Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza
1542

Segmentación automática de vídeos de endoscopias

Cristina Oriol García

Directores:
Ana Cristina Murillo Arnal
Íñigo Alonso Ruiz

Trabajo fin de Grado
Grado en Ingeniería Informática
Computación

Departamento de Informática e Ingeniería de Sistemas
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

Junio 2021

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe entregarse en la Secretaría de la EINA, dentro del plazo de depósito del TFG/TFM para su evaluación).

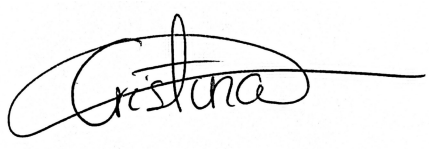
D./Dª. Cristina Oriol García ,en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) Grado (Título del Trabajo)

Segmentación automática de vídeos de endoscopias

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, 22 de Junio de 2021



Fdo: Cristina Oriol García

Resumen

La automatización de ciertos procesos médicos es un tema muy estudiado actualmente, por las muchas posibilidades y beneficios que suponen a la hora de detectar enfermedades de riesgo o de mejorar distintos tratamientos y operaciones. La realización de **endoscopias** es un proceso que ya incluye cierto uso de tecnología por medio de los endoscopios. Añadir un mayor nivel de automatización en estos procedimientos sería muy útil para reducir riesgos o mejorar el diagnóstico. Para ello es necesario analizar estos procedimientos, creando un conjunto de datos sobre endoscopias reales que se puedan estudiar y procesar de manera sistemática.

Recopilar **grandes cantidades de datos reales** supone distintos retos por sus características, no solo por la dificultad de procesar dichas imágenes, la principal **problemática** se debe a que al tratarse de imágenes grabadas en el interior del cuerpo humano a través del endoscopio, aparecen múltiples escenas que no aportan información útil y empeoran la calidad de los procesados, como imágenes con agua, otras en las que se ve borroso ya que el endoscopio ha chocado con las paredes de los órganos. Por ejemplo técnicas de reconstrucción 3D o de algún tipo de reconocimiento (lesiones, tumores...) se verían beneficiadas de un pre-procesamiento automático de los vídeos que anoten información útil sobre los órganos, como la sección en la que se encuentra.

Así pues, el objetivo principal de este proyecto es **mejorar la calidad** y **descubrir de forma automática información útil** en estas grabaciones para solucionar los problemas mencionados. Dada la tendencia de los últimos años del aumento de las técnicas de aprendizaje automático para clasificación de imágenes, se ha decidido utilizar uno de estos modelos sobre nuestro conjunto de datos para demostrar su utilidad a la hora de separar distintos tipos de imágenes dentro de este contexto y proporcionar información. Para conseguir estos objetivos, en este proyecto se han desarrollado dos partes principales.

Se ha desarrollado un **sistema** de descripción de imágenes automático utilizado para **detectar el inicio y el final** del procedimiento real de la **endoscopia**. Como resultado, se ha conseguido separar esta parte relevante y descartar los datos inútiles de la secuencia original. Esto resulta esencial para no almacenar partes no necesarias ya que ocupan una cantidad de espacio importante y su uso resulta más sencillo.

Por otra parte, se ha creado otro **sistema** que permite **detectar diferentes patrones y características de las imágenes de la endoscopia**. Para esto, se ha entrenado un modelo de *deep learning* de forma no supervisada con técnicas de aprendizaje por contraste. Con los resultados obtenidos se comprobó que el modelo es capaz de separar las imágenes que no pertenecen a las endoscopias, pudiendo utilizarse para realizar también el trabajo del primer prototipo. Pero lo más interesante, además es capaz de separar muchos eventos relevantes que ocurren a lo largo de toda la grabación, siendo útil para otras tareas como el etiquetado en secciones o la eliminación de imágenes no informativas.

Este trabajo se desarrolla dentro del ámbito del proyecto europeo EndoMapper, el cual está realizando numerosas grabaciones en colaboración con médicos del hospital Clínico Universitario Lozano Blesa. El módulo de grabación con recorte implementado ya está instalado y funcionando en los sistemas de grabación reales de este proyecto, mientras que el módulo de segmentación automática está siendo evaluado con los datos reales del proyecto recogidos actualmente.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Este capítulo motiva y explica el proyecto así como su contexto y los distintos objetivos del mismo.

1.1. Motivación.

En el ámbito de los procedimientos médicos, encontramos que muchos de ellos pueden ser automatizados, de forma parcial o total. Algunos como la monitorización de parámetros biológicos o la automatización de los sistemas de control, aunque también procesos mucho más complejos. Otros, realizados con la ayuda de distintos sistemas y dispositivos, mediante algún tipo de software pueden ofrecer más información durante la realización del procedimiento.

Este proyecto se ha enfocado en la automatización del procesado de las **endoscopias**. En este campo ya nos encontramos la utilización de un instrumento, el endoscopio, que supone un ejemplo de que la utilización de nuevas tecnologías puede suponer una mejora muy considerable en el campo médico, ya que sin él todas las enfermedades que se detectan a día de hoy en estas pruebas no serían diagnosticables. La Figura 1.1 muestra un ejemplo del endoscopio utilizado para realizar estos procedimientos.

Figura 1.1: Ejemplo de endoscopio.

La grabación de grandes bases de datos endoscopias puede ser de gran utilidad mediante su análisis y aplicación en distintas técnicas para conseguir au-

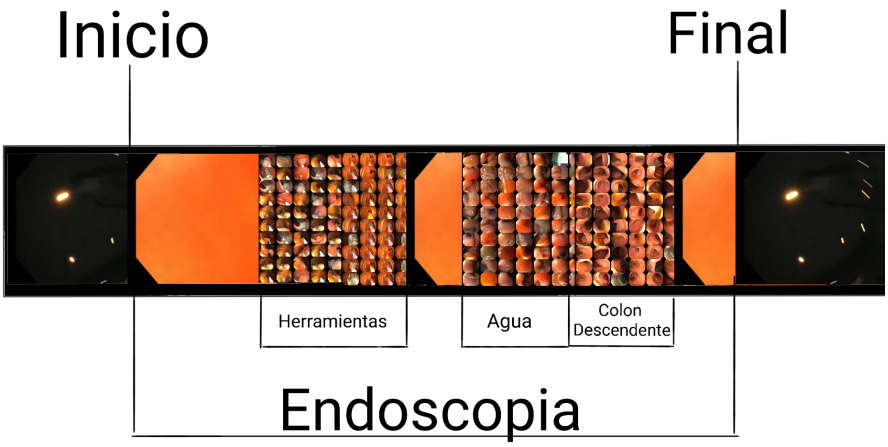


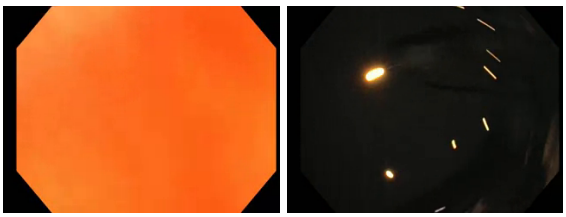
Figura 1.2: Ejemplos de distintos tipo de imágenes que aparecen en una endoscopia y su detección en secciones. Antes y después del procedimiento en si se ven imágenes de la sala del hospital en la que se realiza el procedimiento. Durante este se observan zonas con agua, zonas con herramientas y una sección que corresponde al colón descendiente.

tomatizar estos procedimientos y hacerlos más precisos y seguros. Por ejemplo, para ayudar a desarrollar y evaluar nuevos métodos de reconstrucción 3D o sistemas de reconocimiento de patrones de interés. Los resultado obtenidos pueden suponer descubrimientos que permitan facilitar al personal sanitario el diagnóstico de enfermedades que suponen un riesgo en la salud de sus pacientes, acelerando su detección y tratamiento.

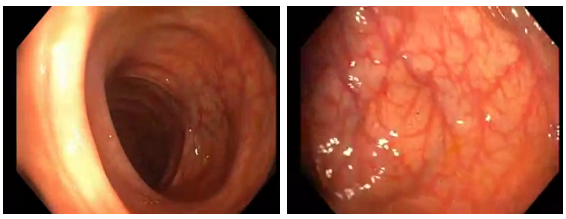
Cualquier procedimiento automático que se quisiera aplicar sobre estas grabaciones reales, necesita algún tipo de pre-procesado automático que analice la información de la que se dispone en el vídeo. En particular, resulta esencial filtrar las zonas del vídeo que no resulten de interés y no sea necesario analizar, o por el contrario secciones concretas en las que ocurre algún tipo de evento interesante. Como se puede observar en la Figura 1.2, donde se muestra un ejemplo de distintas zonas más o menos interesantes detectadas en una misma grabación.

En los vídeos de endoscopia, es frecuente encontrar imágenes en negro, consideradas vacías, o que no aportan ninguna información útil, como imágenes en la que se graba la sala en la que se lleva a cabo el procedimiento. Estas imágenes pueden dificultar el almacenaje o procesado de grabaciones automáticas o llegar a entorpecer de manera significativa las técnicas posteriormente aplicadas para procesar la endoscopia, por ejemplo reconstrucción en 3D, detección de pólipos, etc. La figura 1.3 muestra un ejemplo de algunos de estos *frames* que consideraríamos no útiles en contraposición de aquellos que serían informativos.

En este proyecto se trabaja hacia un sistema de segmentación que permita pre-procesar vídeos reales, completos, grabados en procedimientos médicos ru-



no-informativas



informativas

Figura 1.3: Ejemplos de las imágenes separadas en informativas y no informativas.

tinarios, para anotarlos con la mayor cantidad de meta-datos posible de manera automática.

1.2. Contexto y punto de partida

Este proyecto se ha desarrollado en el grupo de investigación de Robótica, Percepción y Tiempo Real (RoPeRT), en contexto del proyecto **EndoMapper**[4]. El objetivo de este proyecto es establecer los fundamentos para la navegación autónoma y construcción de mapas 3D del interior del cuerpo humano a partir de imágenes de endoscopia médica como se resume visualmente en la Figura 1.4.

La motivación de este proyecto se debe al tipo de datos que se poseen, colonoscopias y gastroscopias, en crudo, tal cual se graban de manera automatizada 24 horas al día en un ordenador instalado en el Hospital Clínico Lozano Blesa de la Universidad de Zaragoza, como parte del proyecto EndoMapper. El reto que suponen estas imágenes es que al pertenecer a grabaciones del interior del cuerpo humano resulta más difícil encontrar características que las diferencien entre ellas. Además hay que notar que no existe ningún mecanismo que indique cuando comenzaba el procedimiento o cuando termina. Debido a esto, las grabaciones generalmente son de una larga duración de la cual, realmente, solía corresponder una pequeña parte a la endoscopia, siendo la mayor parte del vídeo resultante escenas de la sala, sin ningún tipo de interés para este proyecto. Debido a esto a lo hora de realizar técnicas más manuales implica necesitar un mayor análisis de la información que nos aportan de cada imagen, para así decidir cuándo existe una diferencia importante entre varias.

A continuación se describen los objetivos detallados de este proyecto para mejorar este proceso de grabación, pre-procesado y anotación automatizada de datos.

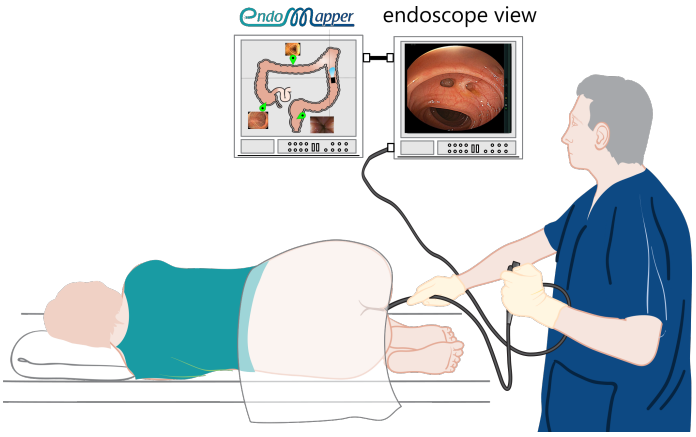


Figura 1.4: Explicación visual del objetivo del proyecto EndoMapper.
Fuente imagen: JMM Montiel, license CC BY-SA 4.0. Figure generated merging next figures: [1] By Cancer Research UK - Original email from CRUK CC BY-SA 4.0 Link ; [2] By MAC 06 - Own work CC BY 4.0 Link ; [3] By melvil - Own work CC BY-SA 4.0 Link ; [4] By Joachim Guntau (=J.Guntau) - Endoskopiebilder.de CC BY-SA 3.0 Link ; [5] By Joachim Guntau (=J.Guntau) - Endoskopiebilder.de CC BY-SA 3.0 Link

1.3. Objetivos y tareas

El principal objetivo de este proyecto consiste en realizar un prototipo de software que nos permita segmentar de la manera más informativa posible las distintas grabaciones, tanto las ya obtenidas dentro de nuestro conjunto de datos, como las nuevas grabaciones que se vayan grabando.

En particular, uno de los resultados que se busca en esta segmentación es eliminar las partes del vídeo que no correspondan al proceso de la endoscopia, como se visualiza en la Figura 1.2. Con esta información conseguiríamos reducir mucho el tamaño de cada vídeo, que es de aproximadamente 60 giga-bytes cuando ya han sido transformados a vídeo y unos 200 cuando todavía son imágenes. Reducir este tamaño supondría una gran ventaja de cara a su almacenamiento. Actualmente esto presentaba numerosos problemas derivados del gran tamaño de cada vídeo.

Otro objetivo es conseguir eliminar las partes no informativas de las grabaciones, imágenes en negro, no pertenecientes al procedimiento, aquellas en las que encontramos herramientas, o algunas otras que por la aparición de agua no se aprecia la zona. En general imágenes que no queremos analizar ya que no tienen utilidad o información útil o necesaria. Para ello, el objetivo es utilizar técnicas de aprendizaje no supervisado para detectar tipos de zonas dentro de una misma grabación. Posteriormente el resultados de estas técnicas permitirán la separación en secciones, de distintas características, de la secuencia.

Tareas. Para alcanzar estos objetivos se han realizado las siguientes tareas:

- **Estudio de las bases teóricas.** Mediante el estudio y análisis de las técnicas de reconocimiento automático con deep learning y tratamiento de imágenes, comprender estas técnicas y saber emplearlas

- **Estudio del manejo básico de las endoscopias.** Analizar el dataset del que disponemos así como comprender el funcionamiento del software utilizado en su grabación y el pre-procesamiento que se aplica a los vídeos obtenidos. Se dedicaron unas 40 horas a este estudio.
- **Diseño e implementación del módulo de grabación.** Desarrollo del módulo incluido en el código de grabación de las secuencias, mejorando el módulo original para superar las limitaciones y errores que presentaba. Se emplearon aproximadamente 50 horas a esta tarea.
- **Diseño e implementación del software del módulo de segmentación.** Desarrollo del software que permite anotar zonas de interés en el procedimiento médico en una fase de post-procesado y Composición de un dataset compuesto por nuestras grabaciones y entrenamiento de un modelo de aprendizaje no supervisado para la segmentación de imágenes buscando la mejora de la segmentación. Se emplearon aproximadamente 150 horas a esta tarea.
- **Evaluación y documentación.** Evaluar los resultados obtenidos en las distintas partes que conforman el proyecto aplicadas a endoscopias reales. Redacción de la memoria. Se dedicaron unas 60 horas a esta tarea.

Figura 1.5: Diagrama temporal que resume el reparto de las tareas a lo largo del proyecto

1.4. Contenido de la memoria

En esta memoria encontramos: en el capítulo 2 se explican las técnicas, ya existentes, utilizadas a lo largo del proyecto. En el capítulo 3 se detalla el sistema original del que se partía inicialmente y el resultado del sistema final realizado. En el capítulo 4 se explica el primer módulo implementado, de grabación y su evaluación. En el capítulo 5 se explica y evalúan los dos bloques pertenecientes al módulo de segmentación. Finalmente en el capítulo 6 se encuentran las conclusiones de este proyecto.

Capítulo 2

Técnicas de descripción de imagen estudiadas

Este capítulo resume los principales trabajos relacionados con las técnicas de visión por computador y *machine learning* aplicadas en los dos módulos principales desarrollados en este proyecto (grabación y segmentación).

2.1. Clasificación y reconocimiento en imágenes con descriptores tradicionales

Para la realización del **módulo de grabación** y del primer bloque del **módulo de segmentación** se han utilizado algunas técnicas muy conocidas, ya existentes, para tareas de reconocimiento visual en imágenes. Les podríamos llamar descriptores tradicionales, frente a los desarrollados más recientemente basados en técnicas de *deep learning* que se describen en la sección siguiente.

OCR. Una de las tareas de reconocimiento necesarias en este proyecto implica la detección de caracteres en imágenes, para ello se ha decidido utilizar las técnicas de OCR, Optical Character Recognition, mediante Tesseract, un OCR de código abierto [1]. Esta tecnología permite el reconocimiento de caracteres, tanto texto manuscrito como redactado a ordenador aunque su resultado depende directamente de la imagen de entrada. En la figura 2.1 se puede ver de forma visual la arquitectura de la herramienta Tesseract. Este OCR se utiliza en este proyecto debido a que el prototipo inicial del que se partía integra Tesseract, por lo que se mantuvo añadiendo una configuración específica de sus parámetros para mejorar su funcionamiento.

Descriptores de color y estructura. Para otras tareas, era necesario un descriptor capaz de capturar la estructura principal de la información en la escena de la imagen. Para ello se han considerado distintos descriptores que nos dan información sobre la estructura de la imagen, algunos de ellos **SIFT**[5], **ORB** [6] y **HOG** [2], todos ellos son utilizados par el reconocimiento de características sobre la estructura de la imagen para, principalmente, detección de objetos.

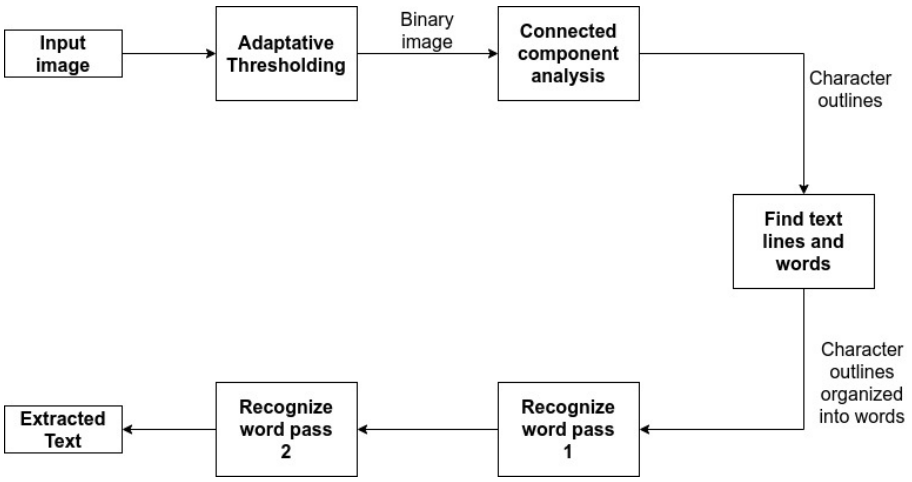


Figura 2.1: Arquitectura de Tesseract. Figura obtenida de [1]

Además se han considerado también otros descriptores que den información sobre los colores de la imágen, principalmente los **histogramas**.

En particular, se ha utilizado Histogram of Oriented Gradients (HOG). HOG es uno de los descriptores más populares y eficientes para la detección de objetos hasta la aparición de las técnicas de *deep learning*. Para su cálculo se realiza una normalización de la imagen, después se obtienen los gradientes en x e y de la imagen y sus correspondientes histogramas. Con los histogramas se lleva a cabo una normalización a través de bloques y finalmente se convierte a un vector de descriptores. En la Figura 2.2 se observa el resultado de aplicar el descriptor de HOG a una imagen. Se ha elegido este descriptor por su capacidad de detectar distintos objetos en base a los contornos que aparecen en una imagen para diferenciar aquellos que aparecen en una imagen que no en otra.

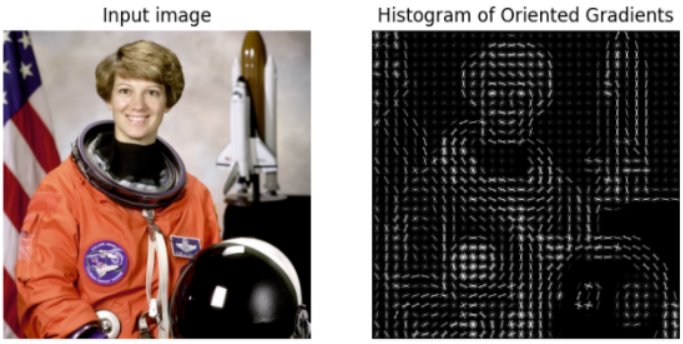


Figura 2.2: Resultado de la aplicación del descriptor de HOG a una imagen. Figura obtenida de [2]

2.2. Métodos no supervisados para aprender descriptores de imagen

En los últimos años la popularidad del *deep learning* ha crecido considerablemente debido a la multitud de campos en los que se puede aplicar, su funcionamiento está basado en el uso de redes neuronales artificiales en las que cada neurona recibe información la transforma y la envía. Actualmente se considera el mejor mecanismo para el clasificado de imágenes, permitiendo detectar e identificar todos los elementos de una imagen. Algunos de los modelos que se han tenido en cuenta son: Doersch et al[7]que funciona extrayendo diferentes trozos de cada imagen de forma desordenada y con ellos entra una red neuronal que adivine cuál es el orden original de los trozos. Gidaris et al[8] en este modelo se entrena una red para que adivine la rotación que se le ha aplicado a una imagen. También se han considerado otros dos modelos que utilizan para su funcionamiento *contrastive learning*, el primero BYOL[3], que solo utiliza ejemplos positivos y SimCLR[9] que utiliza ejemplos tanto positivos como negativos. Concretamente se ha elegido **BYOL** para su uso en el módulo de segmentación desarrollado, se ha decidido utilizar este modelo ya que tiene código abierto.

BYOL. El objetivo principal de **BYOL** consiste en dada una imagen aplicar dos trasformaciones o *image augmentations* a esta, buscando predecir la segunda mediante la primera. Para ello este método utiliza dos redes neuronales: ***online* y *target*** (ver Figura 2.3). La primera de ellas está compuesta por una serie de pesos que se componen en tres pasos: **codificador, proyector y un predictor**. La segunda también es un conjunto de pesos, estos se forman con la media móvil exponencial de los primeros. Dada una imagen se le aplican dos *image augmentations* distintos. Como se puede observar en la Figura 2.3 de la primera *image augmentation* la red *online* calcula una representación y una proyección y la red *target* las calcula también sobre la segunda *image augmentation*. Una vez se han calculado las dos proyecciones sobre la red *online* se calcula un predictor de la proyección de la red *target* y se lleva a cabo una normalización. Finalmente se calcula el error medio cuadrado entre la predicción normalizada y la proyección de *target*.

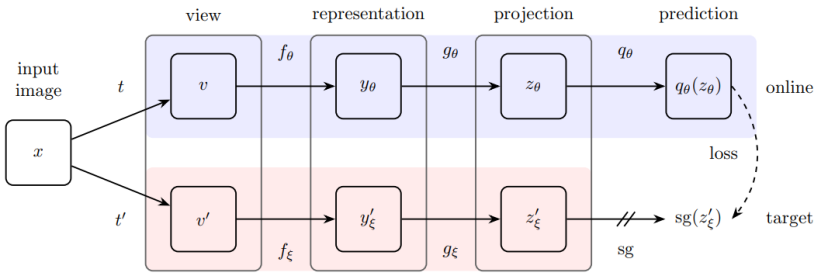


Figura 2.3: Arquitectura de BYOL. Figura obtenida de [3]

Capítulo 3

Descripción del sistema de adquisición de endoscopias

En este capítulo se describe el sistema de grabación original que se encontraba en funcionamiento en el proyecto y la versión final del sistema desarrollado que mejora y optimiza las funciones realizadas por el original, solucionando sus problemas y añadiendo distintas funcionalidades.

3.1. Sistema original

Para el desarrollo de este proyecto se partía de un software base de captura de endoscopias.

Como podemos observar en la Figura 3.1 el sistema original constaba de un módulo de grabación más simple, con las siguientes dos partes:

- **Capturador:** Esta parte está formada por una tarjeta capturadora de vídeo externa, conectada a la pantalla médica en la que los sanitarios ven la cámara del endoscopio. El proceso de captura del procedimiento se realiza guardando imágenes, 40 por segundo, de esta pantalla. Una vez se detecta que el endoscopio está encendido se comienzan a capturar las

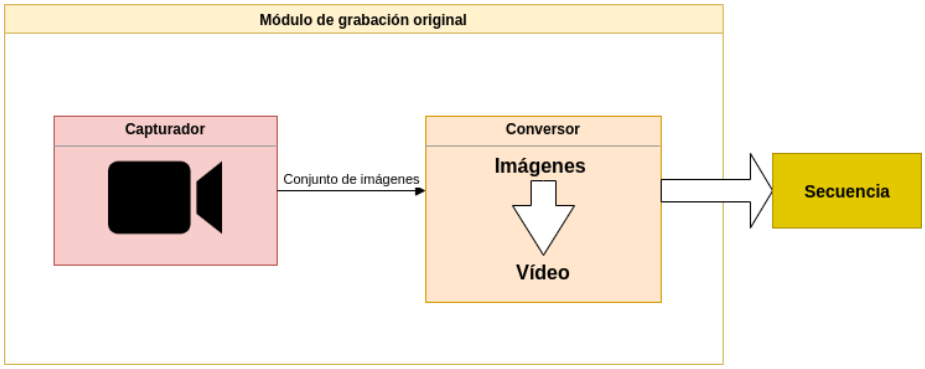


Figura 3.1: Diseño del sistema original de grabación

imágenes. Cada vez que uno de estos *frames* se guarda se procesa de tal forma que la imagen se corta en dos partes, la zona de la endoscopia, que se guarda para construir posteriormente en vídeo del procedimiento y la zona que corresponde a la información de la endoscopia realizada y los datos del paciente. La zona que contiene los datos del paciente se guarda únicamente una vez al inicio de la grabación, por motivos de privacidad y seguridad la imagen visible se guardará en un fichero comprimido al que se tiene acceso mediante una clave, además se le aplica un **OCR** para conseguir en formato texto esta información. A esta información se le aplica una función de cifrado **hash** para obtener el identificador de paciente cifrado y guardarlo en un fichero **JSON** con otra información sobre la grabación: la duración, número de *frames* capturados, etc.

- **Conversor:** Proceso en que se transforman las imágenes a vídeo. Una vez tenemos el conjunto de imágenes que formarán una grabación estas se almacenan en tres vídeos con distintas calidades y resoluciones. (1)***lossless***: vídeos sin compresión que ocupan más de 60 GB de media, tendrá la mejor calidad y ocupará más espacio; (2) ***lossy***: vídeo comprimido con pérdidas, tendrá menor calidad pero ocupará menos espacio, aproximadamente 1 GB de media; (3) ***thumbnail*** este tendrá la menor calidad de los tres pero ocupará mucho menos espacio, no llegando a superar los 100 MB de media. Una vez transformados, los vídeos se copiarán a un disco duro externo del cual se recogen semanalmente.

En esta versión base del sistema una vez se recogían las grabaciones no se realizaba ningún otro tipo de procesamiento, se subían directamente al almacenamiento *online* con el resto de las grabaciones.

Problemas y retos en el sistema original. Este sistema base presentaba varios problemas:

- **Pérdida de endoscopias**. Antes de realizar este proyecto, solo se conseguían grabar un tercio de las endoscopias realizadas en un día,principalmente por motivos de espacio.
- **Fallos en los datos de la endoscopia**. Se presentaban fallos al capturar la información: identificación del paciente, día, hora etc., ya que en muchas de las grabaciones al intentar capturar esta información, los médicos aún no la habían introducido y quedaba perdida, lo que suponía descartar posteriormente la secuencia, esto supuso una gran pérdida de vídeos en nuestro conjunto, dejándolo reducido casi a la mitad.

Con el software final desarrollado se busca solucionar estos problemas consiguiendo ampliar el número de endoscopias grabadas. La mayor dificultad viene de la forma de grabación utilizada, ya que se capturan las imágenes individualmente, ocupando una cantidad de espacio mucho mayor que al guardar únicamente los vídeos. También se ha querido solucionar los problemas con la información para en caso de fallos conseguir recuperar lo perdido y no tener que descartar ninguna de las secuencias.

3.2. Sistema Final

Partiendo del software base de grabación, se ha realizado una nueva versión del sistema de grabación que resuelve los problemas mencionados con anterioridad. El *software* de este sistema está explicado en el anexo A. En este sistema diferenciamos dos módulos: **módulo de grabación** y **módulo de segmentación**

3.2.1. Módulo de grabación

Este módulo es la versión mejorada del original. Con respecto al diseño inicial se han realizado modificaciones en el software del capturador y se ha añadido un nuevo proceso como podemos observar en el diagrama de la Figura 3.2. Aquí

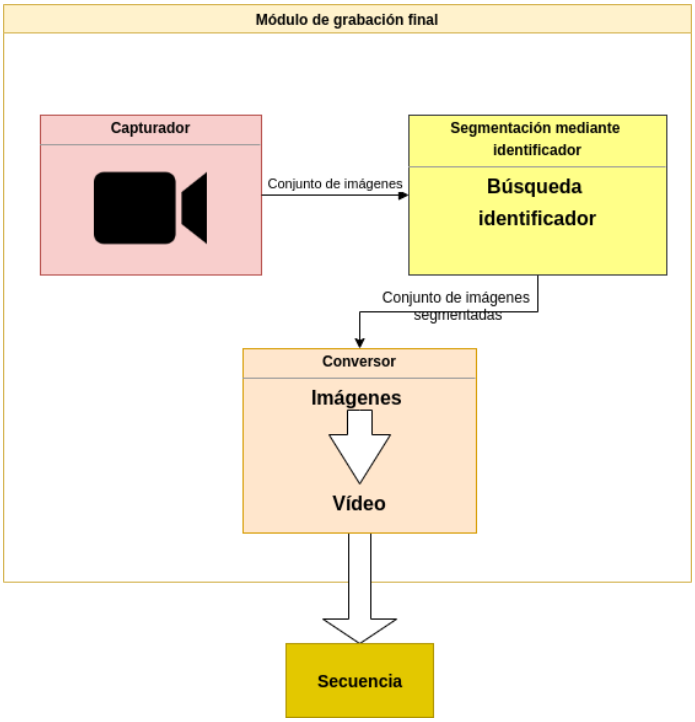


Figura 3.2: Diseño del módulo final de grabación

se presenta un breve resumen, el contenido del módulo y su evaluación está detallada en el capítulo 4. Las partes que conforman el nuevo módulo son:

- **Capturador:** La obtención de las imágenes se realiza del mismo modo comentado en el software inicial. El cambio se ha realizado a la hora de guardar la información de los pacientes. Esta información se guarda en formato de imagen, periódicamente a lo largo de toda la grabación, cada quince segundos. Al permitir cambios en ese tiempo en la información e ir almacenándolos se consigue tener todos los detalles que han ido modificándose a lo largo de la secuencia y no tener perdidas, sin calcular para cada una el **OCR**, que nos transforma estas imágenes a texto.

- **Segmentación mediante el identificador:** Este nuevo módulo es introducido en la versión mejorada del sistema. Una vez terminada la captura se busca en que secuencia se han introducido los datos del paciente, este *frame* se considerará el inicio válido del procedimiento. Se considera así ya que significa que el médico ha introducido los datos del paciente y se dispone a comenzar con la endoscopia, a partir de ese momento pero no antes.

- **Conversor:** Este módulo no ha sufrido ningún cambio con respecto a la versión inicial.

3.2.2. Módulo de segmentación

Este módulo se ha creado para intentar encontrar distintas secciones o información útil en la secuencia de manera automatizada. Aquí se presenta un breve resumen, el contenido del módulo y su evaluación está detallada en el capítulo 5. Como se puede observar en la Figura 3.3 este módulo consta de dos bloques distintos:

- **Segmentación supervisada:** En este primer sub-módulo se utiliza la combinación de distintos descriptores, que nos dan información de las imágenes como contornos, color, etc. Este bloque recibe la endoscopia completa y la devuelve recortada, eliminando las partes al inicio y al final que no corresponden al procedimiento médico.

- **Segmentación no supervisada:** Mediante el entrenamiento de un modelo de **deep learning** para la representación de imágenes **BYOL** se busca mejorar la segmentación obtenida de forma supervisada. Este bloque recibe también la endoscopia completa y busca patrones entre los distintos *frames* buscando secciones distintas del cuerpo, imágenes que no sean útiles, algunas en las que se pueda observar el uso de herramientas, etc.

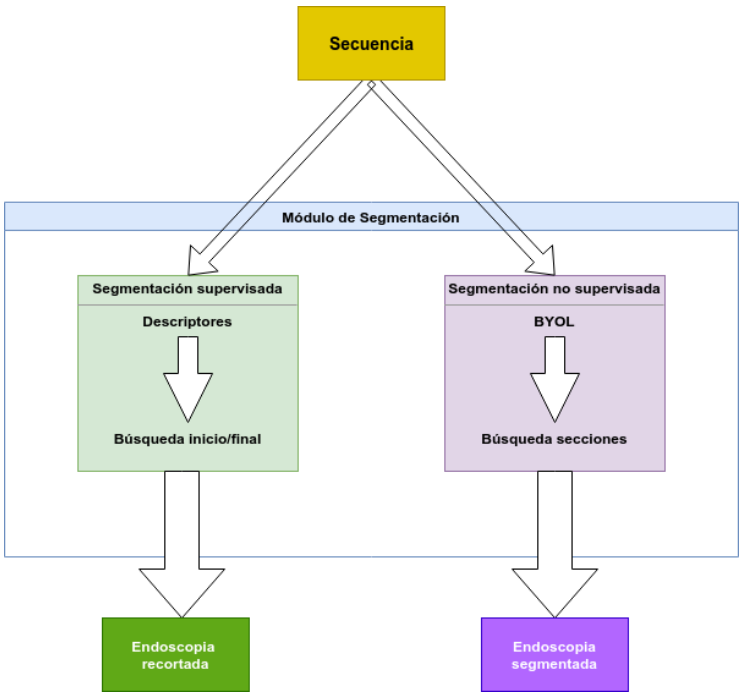


Figura 3.3: Diseño del módulo final de segmentación

Capítulo 4

Módulo de grabación

En este capítulo se va a explicar el proceso de grabación de endoscopias, así como el diseño del nuevo módulo de grabación que soluciona los problemas del diseño actual. También se detallará la evaluación llevada a cabo para este sistema.

4.1. Proceso de grabación

¿Cómo se obtienen las grabaciones?. La obtención de las grabaciones con las que trabajamos en este proyecto se han obtenido mediante el software comentado en la sección 3.1. Este sistema se encuentra instalado en el **Hospital Clínico Universitario Lozano Blesa**, conectado en una de las salas en las que se realizan las endoscopias. Cada día se graban los distintos procedimientos hasta que no queda más espacio en el ordenador. Al día se realizan entre diez y quince procedimientos, pero no todos se consiguen grabar por motivos de espacio. Esto sucede ya que que cuando no queda espacio libre, las grabaciones se copian a un disco duro externo que se cambia una vez a la semana. Manualmente, se clasifican las grabaciones en aquellas correspondientes a **colonoscopias y gastroscopias** y se suben al almacenamiento online del proyecto.

Problemas de almacenamiento. Como se ha comentado, uno de los principales intereses de segmentar automáticamente nuestros vídeos es para poder recortar trozos inútiles y reducir el espacio que ocupan, y así poder obtener el máximo número posible de grabaciones. Inicialmente se conseguían grabar unas seis endoscopias diarias, perdiendo las demás realizas ese día principalmente por motivo de espacio. Este problema no era posible solucionarlo incluyendo nuevos discos por las características del espacio en el que se realizan los procedimientos, no había suficiente espacio para introducir un ordenador más grande o muchos discos externos. El objetivo es grabar todas las endoscopias que se lleven a cabo.

Para solucionar esto, se ha introducido un nuevo componente de segmentación dentro de este módulo, comentado en el Capítulo 3 y explicado en detalle en el siguiente capítulo 5. Esta segmentación, o recorte del vídeo, no busca ser precisa para encontrar el momento concreto en que se inicia la endoscopia, ni en el que finaliza. El objetivo es reducir el espacio que ocupan las grabaciones para facilitar su manejabilidad y almacenamiento.

Solución propuesta. Puesto que conocemos el entorno y las condiciones en las que se llevan a cabo los procedimientos y las grabaciones, se posee cierta información para llevar a cabo un primer recorte de las secuencias. Así resulta posible descartar muchas imágenes que corresponden a trozos del vídeo no relevantes. Se sabe que los médicos introducen la información del paciente una vez este se encuentra en la sala y se va a proceder a realizar el procedimiento. Por esto nuestro sistema busca ese evento para descartar toda la información grabada con anterioridad. A continuación se describe como se ha implementado esta idea.

4.2. Segmentación o recorte mediante número de identificación del paciente

Se sabe que el número de identificación de los pacientes aparece en la parte izquierda de la imagen grabada, cuando el personal sanitario de forma manual introduce estos datos. Para detectar el número de identificación del paciente solo podemos acceder al vídeo (no hay otro tipo de metadato disponible). Por lo tanto, se va a utilizar un algoritmo existente de reconocimiento de caracteres: OCR (del inglés *Optical Character Recognition*). En particular, como se ha comentado el capítulo 2, utilizamos *Tesseract* [10].

Detección del ID del paciente en el vídeo y recorte del mismo en un *frame*. El algoritmo para detectar el ID del paciente en un *frame* está resumido en el diagrama de la Figura 4.1, y consta de los siguientes pasos:

1. Recorte inicial de la información de los pacientes en las imágenes guardadas, y su almacenamiento. Cada 12 segundos a lo largo de toda la grabación.

2. Lanzamiento del OCR en estos recortes una vez finalizada la grabación.

3. Hash por privacidad a la información obtenida de la aplicación del OCR.

4. Se busca por votación el resultado de la función *Hash* que más veces se repite y se selecciona como identificador.

5. Se recorta la parte previa de la grabación antes de que el identificador elegido aparezca por primera vez.

El uso de este método ha supuesto un considerable tiempo de pruebas en busca de los ajustes que dieran resultados robustos, ya que se observó que dependiendo de las configuraciones dadas, el número que devolvía no era correcto, en principio por el tamaño y la disposición del campo deseado en la imagen. Finalmente se consiguió ejecutar el algoritmo **OCR** de forma en que no se presentaran fallos obteniendo la predicción de número correcta, añadiendo redundancia en el sistema de obtención de imagen para en caso de fallar este tener más muestras que puedan conseguir el número de forma precisa. No utilizamos tampoco el identificador devuelto directamente por este método, por motivos de privacidad de los pacientes se utiliza una función **hash** que encripta este número como se observa en la Figura 4.1, el resultado de esta función es el finalmente

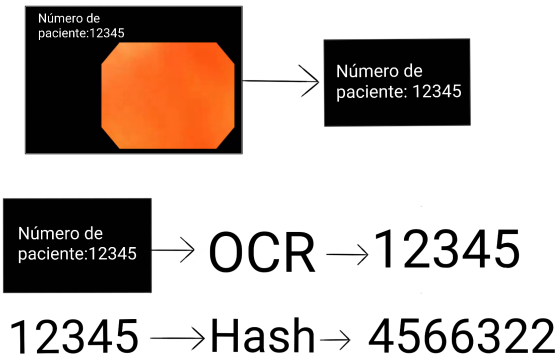


Figura 4.1: Proceso de la obtención del identificador de paciente encriptado

utilizado en todo el sistema, sin que exista la posibilidad de obtener el número de identificación original.

Durante la grabación, cada 12 segundos aproximadamente, se ejecuta el algoritmo de detección del ID del paciente. En este se almacena el recorte con la información del paciente de ese instante y el número de imagen en el que sucede.

Asignación del identificador más probable. Una vez finalizada la grabación, se lleva a cabo un procesamiento de todos los recortes con información para decidir a partir de que momento ya aparece la información del paciente de forma válida, como se ve en la Figura 4.2.

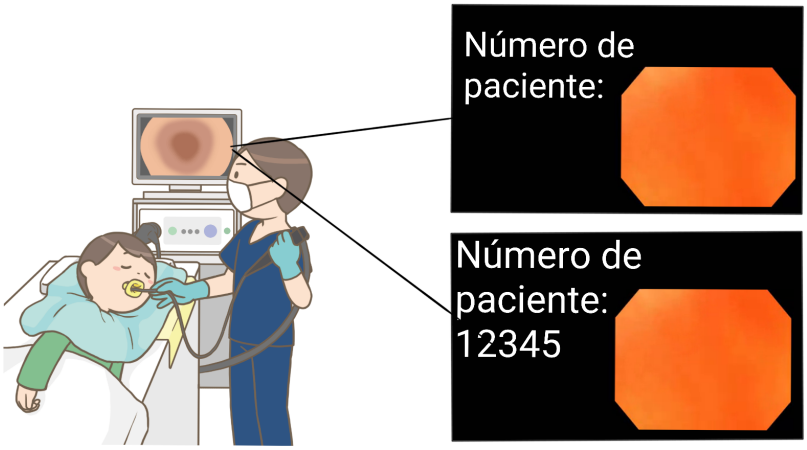


Figura 4.2: Ejemplo de una imagen sin los datos de paciente introducidos y otra en la que ya se han rellenado (imagen extraída de `https://www.pngegg.com/es/png-olhdf`)

Esta comprobación se realiza mediante un sistema de votación, se comprueban cuantos posibles identificadores distintos se han obtenido, descartando aque-

llos que están vacíos, porque aún no se habían introducido los datos. Una vez tenemos todos los posibles se calcula cuál se repite un mayor número de veces a lo largo de toda la grabación, este será el elegido como identificador del paciente.

Otra prueba que se intentó fue capturar los datos múltiples veces desde el comienzo de la grabación, procesar esta información en el momento de captura hasta encontrar el primer identificador válido, es decir no vacío y dejar de capturar la información. Este sistema también presentó fallos debido a que la información se introduce de forma manual, cosa que el sistema no es capaz de tener en cuenta, por lo que en algunas ocasiones, de forma menos habitual que el fallo de la opción anterior, se capturaba la información en el momento justo en el que se introducían los datos quedando en nuestros resultados de forma incompleta o de forma errónea si al introducir el identificador se había cometido un error que se corregía a posteriori.

Versión final (incorporada al sistema real). Finalmente se decide guardar la información a lo largo de toda la secuencia y procesarla al finalizar la captura, la redundancia que se obtiene nos permite acceder a los datos obtenidos una vez finaliza la grabación, lo que supone que en caso de que el programa hubiera fallado en alguna ocasión se podrían llegar a recuperar el resto de identificadores candidatos y encontrar el real sin haberlo perdido como sí sucedía en las anteriores propuestas.

Se ha optado por realizar este sistema de votación tras la realización de numerosas pruebas en las que se intentaba decidir un momento concreto de la grabación para capturar una única vez la información del paciente, a los cinco minutos del inicio de la grabación fue la idea inicial, se comprobó que esta opción fallaba en numerosas ocasiones ya que un importante número de secuencias se iniciaban mucho antes de que el personal médico introdujera los datos, quedando esta información vacía.

Determinar inicio de la secuencia. Una vez hemos determinado el identificador, como sabemos el número de secuencia en el que se ha visto este número por primera vez, se procederá a marcar como inicio esa secuencia, restándole algunos segundos, medio minuto de grabación aproximadamente, como método para prevenir fallos. Posteriormente se eliminarán todas las imágenes de la grabación anteriores a esta marca temporal que hemos creado.

4.3. Experimentos y evaluación

En esta sección evaluamos el correcto funcionamiento del módulo de grabación, tanto de la grabación en sí, como de la capacidad de recortar los vídeos correctamente (es decir, sin tirar información útil).

4.3.1. Evaluación de la Fiabilidad del detector de IDs.

Calidad de la detección de IDs. Para poder evaluar nuestro sistema lo primero que se tiene que comprobar es la fiabilidad del identificador encontrado, sin que se hayan producido fallos, ni en el OCR, ni en el sistema de votación.

Para ello se ha comprobado su funcionamiento durante la grabación de unos trescientos vídeos, en los que de forma manual se ha revisado si el identificador

asignado es el correspondiente. Durante aproximadamente los cien primeros hubo que ajustar la configuración del OCR, ya que presentaba algunos fallos en algunos de los casos, hasta que se consiguió controlar, los principales problemas surgieron por la captura de alguna imágenes en las que se devolvía vacío o cortado por la posición en la escena. Durante los restantes doscientos vídeos, se observó que el sistema que guarda la información de identificación funciona de forma adecuada sin presentar fallos.

En resumen, se comprobó que se realizaba de forma correcta la captura de la identificación.

Calidad de los recortes en el video. Además de comprobar que el identificador quedaba guardado de forma correcta, es importante asegurarse de que el momento temporal marcado como inicio de la grabación no supone ningún problema, es decir, que los cortes se realizan antes haber comenzado la endoscopia. Para ello, se guardó la información del momento en que supuestamente comenzaría, sin llegar a eliminar las imágenes anteriores, durante la recopilación de aproximadamente cien grabaciones.

Realizada esta prueba se observó que efectivamente el momento en que introducen la identificación es previo al comienzo de la endoscopia en la mayoría de los casos, para solucionar los errores que se producían se considero el momento temporal del inicio como veinticinco segundos exactamente antes de la secuencia marcada, así se corrigieron los casos en los que fallaba.

Conclusiones. Realizados estos experimentos se comprobó que el sistema no presentaba fallos en las tareas objetivo. Por tanto, es posible eliminar las imágenes grabadas anteriormente sin perder información, y se añadió al código la eliminación de las imágenes.

4.3.2. Segmentación de vídeo

Comprobación del funcionamiento de la segmentación. Una vez se ha comprobado el correcto funcionamiento del sistema, se eligieron diez grabaciones recopiladas durante la misma semana. Entre las seleccionadas encontramos tanto gastroscopias como colonoscopias de distintas duraciones. De este modo se pretende evaluar si el sistema cumple correctamente con el objetivo principal: disminuir la parte de vídeo almacenada que no es relevante. En la tabla 4.1 se observan los resultados obtenidos de este análisis. El campo **secuencia** corresponde al nombre de la grabación analizada. **Inicio real (**Ini_{real}**)** es el momento en que comienza el procedimiento médico. El campo **corte inicio (**$Ini_{estimado}$**)** corresponde al momento temporal en el que se ha encontrado por primera vez el identificador más votado, es decir, nuestra marca para cortar el vídeo. El campo **Inicio real - Corte inicial** corresponde al tiempo que queda entre esta marca temporal y el inicio real de la endoscopia ($Ini_{real} - Ini_{estimado}$). Como es posible observar en ningún caso este corte se realizaría después del inicio de la endoscopia por lo que no se perdería ningún tipo de información importante.

Análisis de la calidad de los recortes obtenidos. Si observamos la mediana del porcentaje de vídeo que se eliminaría con este corte, vemos que estaría en torno al 30 %. Eliminar este porcentaje de vídeo supone un ahorro en espacio

Secuencia	Duración	Inicio real	Corte inicio	Inicio real - Corte inicial	Porcentaje eliminado	Porcentaje vacío restante
HCULB_00349	0:33:40	0:24:40	0:15:57	0:08:42	47.39 %	25.85 %
HCULB_00358	0:14:33	0:09:25	0:08:40	0:00:45	59.56 %	5.21 %
HCULB_00359	0:12:00	0:06:10	0:03:15	0:02:55	27.08 %	24.30 %
HCULB_00360	0:24:47	0:13:11	0:12:12	0:00:59	49.25 %	3.97 %
HCULB_00361	0:21:14	0:14:08	0:13:40	0:00:28	64.32 %	2.24 %
HCULB_00362	1:15:11	0:35:33	0:24:17	0:11:16	32.31 %	14.99 %
HCULB_00363	0:30:29	0:13:20	0:05:57	0:07:22	19.54 %	24.19 %
HCULB_00364	0:19:12	0:04:40	0:02:00	0:02:10	10.41 %	13.94 %
HCULB_00365	0:22:20	0:06:55	0:02:12	0:04:42	9.89 %	21.09 %
HCULB_00366	0:22:23	0:04:07	0:02:00	0:02:07	8.93 %	9.49 %
Media					**32.87** %	**14.53** %
Mediana					**29.69** %	**14.47** %

Tabla 4.1: Tabla de resultados obtenidos por la segmentación mediante id

bastante considerable. Si además se tiene en cuenta que durante la grabación el formato en el que se almacenan los vídeos es guardando cuarenta imágenes por segundo, sin comprimir, este ahorro es bastante más significativo, pasando de ocupar 200GB de media a 140GB.

Por otro lado, en cada secuencia analizada de media queda un 14.53 % de vídeo no informativo al inicio de la grabación. Esto tiene en cuenta solo las imágenes previas al inicio real de la endoscopia, no tiene en cuanta las finales después de la endoscopia. Esto hace que el porcentaje de vídeo no-informativo almacenado aún sea bastante importante, y haga falta considerar añadir pasos adicionales para eliminarlo, como se comenta en el siguiente capítulo.

Conclusiones. Teniendo en cuenta estos resultados obtenidos se considera un buen primer corte en el vídeo ya que como se ha comentado, no se buscaba precisión, si no eliminar el máximo tiempo posible sin ocasionar perdidas. Eliminar un **30 %** del vídeo en media resulta una gran ayuda inicial y permite grabar una mayor cantidad de grabaciones que las que se podían capturar inicialmente al ocupar más espacio. Así se llegan a grabar en algunas ocasiones todas las endoscopias realizadas un mismo día y si no una gran mayoría de ellas. Además como se ha comprobado en ningún momento se pierde información útil y la proporción de vídeo vacío restante respecto al recortado es considerablemente baja, lo que supone un buen resultado para una heurística de estas características.

Este algoritmo se ha integrado en el módulo de grabación utilizado en el hospital y actualmente se esta utilizando ya en el proyecto para las adquisiciones.

Capítulo 5

Módulo de Segmentación

Una vez realizado el primer recorte del vídeo, descrito en el capítulo 4, mediante la detección del número de identificación del paciente correspondiente, la grabación resultante sigue teniendo mucha información que resultaría útil poder pre-procesar en más detalle de manera automática. Para ello se han implementado dos bloques adicionales.

5.1. Segmentación de inicio-final del procedimiento en el vídeo

El primero de los bloques implementados en este nuevo módulo consiste en una segmentación o recorte de forma supervisada que tiene como objetivo detectar con más precisión el inicio y el final del procedimiento médico de forma automática en el vídeo. Esta segmentación se realiza mediante la comparación de distintas características de cada imagen, comparando una imagen con su consecutiva. La idea general es ir buscando encontrar que sus características sean suficientemente distintas. Al detectar dos imágenes consecutivas que cumplen esa diferencia se determina el inicio de la endoscopia. Del mismo modo pero a la inversa, se detecta el final de la endoscopia.

El resultado de aplicar este proceso a una de las grabaciones es la endoscopia recortada, sin imágenes no informativas ni al inicio ni al final de la misma.

Las características que se han utilizado para analizar las imágenes han sido el **descriptor de HOG** y **el color dominante**, detallados a continuación.

Descriptores utilizados. Para comparar características entre dos imágenes se ha decidido utilizar primero el descriptor de HOG, descrito en 2. Con este se realizará la primera comprobación, buscando una diferencia considerable entre los dos descriptores:

$$d_{HOG} = d_E(d1, d2), \tag{5.1}$$

donde d_E es la la distancia euclídea de los dos descriptores ($d1$, $d2$). El umbral para considerar dos imágenes suficientemente distintas que d_{HOG} sea mayor que 9, comprobado de forma empírica.

El uso de este descriptor únicamente no conseguía cumplir con el objetivo de detectar el inicio y el final de la endoscopia. Esto sucede debido a que las

imágenes que no pertenecen a la endoscopia pueden ser muy distintas entre sí, dando valores de distancia euclídea muy similares a los obtenidos entre una imagen de este grupo y otra que sí pertenece al procedimiento.

Para mejorar la predicción de HOG se añadió una segunda comprobación, el color dominante de la imagen (rojo, azul o verde). Al pertenecer las imágenes de la endoscopia al interior del cuerpo humano, es fácil identificar que su color dominante es el rojo. Por otro lado aquellas imágenes que pertenecen al exterior del cuerpo humano, en estas circunstancias, tenían como color dominante el azul.

En la Figura 5.1 se observa el proceso completo de esta comprobación mediante los descriptores para saber si hay un cambio significativo entre imágenes o no. Es importante notar que no se evalúan imágenes consecutivas, si no *frames* separados por 15 segundos.

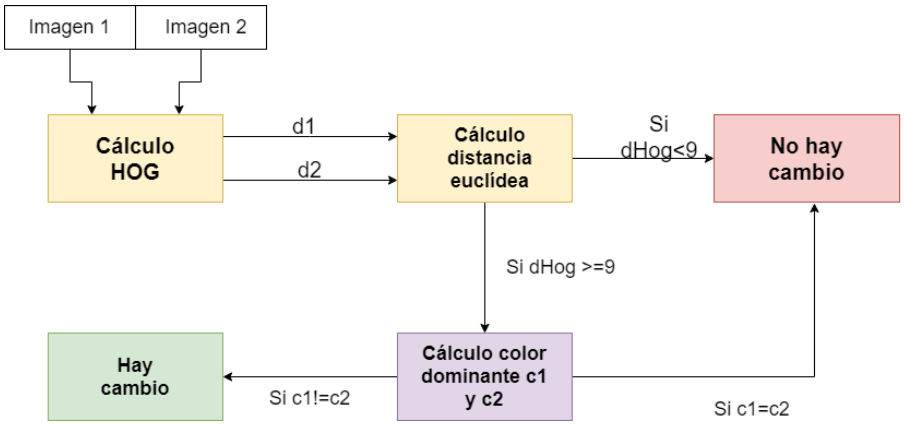


Figura 5.1: Funcionamiento del algoritmo para detectar cambio entre imágenes basado en el descriptor de HOG y el color dominante.

Combinando ambos se puede distinguir si hay un cambio sustancial en el contenido de una imagen respecto a la otra, es decir, en nuestro caso, separar una imagen que no pertenece a la endoscopia, de una que sí.

Detección del inicio. Una vez definido como medir los cambios significativos entre dos frames, para la búsqueda de los puntos de cambio endoscopia/no-endoscopia (es decir, inicio y final), se considera solo una de cada diez imágenes consecutivas, saltando las demás. Comenzando con la primera imagen del vídeo y su consecutiva (la 0 y la 10) se comparan los dos descriptores mencionados, si estos no cumplen los criterios establecidos se pasa al par siguiente (la 10 y la 20) hasta dar con el primer par de imágenes que cumplen, para considerar el inicio válido se comparan los descriptores de la primera de las imágenes de este par con las 10 siguientes. Si para las diez comparaciones se cumplen los requisitos, tanto de HOG como del color, se considera que se ha encontrado el inicio de la endoscopia, si no, se vuelve a comenzar la búsqueda.

Detección del final y recorte del vídeo. De igual modo que para el inicio se van comparando pares de imágenes consecutivas hasta encontrar aquellas en

la que las diferencias entre los descriptores son las requeridas. Igualmente esta condición debe cumplirse diez veces más con la primera imagen del par inicial. Una vez se ha encontrado de este modo el inicio y el final de la endoscopia se eliminan todos los *frames* anteriores al inicio y posteriores al final. Una vez eliminados se transforma a vídeo la secuencia.

Algoritmo desarrollado. Una vez unificados los pasos mencionados con anterioridad se conforma el algoritmo de segmentación con los siguientes pasos:

Algorithm 1: Algoritmo de segmentación inicio-final

```
Input: Lista de imágenes frames de una grabación.
Input: Función cambioEntreImágenes(imagen1, imagen2) que realiza
       las comprobaciones explicadas en 5.1.
Result: De la grabación dada se recorta la endoscopia tras la búsqueda
        de su inicio y final.
i=0;
inicioEncontrado=Falso;
finalEncontrado=Falso;
while no inicioEncontrado and i < longitud(frames) do
    if cambioEntreImágenes(frames[i], frames[i+10]) then
        inicioEncontrado=Verdadero;
        frameInicial=i;
        i=i+100;
    else
        i=i+10;
    end
end
while no finalEncontrado and i < longitud(frames) do
    if cambioEntreImágenes(frames[i], frames[i+10]) then
        finalEncontrado=Verdadero;
        frameFinal=i;
    else
        i=i+10;
    end
end
if inicioEncontrado then
    for i ← 0 to (frameInicial − 10) do
        frames.eliminar(i);
    end
end
if finalEncontrado then
    for i ← (frameFinal + 10) to longitud(frames) do
        frames.eliminar(i);
    end
end
```

5.2. Segmentación mediante el uso de BYOL

Con este segundo bloque se busca utilizar un modelo de aprendizaje automático no supervisado para la representación de imágenes, que nos devuelva un descriptor mucho más general que los utilizados en el bloque anterior, pudiendo ser aplicados también a otros tipos de imágenes. La idea es ver si con este descriptor se consigue encontrar y clasificar distintas secciones en las secuencias, sin buscar ningún tipo concreto, y analizar posteriormente su utilidad en nuestro tipo de escena concreto. Se ha utilizado el modelo **BYOL** [3], Bootstrap Your Own Latent descrito en el capítulo 2, entrenado y evaluado sobre nuestros vídeos de endoscopias para aprender una descripción representativa de las distintas imágenes.

Construcción del *dataset*. Para entrenar este modelo, una parte esencial es construir un buen conjunto de datos de entrenamiento. Se han elegido 25 vídeos de nuestras grabaciones del proyecto, entre ellos algunos contaban ya con algún tipo de meta dato, como las secciones que aparecían. Otros no contaban con ningún tipo de anotación previa. El conjunto final de imágenes para el entrenamiento consiste en aproximadamente 140000 *frames*, de los 25 vídeos.

Entrenamiento. Con el conjunto de datos recopilado utilizando algunas de las grabaciones que ya tenían información previa y otras elegidas aleatoriamente, se realizó un entrenamiento de este modelo con sus parámetros establecidos por defecto, sin darle ningún tipo de etiquetas a las imágenes. Este entrenamiento se ha llevado a cabo en una máquina específica para trabajar con modelos de deep learning, perteneciente al grupo de investigación donde se desarrolla el proyecto, una máquina **DGX-1**[1], poniendo en marcha el código y cargando el *dataset* en este sistema.

Una vez ha finalizado la optimización de los pesos de la red, donde el modelo aprende sobre las texturas, formas, los objetos, etc., la red es capaz de obtener descriptores discriminatorios para imágenes de endoscopias. Estos descriptores discriminativos nos permiten diferenciar los distintos tipos de imágenes de nuestras secuencias, como se analiza a continuación.

Evaluación mediante *clustering*. Para tener una representación más visual de la capacidad de discriminar patrones de interés, una vez calculados los descriptores en el conjunto de datos seleccionado, se aplica sobre estos el algoritmo de **k-means** [11], para agrupar en distintos *clusters* las imágenes. Como dentro de las grabaciones se pueden ver imágenes muy diversas, se aplico el algoritmo para 50 *clusters*.

La idea es ver que tipos de imágenes se agrupan entre sí, de acuerdo a los descriptores aprendidos con este modelo. Así, una vez asignadas las imágenes en los distintos grupos del *clustering*, se puede observar que imágenes se han agrupado juntas y analizar que propiedades tienen en común.

Con estos resultados, podemos dividir una endoscopia en distintas zonas pertenecientes al mismo *cluster*, y analizar si esas zonas comunes corresponden a un evento concreto de interés para luego implementar un detector automático de dicho evento.

[1] `https://www.nvidia.com/en-gb/data-center/dgx-systems/dgx-1/`

5.3. Experimentos y evaluación

En esta sección vamos a evaluar el funcionamiento de los bloques explicados y su capacidad para segmentar nuestras grabaciones.

5.3.1. Evaluación de la segmentación por descriptores

Comprobación de las métricas de los descriptores Lo primero que se evaluado fue el funcionamiento de la detección de cambio entre imágenes comentada en 5.1.

Para ello de forma manual, se separaron las imágenes de una de las grabaciones en aquellas pertenecientes a la endoscopia y las que no. Con estas imágenes separadas, se calculó la distancia euclídea del descriptor de HOG aplicada en los dos grupos por separado y entre las imágenes de un grupo y del otro. Como se puede observar en la tabla 5.1, el problema aparece para distinguir cuando dos imágenes son distintas al pertenecer una a la endoscopia y otra no como se ve en la tabla para el tipo *"no endoscopia"* que la media de la distancia de ese grupo es muy cercana a la distancia en el tipo que compara las imágenes que no pertenecen al procedimiento de las que sí. Siendo mucho más sencillo saber cuando dos imágenes pertenecen ambas a la endoscopia ya que como se observa en la tabla en el tipo *"endoscopia"* su distancia es mucho menor.

Añadiendo el descriptor del color dominante se consigue mejorar la precisión de la detección y solucionar los fallos que aparecían al utilizar solo el descriptor de HOG. Para ello también se comprobó que efectivamente el color dominante de media en las imágenes de la endoscopia era el rojo, con una media de color, en formato **BGR** de [41,2, 80,5, 167,6] y para las no pertenecientes a la endoscopia el azul, con una media de [77,0, 57,1, 34,0]. Con estos datos se comprobó que la

	Tipo	Media	Mediana	Mínimo	Máximo
(a)	Endoscopia	8.38	8.34	0.71	13.7
(b)	No endoscopia	10.87	10.8	0.31	18.36
(c)	Comparativa	10.81	10.41	5.84	18.32

Tabla 5.1: Estudio de los valores típicos de d_{HOG} entre pares de imágenes (a) ambas etiquetadas como pertenecientes al proceso de endoscopia; (b) ambas no pertenecientes a la endoscopia (es decir, las partes de inicio y fin de las grabaciones); (c) una de las imágenes comparadas pertenece a la endoscopia y la otra no pertenece.

combinación de ambos funciona, logrando separar las imágenes que pertenecen a la endoscopia y las que no correctamente.

Calidad de los recortes en el vídeo Una vez se ha comprobado que los descriptores son capaces de marcar el inicio y el final de la endoscopia, se ha seleccionado un conjunto de vídeos sobre los que de forma manual se ha anotado el inicio y final real del procedimiento. Este conjunto de vídeos se ha utilizado como prueba de nuestro programa.

Como se puede observar en la tabla 5.2 la media de vídeo eliminado que se ha conseguido con esta segmentación alcanza el 40 %.

Con respecto a la corrección del corte realizado se observa que se ajusta de forma bastante precisa al inicio y final reales, de todos ellos solo en dos se realiza el corte inicial con un pequeño retraso del inicio real. Comprobando estas pequeñas diferencias se determina que no suponen una perdida considerable de información, ya que las imágenes perdidas corresponden a imágenes no informativas al inicio, por lo que se asume como perdida razonable.

Secuencia	Duración	Inicio endoscopia	Final endoscopia	Corte inicio	Corte final	Porcentaje eliminado
HCULB_00073	0:38:07	0:25:28	0:37:08	0:25:28	0:37:24	68.69 %
HCULB_00074	0:19:37	0:08:58	0:18:30	0:08:58	0:19:37	45.69 %
HCULB_00089	0:41:44	0:24:00	0:40:00	0:23:59	0:40:14	61.07 %
HCULB_00216	0:19:01	0:03:14	0:17:16	0:03:13	0:17:23	25.46 %
HCULB_00301	0:02:26	0:00:14	0:01:32	0:00:13	0:01:32	46.11 %
HCULB_00345	0:06:18	0:00:51	0:04:46	0:00:51	0:04:46	37.77 %
HCULB_00346	0:43:19	0:10:48	0:39:56	0:09:22	0:40:18	28.59 %
HCULB_00357	0:17:39	0:01:51	0:11:50	0:02:07	0:12:18	42.30 %
HCULB_00376	0:12:42	0:04:42	0:10:22	0:05:47	0:11:55	51.73 %
HCULB_00378	0:22:23	0:08:11	0:22:23	0:05:59	0:22:23	26.71 %
HCULB_00379	0:20:26	0:05:10	0:19:35	0:05:07	0:19:46	28.32 %
HCULB_00380	0:17:55	0:04:16	0:15:56	0:04:14	0:16:17	32.85 %
HCULB_00388	0:21:55	0:05:22	0:18:19	0:05:13	0:18:20	40.15 %
Media						**41.19 %**
Mediana						**40.15 %**

Tabla 5.2: Resultados del recorte obtenido mediante la segmentación por descriptores

Conclusiones. Tras la evaluación de este bloque se concluye que se ha diseñado un programa de segmentación basado en descriptores con resultados muy acertados. Se ha conseguido reducir casi por completo las zonas del vídeo no pertenecientes a la endoscopia sin caer en la perdida de información. Esto supone reducir de forma muy importante las imágenes no informativas que podían considerarse ruido, mejorando la aplicación posterior de otro tipo de técnicas. Y además al eliminar el 40 % de la duración de los vídeos el espacio ocupado se reduce de forma muy importante, solucionando los problemas de almacenamiento que aparecían con anterioridad.

5.3.2. Evaluación de los resultados de BYOL

Para evaluar la corrección de los distintos grupos creados tras aplicar **k-means** a los descriptores obtenidos del entrenamiento de **BYOL** se han seleccionado cien fotos de cada grupo al azar para crear de forma gráfica un mosaico con ellas y así comprobar si se parecían entre sí.

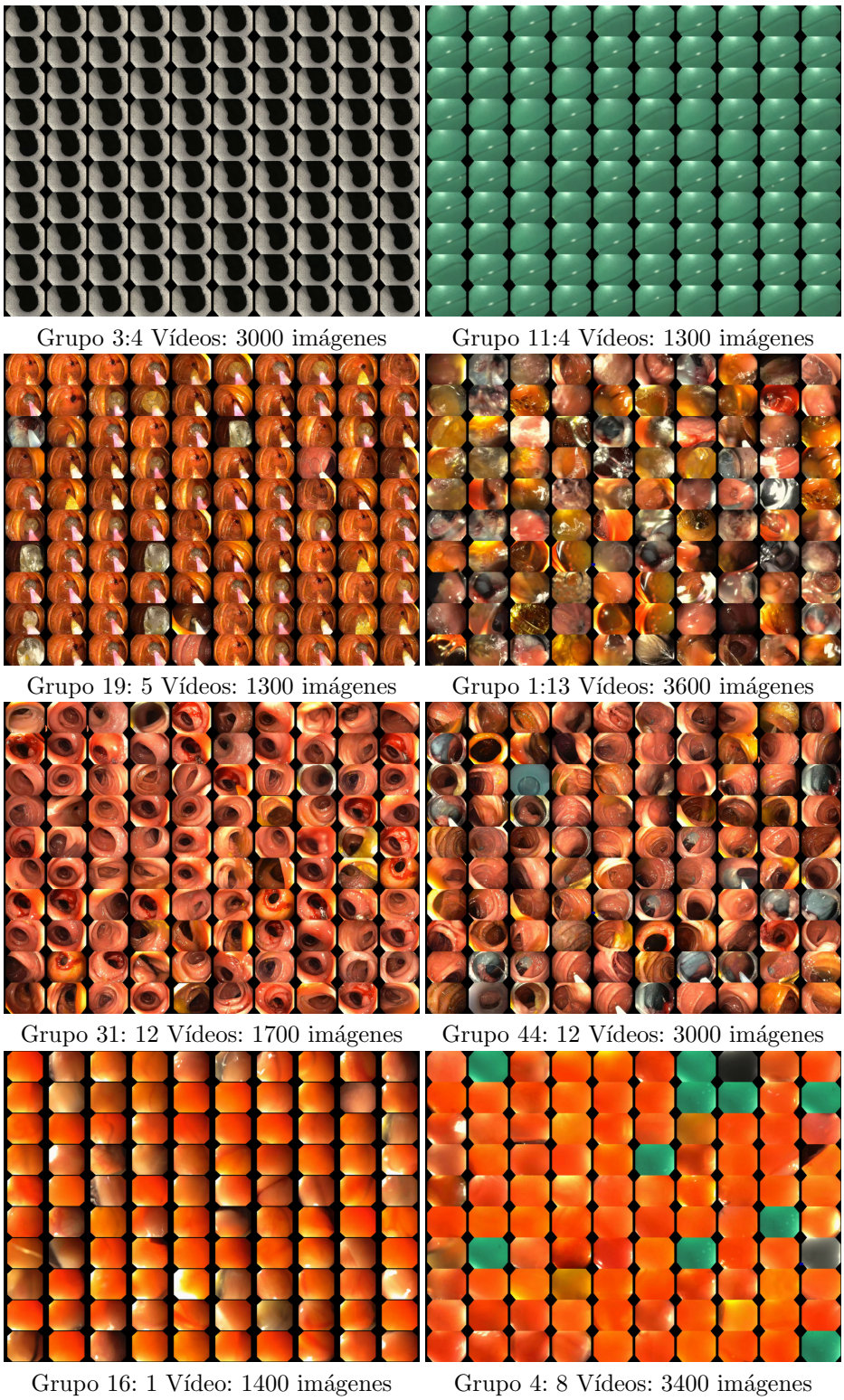
De forma visual, como se aprecia en la Figura 5.2 se puede ver muy claramente que el resultado de agrupar las imágenes de las endoscopias con el descriptor proporcionado por **BYOL** es bastante prometedor. Estos resultados se han obtenido con la evaluación de 13 grabaciones distintas, en los grupos 3 y 11 se observan dos de los grupos en los que se han clasificado imágenes que no pertenecen a la endoscopia. El resto de los grupos mostrados contienen, todos, imágenes pertenecientes al procedimiento. En el caso del grupo 19 se observa

el evento de aparición de herramientas. En el grupo 13, las imágenes agrupadas corresponden al evento de aparición de agua y/o burbujas. Las imágenes de los grupos 4 y 16 que corresponden al instante en que el endoscopio choca con las paredes de los intestinos y no se observa nada concreto. Conseguir identificar estas imágenes resulta muy útil ya que al procesar posteriormente las grabaciones suelen dar bastantes problemas ya que son imágenes puntuales que no tienen patrones similares a los de las distintas imágenes que aparecen cerca temporalmente. Por otro lado, tanto el grupo 21 como el 44 pertenecen a tipos de imágenes informativos del procedimiento que corresponden a distintas zonas del intestino. En el **anexo B** se encuentran las imágenes de los 50 grupos obtenidos para estas grabaciones.

Conclusiones. Gracias a los mosaicos obtenidos,se puede apreciar que las agrupaciones obtenidas con el descriptor aprendido son bastante significativas. Podemos separar, dentro de una misma endoscopia. las imágenes según el grupo en el que se han clasificado. Mediante este análisis se pueden clasificar los grupos que contienen solo imágenes no válidas, eliminándolas de nuestras grabaciones, dejando como resultado de la aplicación de este modelo únicamente las imágenes útiles e informativas de la endoscopia.

Además se podría realizar una clasificación más exhaustiva de los grupos obtenidos, determinando cada uno, si pertenece a una sección concreta vista en la endoscopia (Colon, recto, descendiente...) o si ocurre un evento concreto (extracción de pólipos, aparición de agua...).

En conjunto estas clasificaciones dan una visión mucho más completa de los distintos eventos que suceden en cada secuencia así como qué imágenes son interesantes para procesar.



Grupo 3:4 Vídeos: 3000 imágenes | Grupo 11:4 Vídeos: 1300 imágenes

Grupo 19: 5 Vídeos: 1300 imágenes | Grupo 1:13 Vídeos: 3600 imágenes

Grupo 31: 12 Vídeos: 1700 imágenes | Grupo 44: 12 Vídeos: 3000 imágenes

Grupo 16: 1 Vídeo: 1400 imágenes | Grupo 4: 8 Vídeos: 3400 imágenes

Figura 5.2: Ejemplos de mosaicos de distintos *clusters* obtenidos al aplicar *k-means*. Para cada uno se indica el número de imágenes en ese clúster, y de cuantos vídeos distintos vienen.

Capítulo 6

Conclusiones

En esta sección se discuten las conclusiones extraídas del sistema desarrollado en este proyecto así como de las futuras mejoras.

6.1. Conclusiones técnicas

Respecto a los objetivos iniciales establecidos, todos han sido cumplidos.

En primer lugar, se ha desarrollado un nuevo módulo de grabación que mejora y solventa los problemas del sistema original, pérdidas de endoscopias y de la información de los pacientes. Así, ahora el sistema es capaz de guardar todas las endoscopias realizadas a diario gracias a reducir su tamaño, mediante el recorte basado en la obtención del identificador del paciente.

El principal problema de este apartado ha sido la elección del momento en que cortar el vídeo de forma inicial, lo que también supone una limitación, ya que el momento en que aparece el identificador no es lo suficientemente preciso para acotar el inicio y dependiendo de la grabación aparece antes o después.

Además, con el nuevo módulo de segmentación se ha conseguido acotar el inicio y el final sobre las endoscopias, recortando automáticamente las partes que no pertenecían al procedimiento, lo que reduce aún más el espacio que ocupan nuestra grabaciones y mejora y facilita el uso de las secuencias en trabajos posteriores.

Finalmente se ha analizado la capacidad del modelo utilizado, **BYOL**, para ayudarnos a discriminar de manera automática los distintos tipos de imágenes que contienen nuestros vídeos. Este tipo de métodos no supervisados, no solo generan descriptores capaces de encontrar las secciones de los videos sin información para poder eliminarlas sino que también son capaces de separar los diferentes segmentos y eventos de las endoscopias como se ha visto en este trabajo.

Parte de este sistema desarrollado, el módulo de grabación, se encuentra instalado y en pleno funcionamiento en el Hospital Clínico de la Universidad de Zaragoza, está planificado que sea instalado en otros hospitales que colaboran con el proyecto.

6.2. Trabajo Futuro

Como futuras mejoras se plantea integrar los dos módulos desarrollados. De esta forma el bloque de segmentación por medio de identificadores planteado en el módulo de segmentación se añadirá a la grabación, para realizar el recorte de la endoscopia de forma directa. Una vez integrados los módulos se instalará el sistema completo en los hospitales mencionados para realizar una mejor recopilación de las endoscopias reales.

Además analizando los resultados obtenidos con **BYOL**, se planea desarrollar un sistema para la total segmentación de las grabaciones de forma plenamente automática. Este sistema marcaría los momentos temporales en que inicia y termina una nueva sección y les daría etiquetas a cada tipo de ellas, ofreciendo una versión final de la endoscopia clasificada en distintos segmentos.

Bibliografía

[1] Ravina Mithe, Supriya Indalkar, and Nilam Divekar. Optical character recognition. *International journal of recent technology and engineering (IJRTE)*, 2(1):72–75, 2013.

[2] official documentation Scikit image. Skimage feature hog. `https://scikit-image.org/docs/dev/api/skimage.feature.html#skimage.feature.hog`. Consultado el 10/03/2021.

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[4] Real-time mapping from endoscopic video. https://sites.google.com/unizar.es/endomapper, 2020.

[5] Liang-Chi Chiu, Tian-Sheuan Chang, Jiun-Yen Chen, and Nelson Yen-Chung Chang. Fast sift design for real-time visual feature extraction. *IEEE Transactions on Image Processing*, 22(8):3158–3167, 2013.

[6] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[7] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2016.

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[10] Chirag Patel, Atul Patel, and Dharmendra Patel. Optical character recognition by open source ocr tool tesseract: A case study. *International Journal of Computer Applications*, 55(10):50–56, 2012.

[11] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.