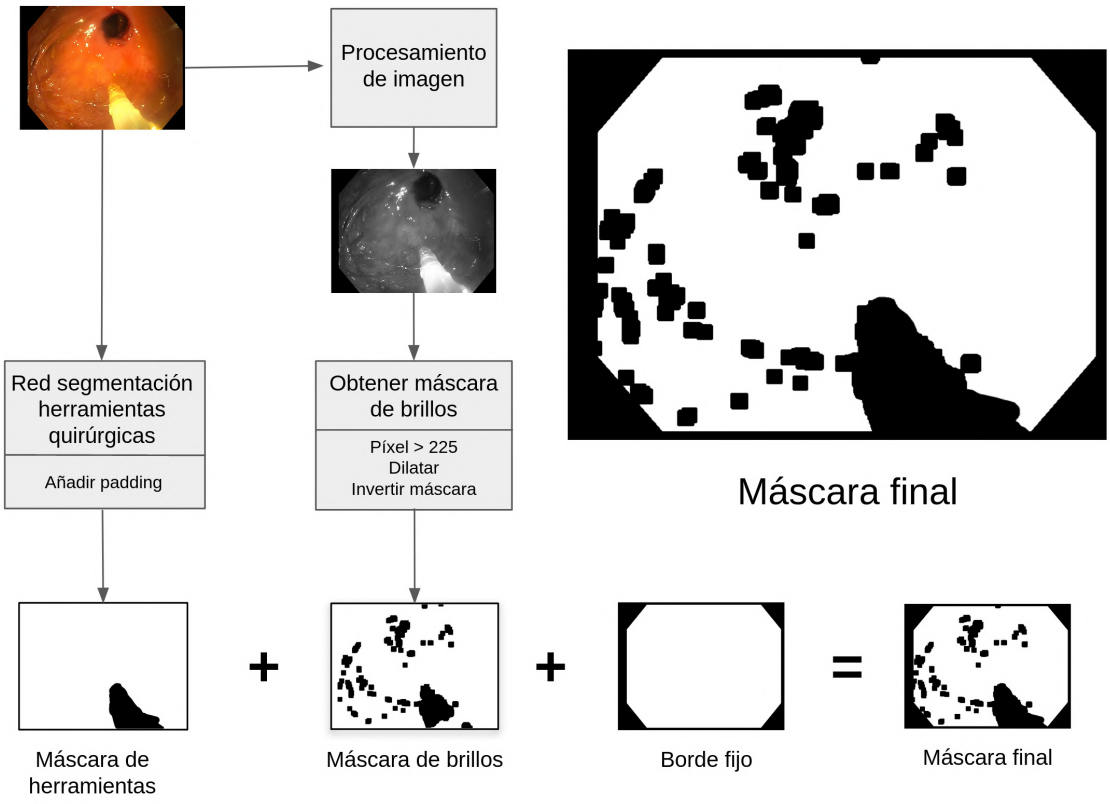


Figura 3.8: Representación del proceso de obtención de la máscara.

3.4. Consistencia geométrica

Los emparejamientos putativos obtenidos antes de esta fase no siguen ningún tipo restricción geométrica formal que asegure una cierta coherencia entre ellos. Para imponerla, puede utilizarse la restricción epipolar, que restringe la proyección sobre el plano de imagen de un mismo punto X desde dos posiciones y orientaciones de cámara distintas.

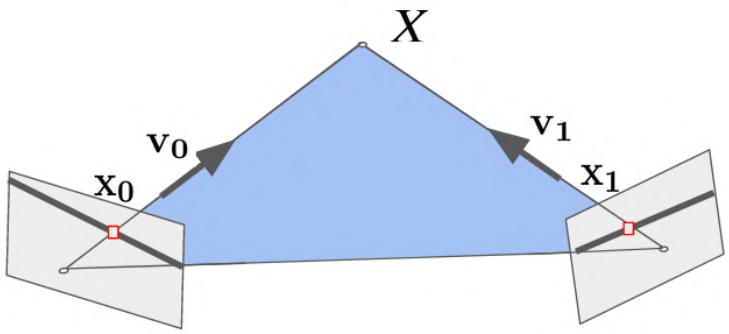


Figura 3.9: Se refleja en el diagrama la visión de un mismo punto tridimensional X desde dos imágenes distintas. La restricción epipolar puede aplicarse sobre la proyección de X en ambos planos de imagen, x_0 y x_1.

Concretamente, con esta puede calcularse, dada la proyección sobre el plano de imagen del punto en la primera imagen x_0, y la matriz esencial, la línea sobre el otro fotograma en la que debe aparecer la proyección x_1 del mismo punto. La representación gráfica de los mismos puede apreciarse en la figura 3.9. La ecuación que define la restricción es la siguiente:

$$x_1^T \, E \, x_0 = 0 \tag{3.1}$$

En la ecuación anterior se utilizan las proyecciones en el plano de imagen $x_0, x_1 \in \mathbb{R}^3$ del punto X en coordenadas homogéneas, es decir, el valor de la tercera coordenada es 1. Esto entra en conflicto con el uso de cámaras tipo ojo de pez en los endoscopios. Las cámaras ojo de pez permiten un ángulo de visión muy grande, incluso mayores que 180°. Por tanto, pueden extraerse características situadas a 90^{o} del eje óptico, en las que su proyección sobre el plano de imagen tendría un 0 en la tercera coordenada. Por tanto, no es posible utilizar las coordenadas homogéneas en este caso.

Para solventar este inconveniente se ha optado por emplear directamente la matriz esencial sobre los vectores directores unitarios $v_i \in \mathbb{R}^3$ de cada par de puntos emparejados. Para ello, los últimos se proyectan al espacio tridimensional empleando los parámetros de la cámara, siguiendo un modelo de Kannala-Brandt [10]. Los vectores directores obtenidos deben cumplir la siguiente restricción:

$$v_1^T \, E \, v_0 = 0 \tag{3.2}$$

Para obtener la E final se recurre a un RANSAC que trata de obtener el modelo con el mayor número de emparejamientos legítimos posible. Al tratarse de un entorno en continua deformación, aparecen variaciones en la estructura tridimensional de la escena, siendo necesario tolerar este tipo de deformaciones. Esta tolerancia es introducida a la hora de decidir si los vectores directores de un par de puntos emparejados están de acuerdo con el modelo calculado.

Se puede aprovechar el haber trasladado los puntos al espacio 3D mediante la obtención de sus vectores directores para establecer cuánta desviación se permite entre las vectores ideales marcados por la restricción epipolar y los reales. El primer producto de la ecuación 3.2 proporciona la normal del plano formado por el punto X y ambos centros ópticos. En la teoría, sin existencia de deformación, el ángulo entre v_0 y la normal debería ser 90^{o}. En la práctica, se ha establecido un umbral bajo el cual el emparejamiento acepta el modelo.

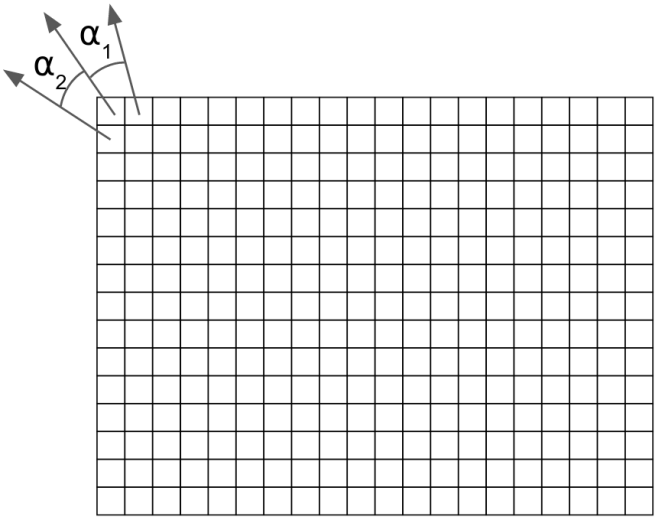


Figura 3.10: En la imagen se muestra representado el ángulo entre los vectores directores de un par de píxeles contiguos en horizontal, α_1, así como el existente entre los de un par de píxeles contiguos en vertical, α_2.

Para obtener el umbral, primero se ha medido el ángulo entre los vectores directores de cada par de píxeles contiguos de una imagen de endoscopia médica tanto vertical como horizontalmente, tal y como se representa en la figura 3.10.

Acto seguido, se calcula la mediana de todos los ángulos extraídos, obteniendo una medida del ángulo que abarca cada píxel individual de la imagen. La obtención de este valor se refleja en la ecuación 3.3.

$$\delta \; = \; median(\alpha_i) \; = \; 0{,}001553343 \; \frac{rad}{píxel} \tag{3.3}$$

Finalmente, se calcula el umbral definitivo th de acuerdo a la ecuación 3.4, en la que n representa el número de píxeles de desviación permitidos.

$$th \; = \delta \; * \; n \tag{3.4}$$

Empíricamente se ha ajustado n a un valor de 10 píxeles para el método que emplea características basadas en redes neuronales convolucionales y de 14 píxeles en el caso del método que emplea características artesanales. La diferencia en los valores se puede atribuir a la coherencia geométrica impuesta por la red SuperGlue sobre los emparejamientos, que exige un filtrado más estricto para descartar los falsos emparejamientos. Cualquier par de puntos cuyos vectores directores se desvíen de la posición ideal menos de th radianes son considerados emparejamientos legítimos y aceptan el modelo calculado.

Capítulo 4

Reentrenamiento de SuperPoint

La red SuperPoint [4] cuenta con unos pesos predeterminados que pueden ser utilizados directamente en secuencias de endoscopia obteniendo resultados aceptables. No obstante, se ha intentado lograr un refinamiento de los pesos de la misma especializado para el tipo de secuencias mencionado. Para poder comprender en qué consiste este proceso, primero se expone el entrenamiento original realizado a SuperPoint.

4.1. Entrenamiento original de SuperPoint

El entrenamiento original de la red SuperPoint se divide en tres fases, representadas en la figura 4.1.

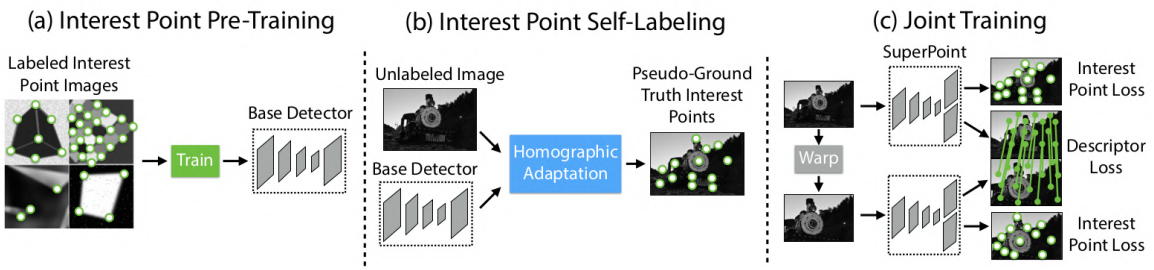


Figura 4.1: Entrenamiento original de SuperPoint [4]. Se muestran las tres fases empleadas: la primera basada en imágenes sintéticas, la segunda que emplea el proceso de *Homographic Adaptation*, y la tercera, que entrena conjuntamente la detección de características y la extracción de descriptores.

La primera fase empieza a entrenar tanto el codificador como el decodificador de puntos de interés, estableciendo los lugares donde deben extraerse las características. Para ello, se utiliza un conjunto de imágenes sintéticas que consiste en combinaciones de figuras geométricas 2D, el cual no ha sido publicado. Las etiquetas de los puntos de interés se establecen en las intersecciones, los extremos de líneas rectas y el centro de

pequeñas elipses. Estas imágenes sintéticas se generan al vuelo, de forma que durante la primera fase del entrenamiento la red no ve dos veces la misma imagen.

Acto seguido se lleva a cabo la segunda fase del entrenamiento en la que se intenta aumentar la capacidad de generalización a las detecciones de la red SuperPoint, buscando que se adapte a imágenes del mundo real. En esta fase se emplea el conjunto de imágenes no etiquetadas MS-COCO 2014 [11] redimensionadas a una resolución de 240x320 píxeles y convertidas a escala de grises. Para generar las etiquetas de puntos de interés se emplea el proceso llamado *Homographic Adaptation*, expuesto en la figura 4.2.

El proceso de *Homographic Adaptation* comienza aplicando a una imagen un conjunto de homografías para deformarla, en este caso 100. Empleando el detector resultante de la primera fase, se extraen características en cada una de las deformaciones. A las características detectadas se les aplica la transformación inversa a la utilizada para deformar la imagen, obteniendo su proyección sobre la imagen original. Al agregar las características obtenidas en todas las deformaciones se obtiene un conjunto más amplio de etiquetas de puntos de interés. Con las imágenes originales y las etiquetas obtenidas se lleva a cabo un aprendizaje supervisado. Una vez finalizado, se vuelven a generar etiquetas con la red resultante, volviéndose a repetir esta fase una segunda vez.

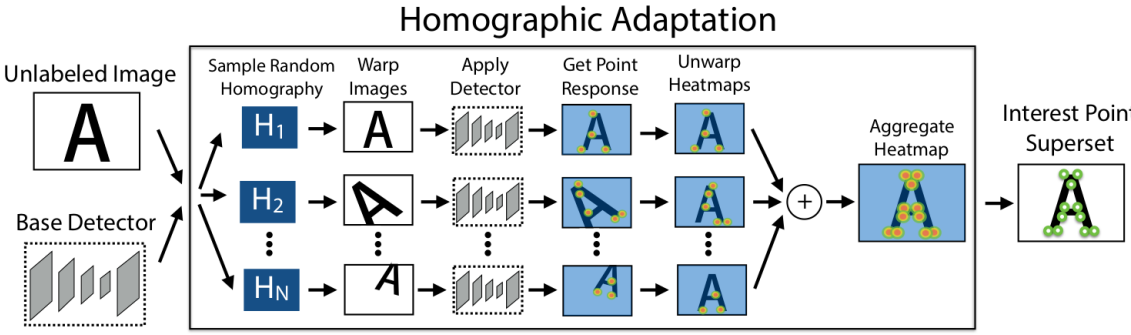


Figura 4.2: Proceso de *Homographic Adaptation* [4]. Proceso que obtiene las etiquetas de características para una imagen mediante el uso de múltiples homografías para deformar la imagen. A partir de las deformaciones se extraen las características, que son proyectadas sobre la imagen original mediante la transformación inversa.

Por último, se lleva a cabo un entrenamiento conjunto de la red completa, incluyendo el decodificador de descriptores. En este, mostrado en la tercera parte de la figura 4.1, se vuelven a utilizar las imágenes del conjunto MS-COCO a resolución 240x320 píxeles en escala de grises.

Durante la última fase, cada muestra está compuesta de dos imágenes, con sus respectivos conjuntos de puntos de interés y una homografía que las relaciona. Para obtener las muestras a partir de las imágenes originales, para cada imagen se extraen puntos de interés con el detector resultante de la segunda fase. Además, se genera un recorte aleatorio compuesto de transformaciones sencillas como rotación, traslación o escalado. Este recorte delimita una región de la imagen que es proyectada a las dimensiones originales del fotograma mediante una homografía, formando así la segunda imagen. Las etiquetas de puntos de interés del nuevo fotograma se obtienen aplicando la homografía a las del primero, buscando con ello detecciones más robustas.

Se lleva entonces a cabo un aprendizaje supervisado en el que, al pasar cada muestra por la red se calcula una función de coste de puntos de interés sobre ambas imágenes. Además, se calcula una función de coste de descriptores. Es la homografía que relaciona ambas imágenes lo que permite determinar qué descriptores deben parecerse y cuáles no, tratando de obtenerse descriptores invariantes a transformaciones sencillas. Con esta combinación de funciones de coste puede llevarse a cabo el entrenamiento conjunto del decodificador de puntos de interés y del decodificador de descriptores, finalizando el entrenamiento de SuperPoint.

4.2. Transferencia de aprendizaje a secuencias de endoscopia

El conjunto de datos sintéticos de figuras geométricas del entrenamiento original de SuperPoint no ha sido publicado. La ausencia de este ha llevado a optar por transferir el conocimiento de la red a las escenas de endoscopia en lugar de llevar a cabo un entrenamiento desde cero. No obstante, aparece una dificultad adicional, ya que tampoco se contaba con un conjunto de imágenes de endoscopia con etiquetas de puntos de interés.

El entrenamiento realizado ha sido dividido en dos fases. En la primera, se intenta especializar la extracción de puntos de interés reutilizando la idea de emplear imágenes generadas sintéticamente. Esta primera fase actúa como sustituto de las fases primera y segunda del entrenamiento original. Mientras tanto, en la segunda, se emplean las imágenes del colon y se intenta refinar la obtención de descriptores, imitando la tercera fase del entrenamiento original.

4.2.1. Fase venas

Una de las principales dificultades aparece al intentar determinar qué puntos puede ser interesante extraer en entornos con tan poca textura, ya que hay pocos elementos reconocibles que puedan ser utilizados durante el emparejamiento, siendo el más destacable, y en el que se centra esta fase, las venas.

El entrenamiento de esta fase emplea el conjunto de imágenes MS-COCO 2014 [11], etiquetado con los puntos de interés generados por SuperPoint, junto con imágenes generadas sintéticamente por nosotros que tratan de emular la geometría de las venas, con los puntos de interés localizados en las intersecciones de las mismas, tal y como puede observarse en la figura 4.3. Se genera una imagen de venas sintéticas por cada 3 del conjunto MS-COCO. De esta forma se intenta que la red afiance y amplíe su capacidad de extracción de los puntos en las venas sin alejarse de su capacidad previa. Adicionalmente, se incluyen en las imágenes de venas sintéticas artefactos que emulan los brillos de las secuencias, con el objetivo de paliar la extracción de características en estas zonas. Todas las imágenes se fijan a un tamaño de 240x320 píxeles y son transformadas a escala de grises.

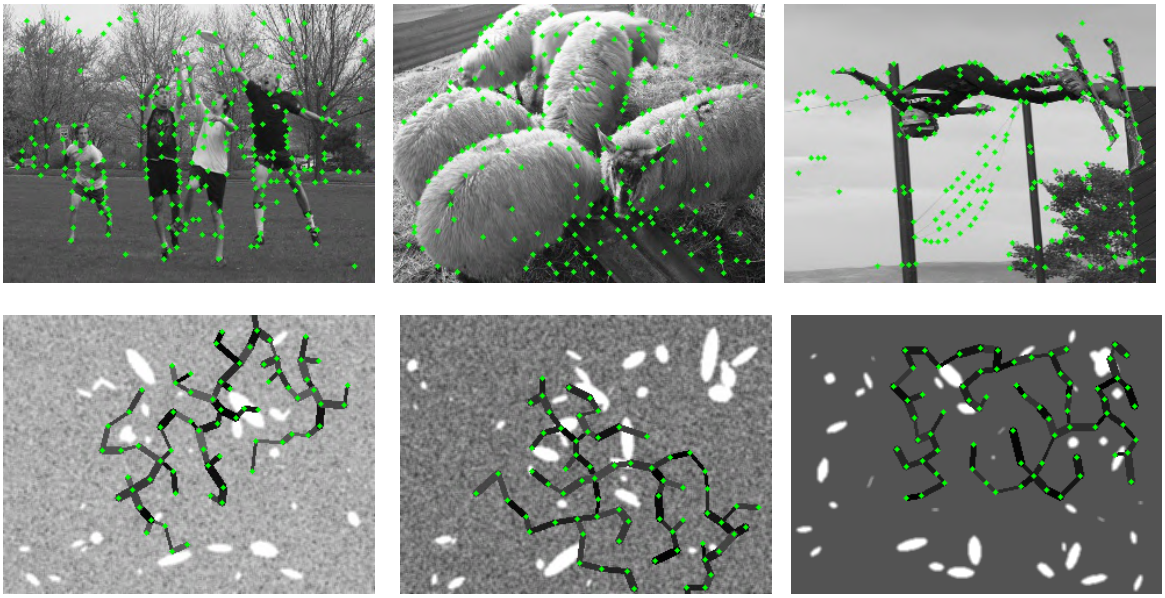


Figura 4.3: Ejemplo de imágenes empleadas durante la primera fase del entrenamiento. Las ubicaciones de las etiquetas de puntos de interés aparecen marcadas en verde.

4.2.2. Fase colon

Habiendo modificado los puntos de interés extraídos por la red, se puede comenzar el entrenamiento conjunto de los puntos y sus descriptores en fotogramas de secuencias

de endoscopia. Para ello, ha de formarse un conjunto de imágenes de este tipo. Con este objetivo en mente, se ha recurrido a obtener fotogramas de estas secuencias separados por un cuarto de segundo, seguido de un filtrado manual de aquellos en los que la cámara se encuentra pegada a la pared u ocluida. Pueden observarse ejemplos en la figura 4.4. Cada una de estas imágenes, una vez transformadas a escala de grises, se divide en 4 fragmentos de igual tamaño que son redimensionados a un tamaño de 240x320 píxeles.

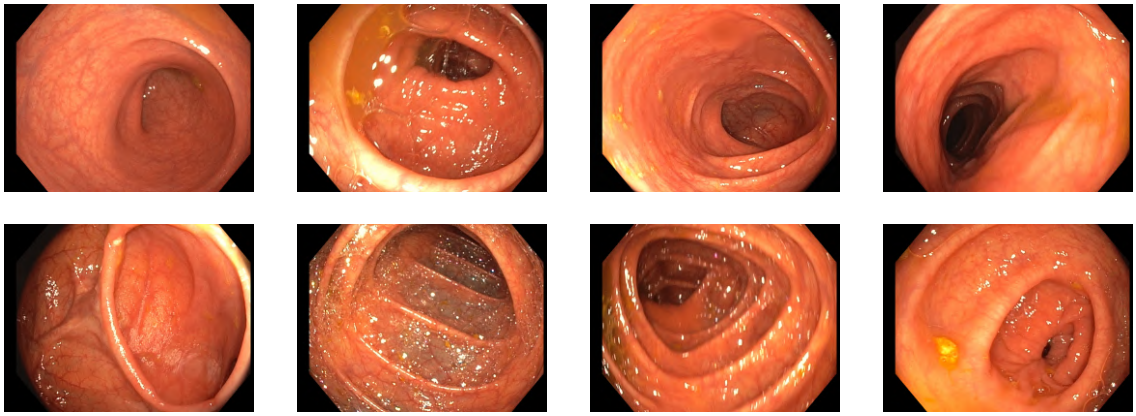


Figura 4.4: Ejemplo de imágenes utilizadas para generar los datos de entrenamiento.

Una vez obtenidas las imágenes han de generarse las muestras para el entrenamiento. Para ello, se emplea el mismo mecanismo utilizado en la tercera fase del entrenamiento original de SuperPoint, obteniendo resultados como los mostrados en la figura 4.5.

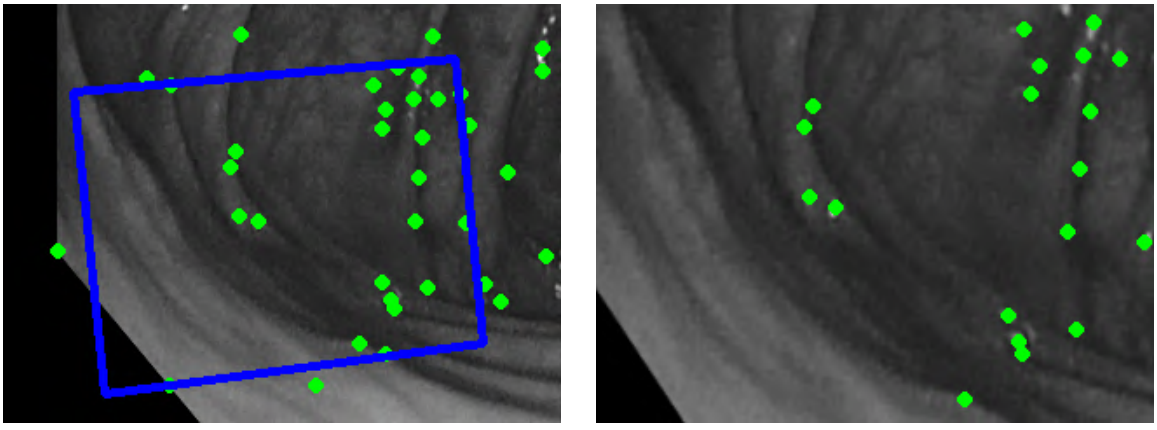
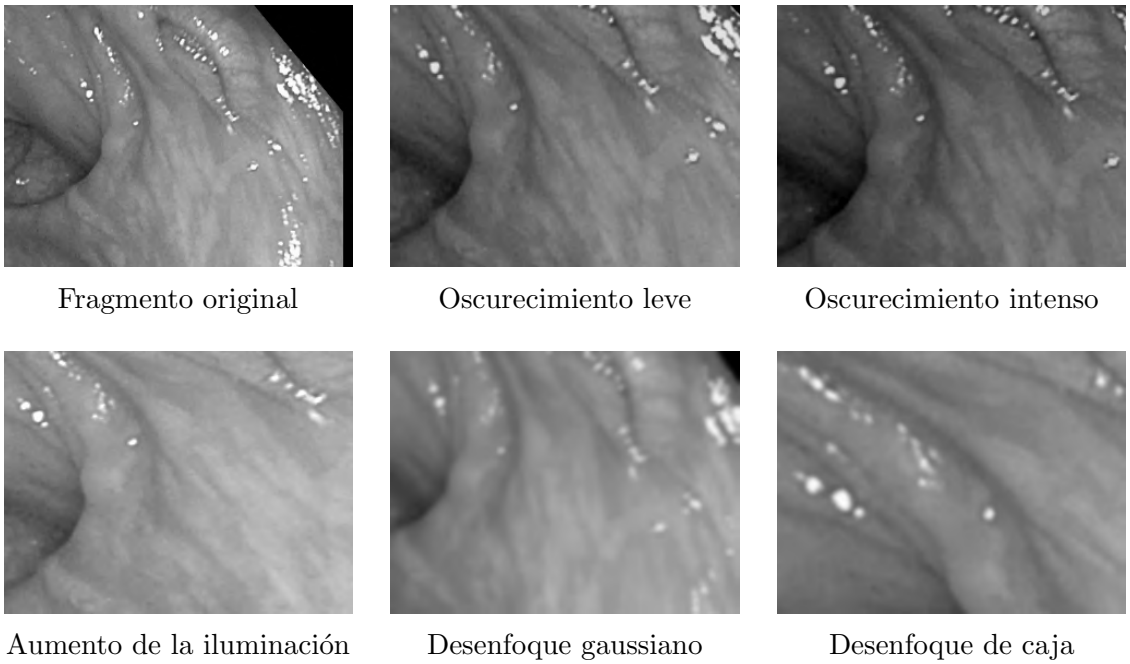


Figura 4.5: Ejemplo de muestra generada para la fase del colon del entrenamiento. A la izquierda, fragmento original en escala de grises redimensionado a 240x320 píxeles con el recorte dibujado en azul. A la derecha, imagen resultante de aplicar la homografía dada por la proyección del recorte a las dimensiones del fragmento original. En ambas imágenes se representan en verde las localizaciones de características.

4.2.3. Aumentación de datos

A la hora de llevar a cabo el entrenamiento se han intentado tener en cuenta otro tipo de alteraciones comunes en las secuencias de endoscopia. La primera de ellas se relaciona directamente con la luz incorporada en el endoscopio. Variaciones en la orientación de la misma generan cambios en la iluminación de la escena, por lo que centrarse en este tipo de entornos exige tenerlas en cuenta. El segundo se relaciona con la aparición de desenfoque en las imágenes, que puede aparecer tanto debido al movimiento de la cámara como a fluidos, ya sean del interior del colon o de las herramientas utilizadas por el especialista.

Para tratar de paliar estos problemas, se han generado nuevas muestras para cada uno de los fragmentos que son parte del conjunto de datos original. Concretamente, se han generado nuevas muestras en las que a la imagen resultante de aplicar una nueva homografía al fragmento original se le han aplicado efectos como oscurecimiento, aumento de la iluminación o desenfoque.



Fragmento original | Oscurecimiento leve | Oscurecimiento intenso

Aumento de la iluminación | Desenfoque gaussiano | Desenfoque de caja

Figura 4.6: Ejemplos de efectos de aumentación de datos aplicados durante la generación de las muestras. Para cada uno de los 5 efectos mostrados, se ha generado a partir del fragmento original una nueva imagen, a la cual se le ha aplicado el efecto indicado en cada caso.

4.2.4. Efecto de la transferencia de aprendizaje en la detección de características

Uno de los aspectos que difieren entre la red antes y después de su reentrenamiento son los lugares en los que se detectan las características. El propósito de la fase de las venas es obtener mejores detecciones en estas estructuras. Por tanto, si la fase de las venas cumple con su propósito, debe poder apreciarse esta variación en las ubicaciones de las características extraídas. En la figura 4.7 se puede apreciar la diferencia.

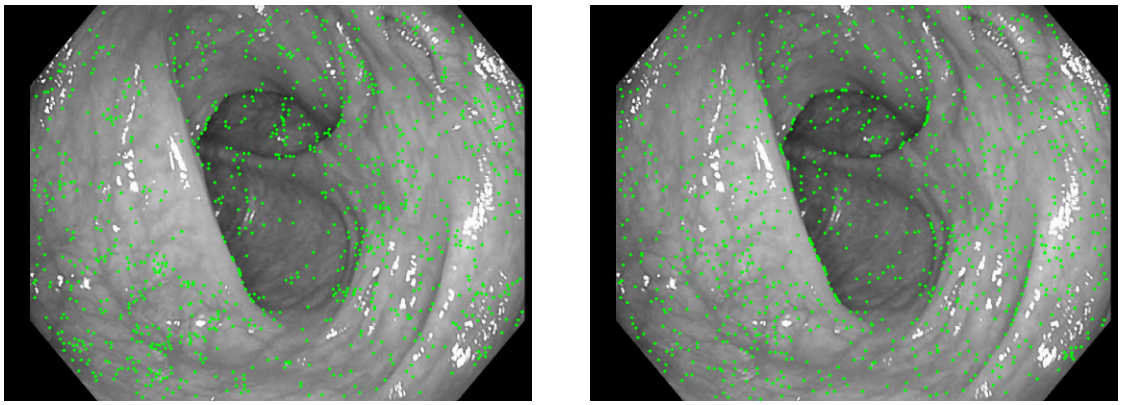


Figura 4.7: Diferencia en la detección de características en las venas antes (izquierda) y después (derecha) del reentrenamiento. En ambos casos se ha aplicado la máscara completa y se ha extraído el mismo número de características. Se observa cómo en la segunda imagen los puntos aparecen más distribuidos por las venas de la zona inferior izquierda, sustituyendo a agrupaciones de puntos de la primera imagen.

Capítulo 5

Resultados

5.1. Metodología de evaluación

A la hora de seleccionar métricas para la evaluación ha de tenerse en cuenta la aplicación para la que se ha desarrollado el procedimiento. En el caso de reconocimiento de lugares en un sistema de SLAM visual en endoscopias, ha de considerarse el coste de los falsos emparejamientos. Al declararse un emparejamiento de este tipo, el sistema trataría de ajustar la reconstrucción obtenida hasta ese momento para que encaje con la nueva posición y orientación de la cámara. Además, corregir este error supondría un coste temporal notable, que es crucial en sistemas que operan en tiempo real.

Ante la ausencia de un *ground truth* de puntos de interés y de las transformaciones relativas entre las imágenes, se restringen las posibles métricas a emplear. La herramienta principal utilizada para la evaluación serán las curvas de *precision–recall*, calculadas variando el número de puntos emparejados necesarios para declarar un emparejamiento positivo. La métrica principal será la máxima exhaustividad (*recall*) que mantenga la precisión (*precision*) al 100 %. Es decir, se maximizará el número de reconocimientos de lugares verdaderos positivos sin que aparezca ningún falso positivo.

Adicionalmente, al ser un sistema pensado para actuar como módulo de otro más amplio, se va a llevar a cabo un análisis del tiempo empleado por fase, para así poder valorar su adecuación o posible implicación en otros sistemas.

Durante el desarrollo del proyecto han sido utilizadas múltiples secuencias de endoscopia. A partir de grupos de secuencias distintos han sido obtenidos los datos tanto para el reentrenamiento de la red SuperPoint, como para evaluar los distintos modelos. Las secuencias seleccionadas con estos fines pueden observarse en la tabla 5.1.

Secuencia	Evaluación		Rentrenamiento de SuperPoint
	Pares positivos	Pares distractores	
HCULB_00039	✓	✓	
HCULB_00045		✓	
HCULB_00053		✓	
HCULB_00033			✓
HCULB_00034			✓
HCULB_00048			✓
HCULB_00089			✓
HCULB_00327			✓
HCULB_00357			✓
HCULB_00372			✓

Tabla 5.1: Uso dado a las secuencias de endoscopia médica utilizadas.

5.2. Conjunto de datos de evaluación

El conjunto de datos de evaluación preparado está formado por un total de 2122 pares de imágenes, que se dividen en 1061 pares positivos y 1061 pares distractores.

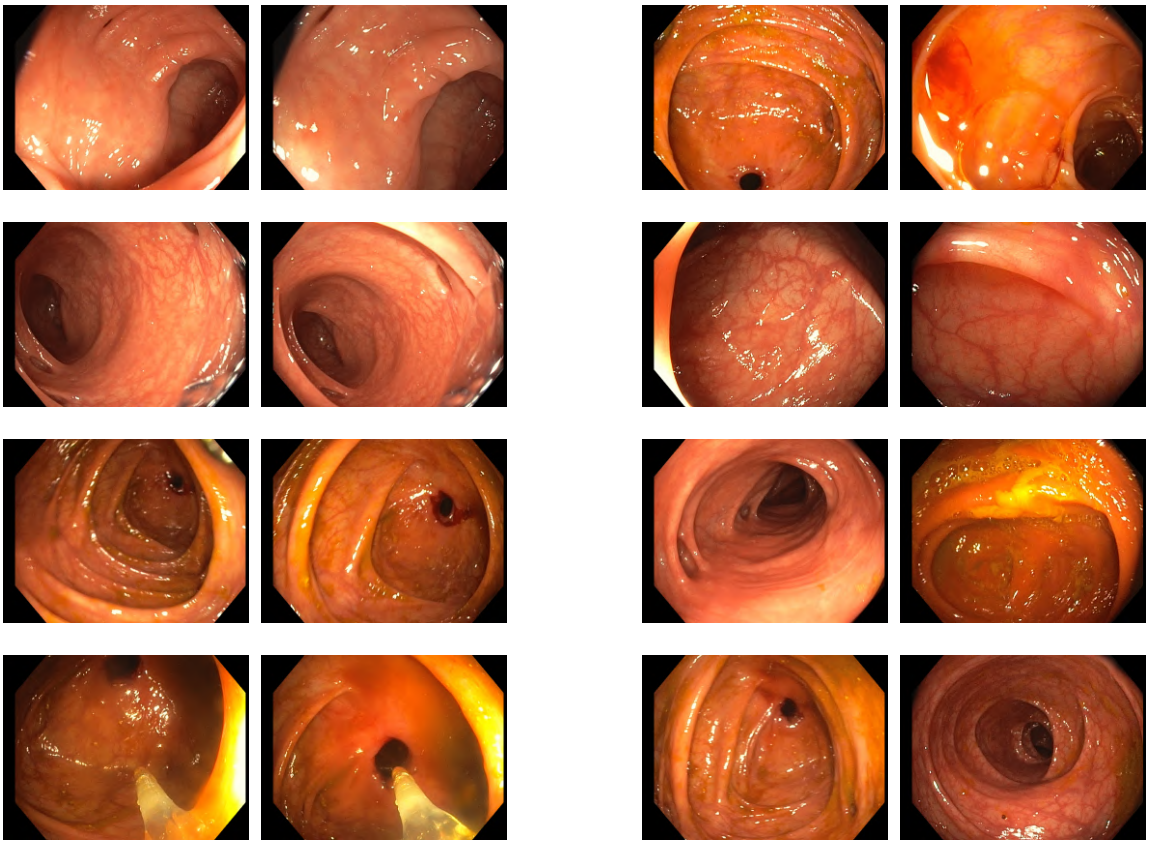


Figura 5.1: Ejemplos del conjunto de datos de evaluación. A la izquierda pares positivos formados por imágenes de la secuencia HCULB_00039. A la derecha, pares distractores formados por una imagen de la secuencia HCULB_00039 y otra de la secuencia HCULB_00045 o HCULB_00053.

Los pares positivos se han formado con imágenes de una misma secuencia separadas entre sí medio segundo, en cambio, los distractores se componen de imágenes de secuencias distintas. Se ha llevado a cabo un filtrado manual previo para eliminar aquellas imágenes en las que la cámara se encuentra totalmente ocluida o pegada a la pared del colon.

5.3. Resultados de los distintos métodos

A continuación se muestran los resultados de la evaluación de los distintos métodos:

Modelo	Procesado imagen	Consistencia geométrica	Recall (%) @precision 1	Emparejamientos necesarios
SIFT	Escala grises		23.19	94
SIFT	Escala grises	✓	29.88	17
SIFT	Verde + *resize*		35.63	70
SIFT	Verde + *resize*	✓	38.27	13
SIFT	Grises + *resize*		31.16	78
SIFT	Grises + *resize*	✓	37.87	13
ORB	Escala grises		18.57	56
ORB	Escala grises	✓	20.54	13
ORB	Verde + *resize*		32.52	59
ORB	Verde + *resize*	✓	35.63	11
ORB	Grises + *resize*		30.41	61
ORB	Grises + *resize*	✓	33.14	13
AKAZE	Escala grises		34.40	131
AKAZE	Escala grises	✓	50.05	18
AKAZE	Verde + *resize*		58.62	60
AKAZE	Verde + *resize*	✓	59.85	12
AKAZE	Grises + *resize*		58.83	64
AKAZE	Grises + *resize*	✓	**61.47**	21
SuperPoint	Escala grises		52.97	60
SuperPoint	Escala grises	✓	54.00	16
SuperPoint	Canal verde		55.80	47
SuperPoint	Canal verde	✓	56.83	12
SuperPoint (reentrenado)	Escala grises		64.09	98
SuperPoint (reentrenado)	Escala grises	✓	**70.69**	15
SuperPoint (reentrenado)	Canal verde		66.82	82
SuperPoint (reentrenado)	Canal verde	✓	69.09	13

Tabla 5.2: Evaluación de los distintos métodos

Entre los puntos de interés clásicos destaca claramente AKAZE, con un rendimiento superior a ORB y SIFT. En todos los casos, utilizar una imagen redimensionada a la mitad mejora algo los resultados, lo que puede ser debido a la pérdida de resolución que implica el patrón de Bayer de las imágenes en color originales. Respecto a los canales

utilizados, el canal verde ofrece una cierta ventaja sobre el canal de niveles de gris para SIFT y ORB, pero la diferencia se invierte en el caso de AKAZE, lo que probablemente se debe a los diferentes procesamientos de imagen realizados por cada detector.

En el lado de SuperPoint antes de su reentrenamiento, el mejor resultado se obtuvo con el canal verde, situándose ligeramente por debajo de AKAZE. No obstante, tras el reentrenamiento, que se realizó en escala de grises, Superpoint con el canal gris asciende para colocarse con diferencia como el mejor de los métodos evaluados, alcanzando el 70 % de *recall* manteniendo la *precision* al máximo. Es de destacar el hecho de que, a excepción de AKAZE, el resto de características artesanales evaluadas no se acercan al rendimiento obtenido por las características basadas en redes neuronales convolucionales. En la figura 5.2 pueden observarse las curvas *precision–recall* obtenidas con el mejor resultado con cada característica.

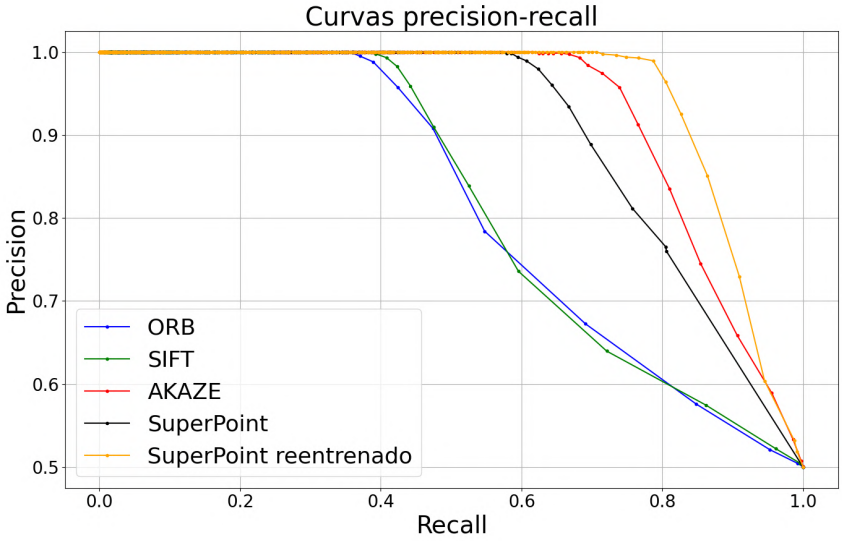


Figura 5.2: Curvas precision-recall de las mejores versiones de cada punto de interés.

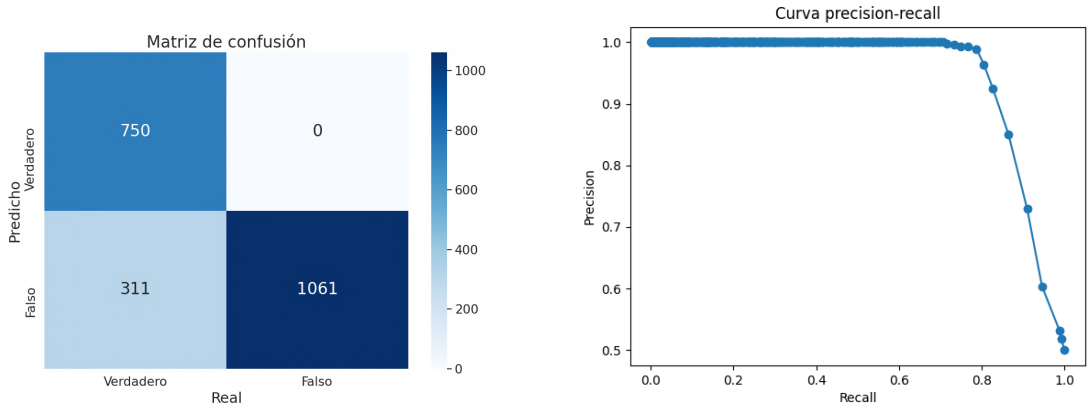
Debido a estar el entrenamiento de la red dividido en fases, puede ser interesante determinar cómo afecta cada una de ellas al resultado final. En la tabla 5.3 se muestran los resultados de evaluar la red tras distintos tipos de entrenamiento que implican una o varias fases de las expuestas anteriormente, aplicando la fase de consistencia geométrica y la escala de grises como procesamiento de la imagen en todos ellos.

Modelo	Reentreno fase venas	Reentreno fase colon	Aumentación de datos	Recall (%) @precision 1	Emparejamientos necesarios
SuperPoint				54.00	16
SuperPoint		✓		47.50	18
SuperPoint		✓	✓	60.60	13
SuperPoint	✓	✓		57.96	16
SuperPoint	✓	✓	✓	68.52	15
SuperPoint	✓	✓	✓✓	**70.69**	**15**

Tabla 5.3: Evaluación de los distintos entrenamientos

Uno de los aspectos que resaltan es cómo las tres fases del entrenamiento contribuyen en gran medida a la obtención del mejor resultado. Es especialmente notable el efecto de la aumentación de datos, que logra situar los modelos que la incluyen como los que proporcionan los mejores resultados. Se aprecia esta diferencia entre los dos últimos modelos, en los que al aumentar tanto el número de muestras empleadas durante el entrenamiento, como el rango de variación aplicado a los efectos utilizados en ellas, un 68.52 % de *recall* se convierte en un 70.69 %. Otro elemento a resaltar es la influencia de la inclusión de las imágenes sintéticas generadas en la primera fase del entrenamiento, que proporcionan modelos que ofrecen mejores resultados.

El entrenamiento completo obtiene la puntuación más alta de entre todos los modelos. Con un umbral de 13 emparejamientos el modelo alcanza una exactitud del 85.34 %. La matriz de confusión del modelo puede observarse en la figura 5.3 junto a la curva *precision-recall* calculada.



(a) Matriz de confusión del mejor modelo (b) Curva *precision–recall* del mejor modelo

Figura 5.3: Resultados obtenidos con el mejor modelo con un umbral de 13 emparejamientos para declarar un reconocimiento de lugares positivo.

El análisis anterior permite determinar el mejor modelo de acuerdo con su capacidad de reconocimiento de lugares en el colon. Al considerar su introducción en un sistema mayor, capaz de llevar a cabo el proceso completo de SLAM visual, ha de tenerse también en cuenta el tiempo necesario para llevar a cabo el proceso. Debido a esto, se han desgranado ambos procesos obteniendo el tiempo necesario para cada fase[1], como puede observarse en la tabla 5.4. Además, se dividen las mediciones en aquellas que calculan la máscara de herramientas en ejecución, y aquellas que la precalculan, debido a que este proceso llega a consumir una gran parte del tiempo total de cada uno de los métodos, impidiendo la apreciación de las diferencias entre modelos.

Modelo	Máscara herramientas precalculada	Procesar imagen	Obtener máscara	Extracción + aplicar máscara	Matches putativos	Consistencia geométrica	Total (ms)	FPS
SIFT		37.40	321.75	182.90	6.93	4.47	553.45	1.81
ORB		38.17	320.53	40.47	1.33	3.49	403.99	2.48
AKAZE		38.16	316.30	100.70	2.45	15.06	472.69	2.12
SuperPoint		42.38	395.22	61.06	27.43	33.58	559.67	1.79
SIFT	✓	37.37	8.93	178.74	6.92	4.53	236.49	4.23
ORB	✓	37.35	8.97	39.52	1.31	3.55	90.70	11.03
AKAZE	✓	38.16	9.04	99.17	2.45	15.31	164.13	6.09
SuperPoint	✓	41.94	52.29	58.22	26.58	32.79	211.82	4.72

Tabla 5.4: Tiempo medio de los distintos modelos en cada una de las fases (ms)

Ninguno de los modelos desarrollados es capaz de actuar a la frecuencia de obtención de los fotogramas. El más cercano a ellos es ORB, que es también el que proporciona peores resultados. El modelo de SuperPoint se sitúa algo por detrás de AKAZE, la característica artesanal que mejor ha funcionado.

Con estos resultados a la vista se puede concluir que la combinación de SuperPoint y SuperGlue, es capaz de funcionar mejor que los métodos que emplean características artesanales en secuencias de endoscopia una vez se ha llevado a cabo un entrenamiento adecuado para estas. Además, todas las fases del mismo propuestas contribuyen a este resultado. Por último, se destaca que no hay una gran diferencia entre el tiempo requerido por este modelo y AKAZE.

[1]Resultados obtenidos en un computador con procesador Intel(R) Core(TM) i7-10700K CPU 3.80GHz y GPU GeForce RTX 2080 Ti.

5.4. Evaluación del método para reconocimiento de lugares en SLAM visual

En los apartados anteriores se ha analizado el comportamiento del método propuesto utilizando pares de imágenes sueltas. En este apartado vamos a analizar su potencial integración en un sistema de SLAM visual. Para ello, suponemos que tenemos un mapa de una sección del colon compuesto por un conjunto de imágenes clave (*keyframes*), y queremos localizar dentro de ese mapa la secuencia de imágenes que nos están llegando actualmente del endoscopio. El objetivo es determinar si el sistema sería capaz de llevar a cabo el reconocimiento de lugares correctamente. Un resultado positivo permitiría confirmar su validez como una alternativa apropiada para reconocimientos de lugares en entornos deformables.

Con el objetivo de llevar a cabo el experimento, los miembros del proyecto EndoMapper han proporcionado dos conjuntos de fotogramas de una misma secuencia de endoscopia con los cuales han conseguido reconstruir mapas del entorno. La unión de estos se corresponde con una parte de la operación de retirada del endoscopio. Las imágenes del primer grupo actúan como *keyframes* empleados para formar el mapa, mientras que las del segundo actúan a modo de secuencia recibida del endoscopio. Además, se conoce que las imágenes que actúan como secuencia comienzan en el mismo lugar en el que termina la construcción del mapa. Un subconjunto de ambos grupos puede observarse en el Anexo B.

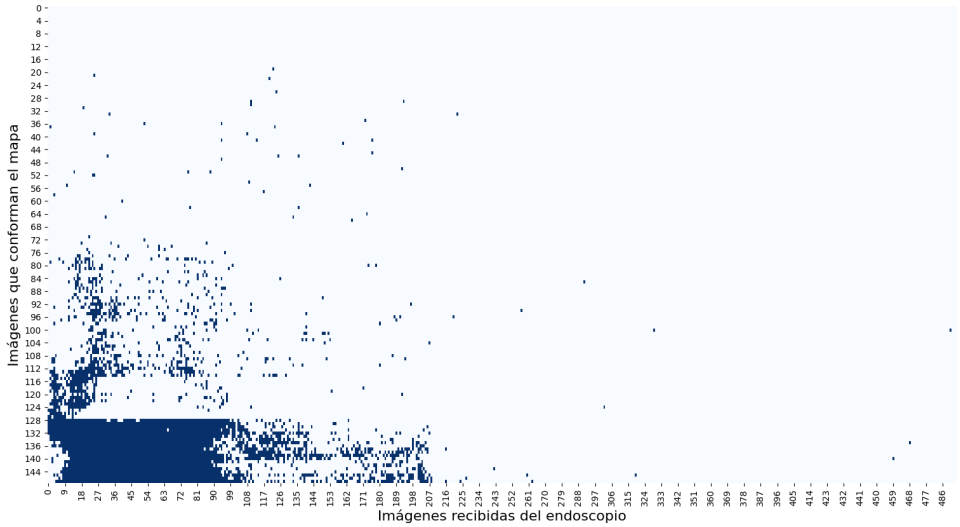
Anteriormente se ha obtenido un umbral de 13 puntos de interés emparejados para declarar que el lugar ha sido reconocido. Para afianzar la seguridad en estos, se va a optar por buscar una consistencia temporal entre cada una de las imágenes recibidas de la secuencia y las empleadas para simular el mapa. Para ello, se declarará un emparejamiento entre una imagen de la secuencia y otra del mapa si y solo si se cumple una condición adicional. Esta consiste en que existan al menos otros dos *keyframes* con los que también se supere el umbral de 13 emparejamientos. Además, estos deben situarse a una distancia de menos de 30 imágenes del *keyframe* del par de imágenes evaluado.

La implementación del experimento se basa en tratar de emparejar las 148 imágenes de la primera secuencia con cada una de las 494 de la segunda, haciendo un total de 73112 emparejamientos potenciales. El modelo empleado para este se corresponde con el mejor de los obtenidos en el apartado anterior, es decir, SuperPoint reentrenado, aplicando la fase de consistencia geométrica y con las imágenes en escala de grises.

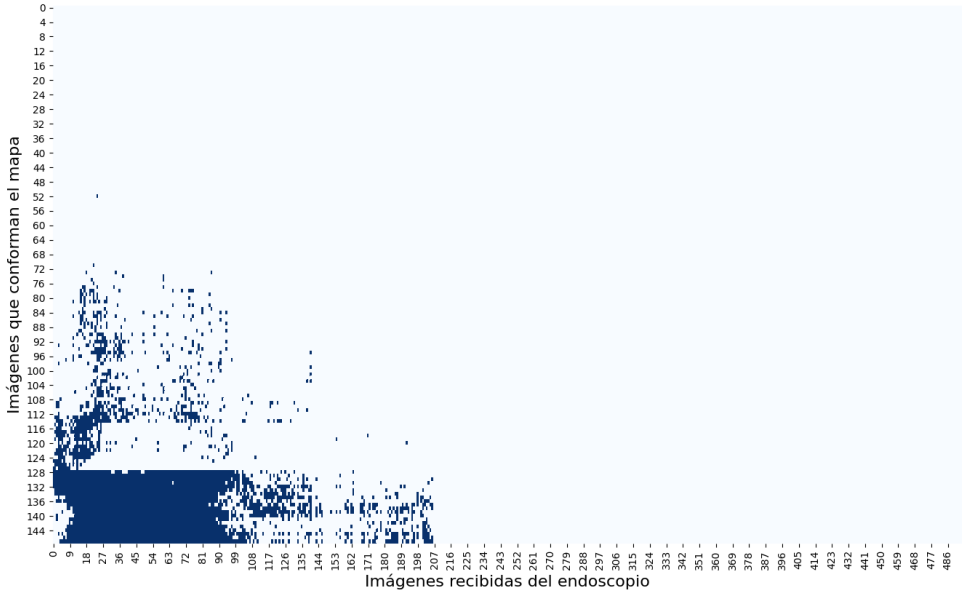
No obstante, en este caso se ha prescindido del cálculo de la máscara de herramientas, dado que estas no aparecen en la secuencia. El resultado obtenido tanto con consistencia temporal como sin ella se muestra en la figura 5.4.

En los resultados obtenidos se observa cómo los pocos emparejamientos dados con fotogramas de la secuencia más allá del 210 son descartados por la restricción de consistencia temporal. Es en torno a estas imágenes en las que se empieza un giro del endoscopio en su retirada, dejando de ser visible la zona inicial. Cumple por tanto con el resultado deseado en este aspecto. El efecto de esta restricción puede observarse más claramente con el ejemplo de la figura 5.5.

Este experimento refleja más exhaustivamente el comportamiento del sistema en zonas distintas pero muy cercanas en tiempo, espacio y apariencia. En ellas, la unión de aplicar el umbral obtenido en la sección anterior y la búsqueda de consistencia temporal ha proporcionado buenos resultados. Por consiguiente, el sistema creado es capaz de llevar a cabo el reconocimiento de lugares en los entornos deformables del interior del colon.



(a) Pares aceptados como un reconocimiento de lugares positivo sin aplicar la restricción de consistencia temporal.



(b) Pares aceptados como un reconocimiento de lugares positivo aplicando la restricción de consistencia temporal.

Figura 5.4: Representación de los pares de imágenes en los que se ha llevado a cabo el reconocimiento de lugares. En azul se muestran aquellos pares que han sido declarados como un reconocimiento de lugares positivo. En el caso de la primera imagen, estos pares son aquellos con más de 13 emparejamientos. En el caso de la segunda imagen, los pares de imágenes también deben superar la restricción de consistencia temporal, es decir, el fotograma de la secuencia de ese par debe ser capaz de llevar a cabo el reconocimiento de lugares con al menos otros dos *keyframes* a una distancia de menos de 30 imágenes del *keyframe* del par evaluado.

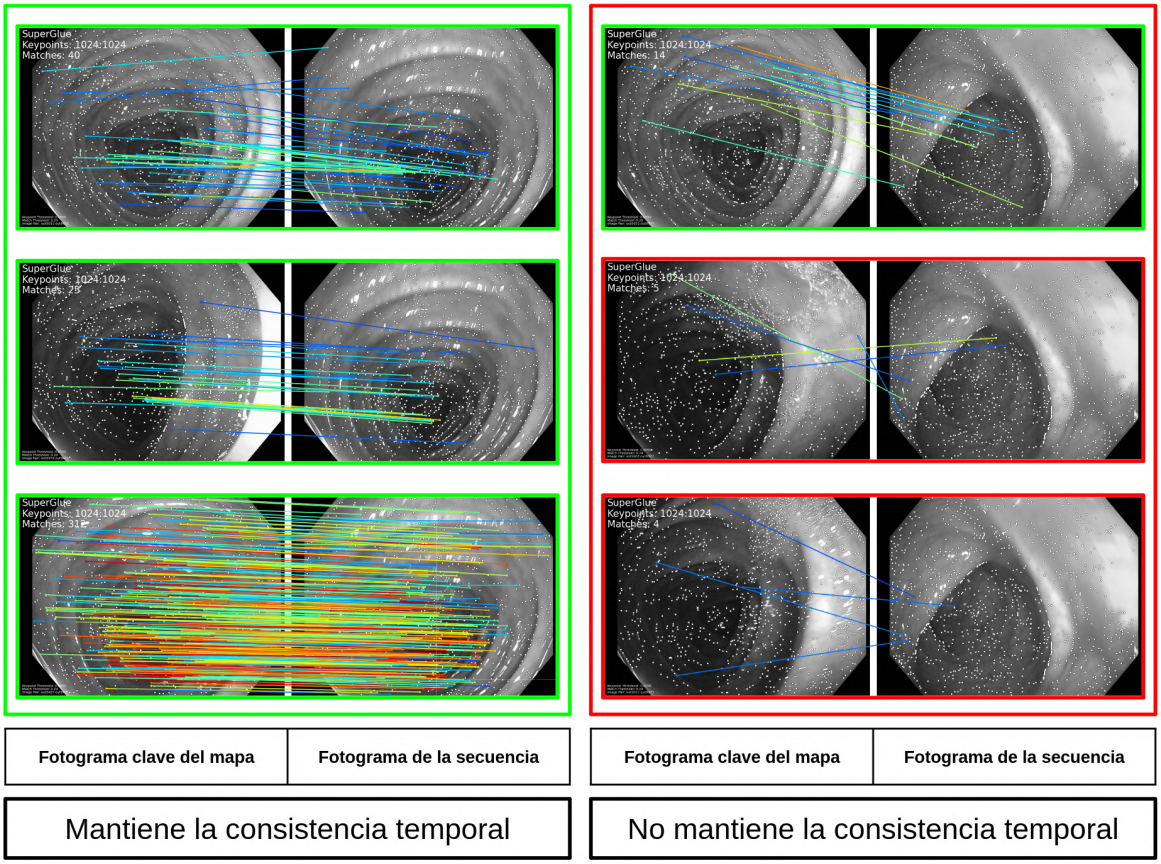


Figura 5.5: En la columna de la izquierda se muestra el fotograma 32 de la secuencia junto a 3 *keyframes* del mapa con los que supera la restricción de consistencia temporal. En la columna derecha se emplea el fotograma 491 de la secuencia, que únicamente supera el umbral con un *keyframe*, no superando la restricción de consistencia temporal y siendo rechazado el reconocimiento de lugares. Aquellos pares con al menos 13 emparejamientos son representados con un marco verde. El resto se representan con un marco rojo.

Capítulo 6

Conclusiones

El reconocimiento de lugares en entornos deformables es un campo en el que todavía no se había investigado. Sin embargo, se ha logrado desarrollar un método capaz de obtener buenos resultados en el interior del colon, un entorno de estas características. Para ello, ha sido necesario el estudio previo de las técnicas existentes concebidas para entornos rígidos, que han servido de punto de inicio para el sistema creado.

Durante el proyecto, la división entre características artesanales, y aquellas basadas en redes neuronales convolucionales ha permitido comparar el rendimiento de métodos basados en cada uno de los dos grupos en conjuntos de datos de endoscopia. Han demostrado una capacidad similar en ausencia de un entrenamiento específico de las características basadas en redes profundas. Al llevarlo a cabo son capaces de diferenciarse del resto, alcanzando un 70 % de *recall* manteniendo la *precision* al 100 % en la evaluación. Puede concluirse por tanto que, con un reentrenamiento adecuado, pueden obtener un rendimiento muy superior en este tipo de secuencias. Dentro del método desarrollado destacan la necesidad de enmascarar ciertos elementos en secuencias de endoscopia, así como la inclusión de una fase de consistencia geométrica.

El procedimiento expuesto se ha mostrado capaz de realizar su función en sistemas de SLAM visual que actúan en el interior del cuerpo humano, abriendo la puerta a su inclusión en sistemas de este tipo.

6.1. Trabajo futuro

El sistema desarrollado se encuentra totalmente desacoplado de cualquier sistema de SLAM visual. Esto permite evaluar su comportamiento como módulo independiente, pero no como parte del conjunto completo. El paso lógico sería su integración en un sistema de SLAM visual. Para ello, sería imprescindible estudiar cuál es la forma óptima de integrarlo, estudiando posibles formas de inclusión, además de las limitaciones temporales y de capacidad de cómputo impuestas por el sistema. Estas se corresponderían con los límites en los que puede actuar el método desarrollado, determinando cuántos pares de fotogramas pueden ser tratados cada segundo. Por ende, cobraría una gran importancia la reducción del tiempo de ejecución requerido, pudiendo ser necesaria la optimización del código. No obstante, estos aspectos quedan fuera del alcance de este trabajo de fin de grado, quedando como posibilidades que podrían llegar a estudiarse en el futuro.

Capítulo 7

Bibliografía

[1] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[2] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*. IEEE, November 2011.

[3] Pablo Alcantarilla, Jesus Nuevo, and Adrien Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *Procedings of the British Machine Vision Conference 2013*.

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[7] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison,

Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[9] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018.

[10] Juho Kannala and Sami S Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340, 2006.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Lista de Figuras

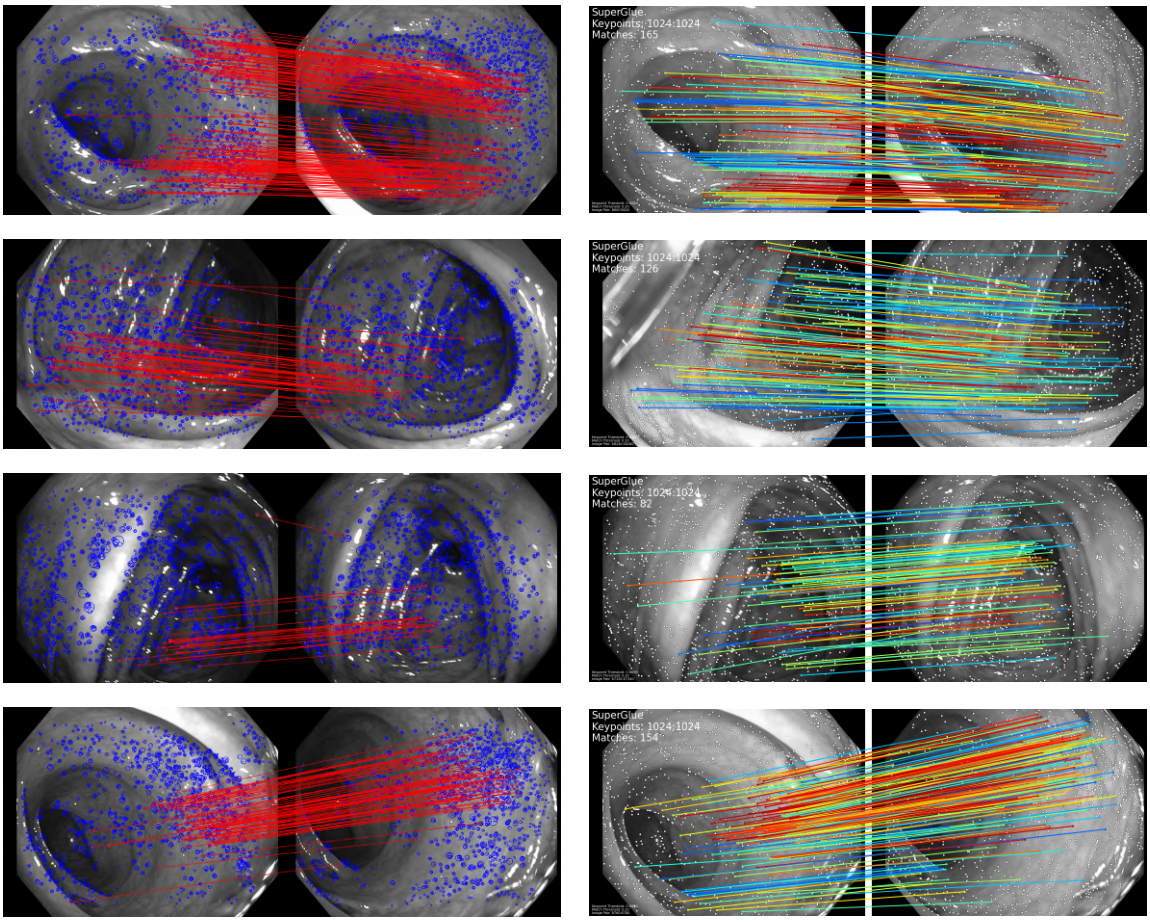
Lista de Tablas

Anexos

Anexos A

Ejemplos de reconocimiento de lugares en endoscopias

En la figura A.1 se muestran los emparejamientos obtenidos al llevar a cabo el reconocimiento de lugares en endoscopia sobre distintos pares de imágenes con los dos métodos desarrollados.



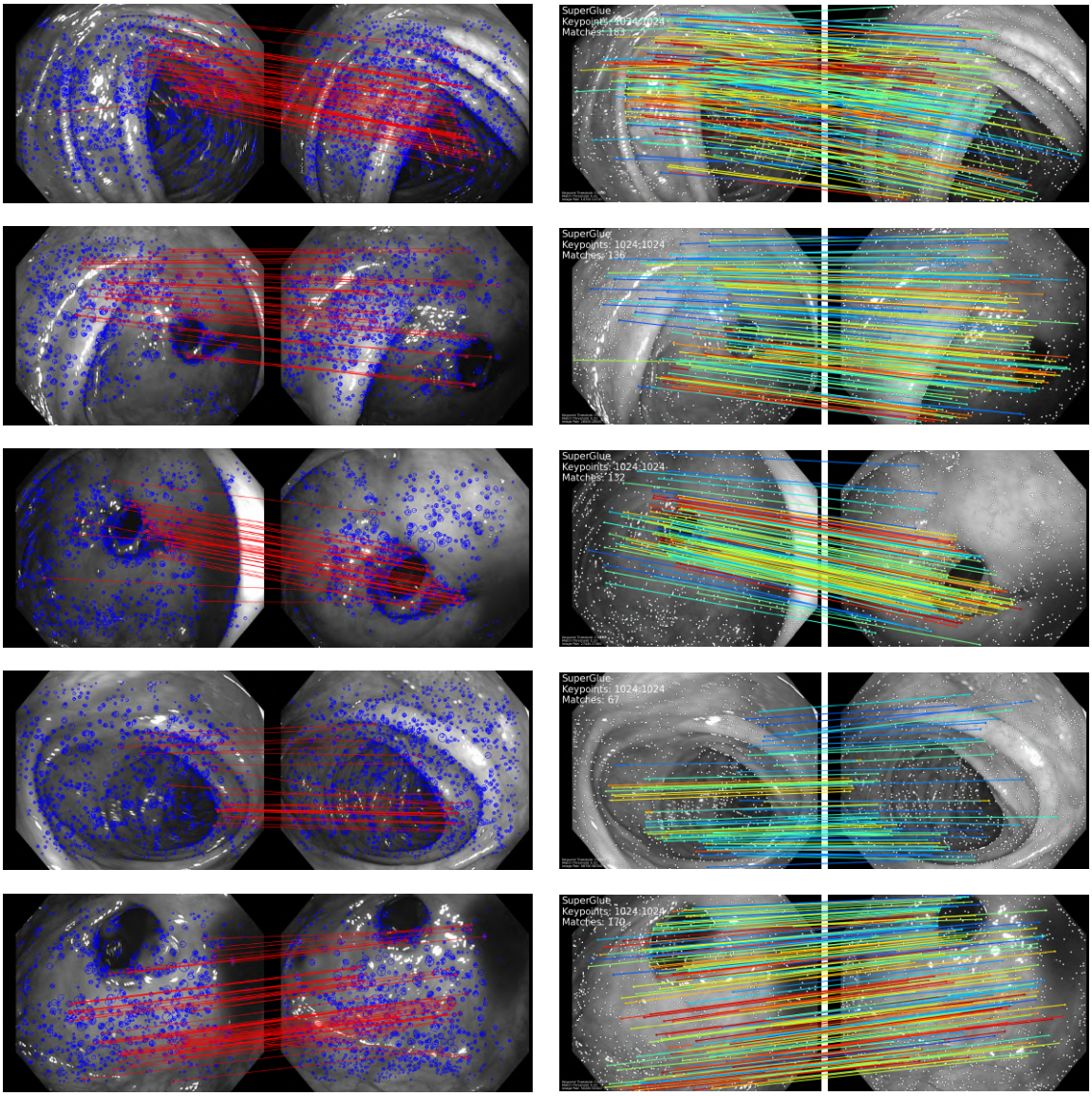


Figura A.1: Ejemplos de reconocimiento de lugares. En la columna izquierda se muestran los emparejamientos obtenidos mediante el procedimiento basado en características artesanales para distintos pares de imágenes, empleando puntos de interés AKAZE. En la columna de la derecha aparecen los emparejamientos para los mismos pares obtenidos mediante el método que emplea características basadas en CNN en su versión reentrenada. En esta segunda columna, los colores de los emparejamientos representan la confianza predicha por SuperGlue. Rojo indica mayor confianza, mientras que azul indica una confianza menor. Todos los pares de imágenes pertenecen al conjunto de datos de test.

Se adjunta a continuación la representación de los emparejamientos obtenidos con el método que utiliza características artesanales para los 4 primeros pares de la figura A.1, empleando tanto características SIFT como ORB.

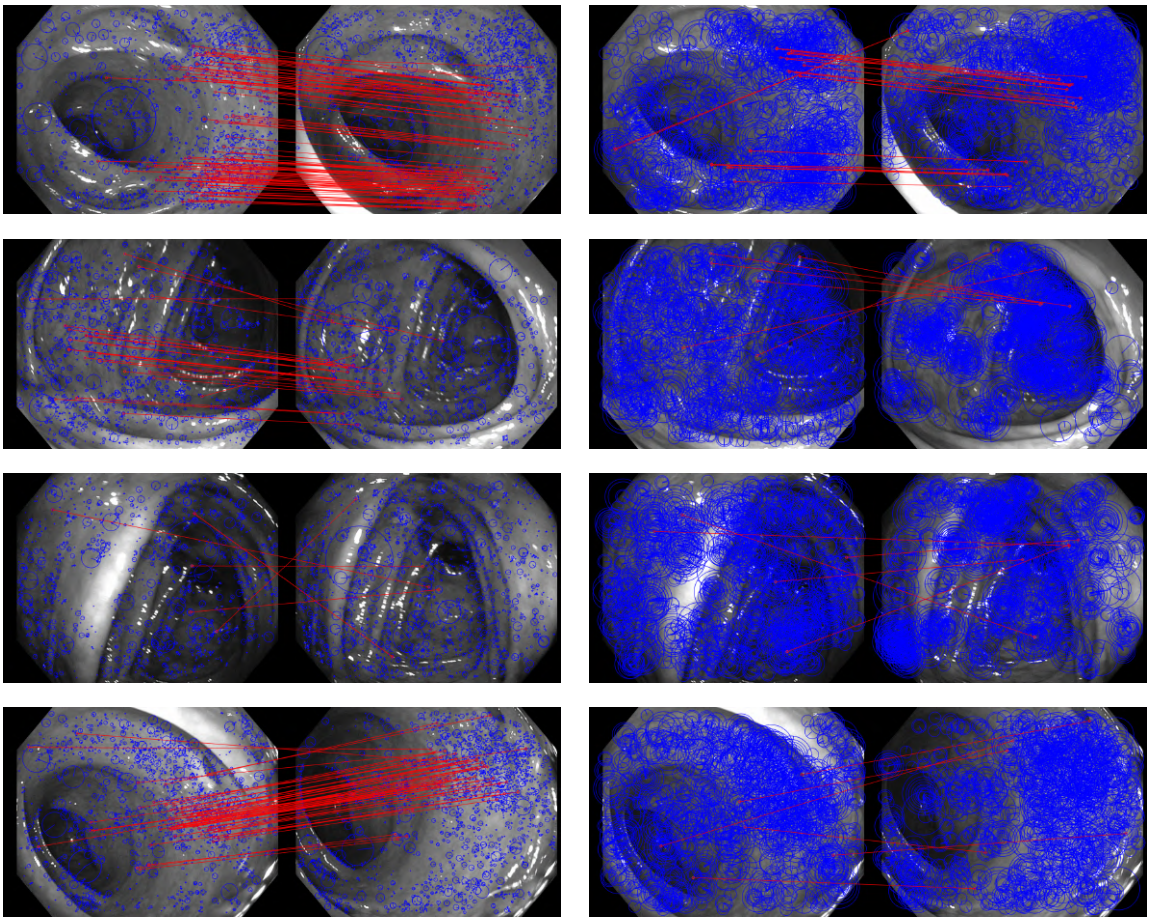


Figura A.2: Ejemplos de reconocimiento de lugares con el método que utiliza características artesanales. En la columna izquierda se muestran los emparejamientos obtenidos empleando SIFT, mientras que aquellos obtenidos con ORB aparecen en la derecha.

Anexos B

Secuencias para la evaluación del reconocimiento de lugares en SLAM visual

En la figura B.1 se muestran fotogramas del conjunto de imágenes que actúa como mapa durante la evaluación descrita en la sección 5.4. Por otro lado, en la figura B.2 se muestran fotogramas que forman parte del conjunto utilizado a modo de secuencia recibida desde el endoscopio.

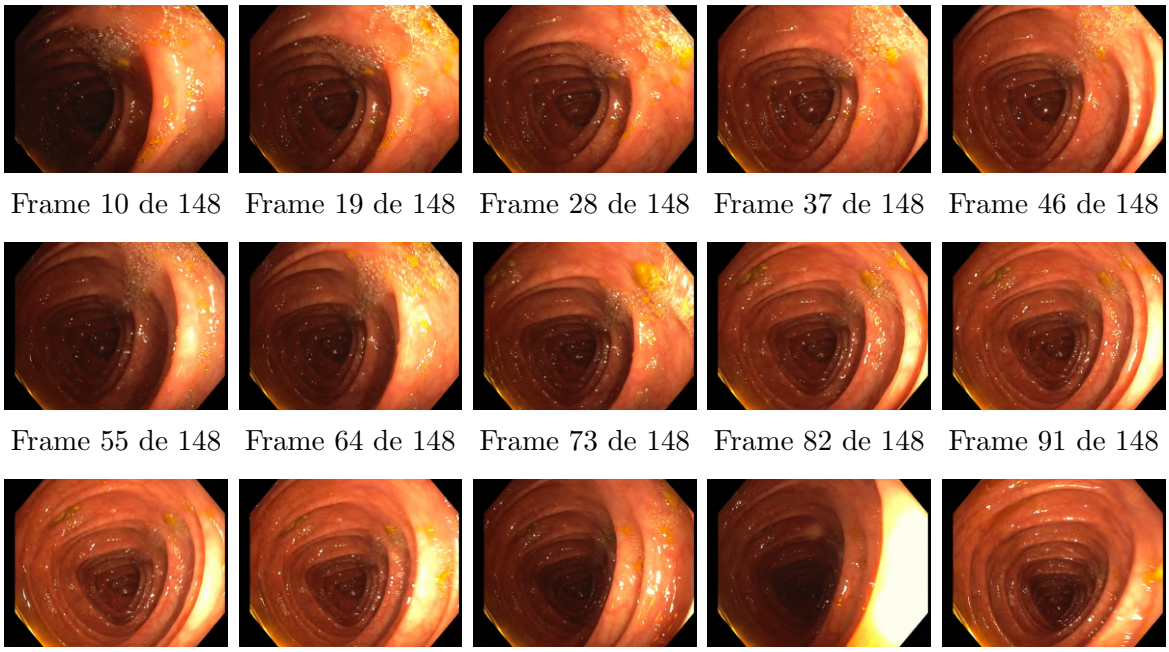
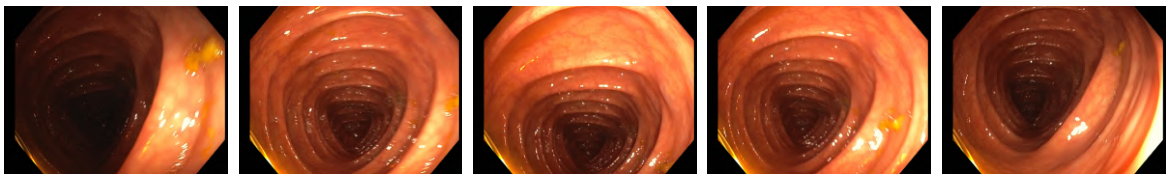
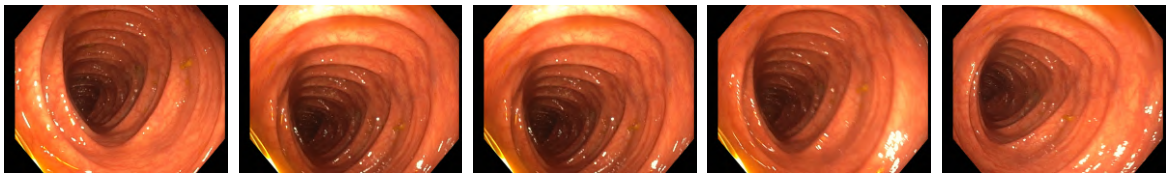


Figura B.1: Fotogramas del conjunto que actúa como mapa del SLAM visual.



Frame 1 de 494 Frame 26 de 494 Frame 51 de 494 Frame 76 de 494 Frame 101 de 494



Frame 126 de 494 Frame 151 de 494 Frame 176 de 494 Frame 201 de 494 Frame 226 de 494

Frame 251 de 494 Frame 276 de 494 Frame 301 de 494 Frame 326 de 494 Frame 351 de 494

Frame 376 de 494 Frame 401 de 494 Frame 426 de 494 Frame 451 de 494 Frame 476 de 494

Figura B.2: Fotogramas del conjunto que actúa como secuencia recibida del endoscopio

Anexos C

Gestión del proyecto

Este trabajo se basa en la investigación realizada con una Beca de Colaboración del Ministerio de Educación y Formación Profesional, dentro del Departamento de Informática e Ingeniería de Sistemas de la Universidad de Zaragoza. La dedicación diaria aproximada ha sido de 3 horas, haciendo un total de 15 horas semanales.

El proyecto se ha llevado a cabo durante el periodo comprendido entre la segunda quincena de septiembre de 2020 y junio del 2021. La carga de trabajo, que puede observarse fragmentada en tareas en la tabla C.1, asciende a un total de 481 horas.

Tarea	*Horas dedicadas*
Configuración del entorno	35 h
Estudio previo	30 h
Obtención de los datos de evaluación	20 h
Desarrollo del método con características artesanales	45 h
Desarrollo del método con características basadas en CNN	45 h
Reentrenamiento de SuperPoint	140 h
Obtención de resultados	50 h
Reconocimiento de lugares en SLAM visual	20 h
Memoria	60 h
Reuniones	36 h
Total:	481 h

Tabla C.1: Horas totales por tarea dedicadas al proyecto.

En la figura C.1 se muestra el diagrama de Gantt que refleja a qué tareas ha sido dedicada cada semana durante la duración del proyecto. En el diagrama, la configuración del entorno se encuentra dividida debido a la habilitación de un nuevo entorno de trabajo con el fin de reentrenar la red neuronal SuperPoint.

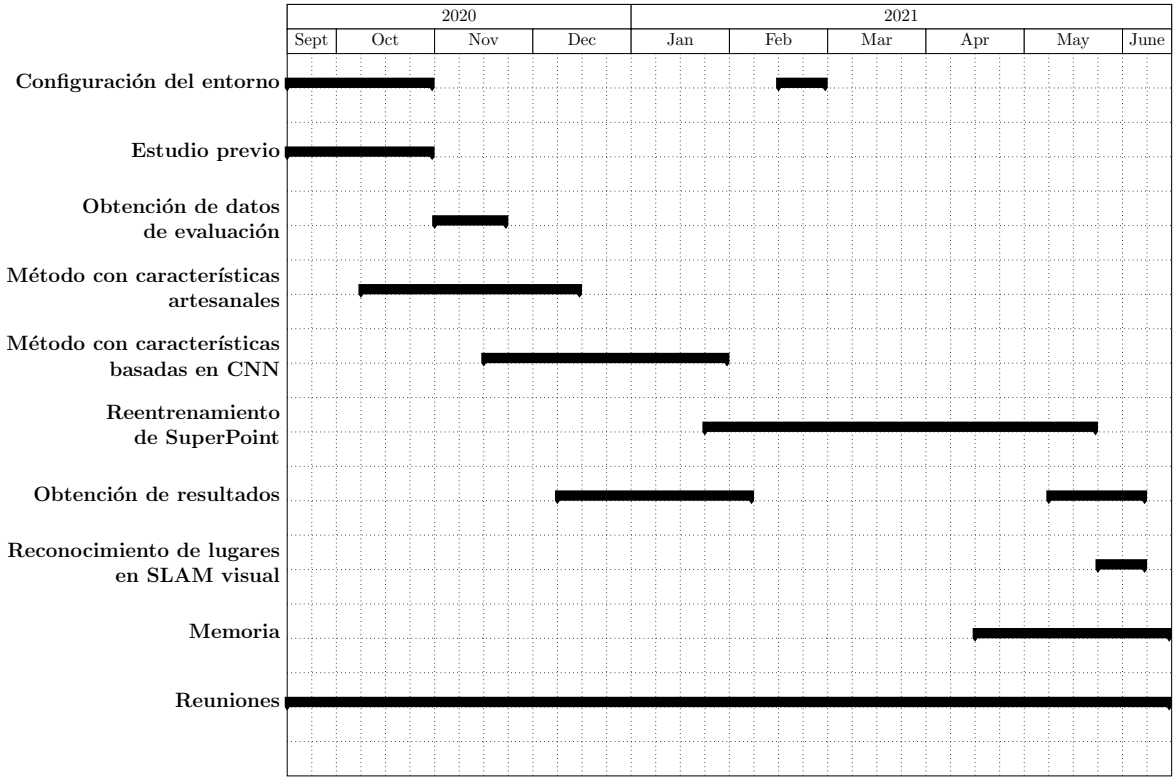


Figura C.1: Diagrama de Gantt del proyecto.