XLIV Jornadas de Automática 2023

https://jautomatica.es

# Cálculo de flujo óptico denso en colonoscopia mediante aprendizaje no supervisado

Gonzalo, Iván[1,*], Morlana, Javier[1], Montiel, JMM[1]

*Instituto de Investigación en Ingeniería de Aragón (I3A). Universidad de Zaragoza.*

To cite this article: Gonzalo, I., Morlana, J., Montiel, J.M.M., 2023. Dense optical flow estimation in colonoscopy images using an unsupervised learning approach. XLIV Jornadas de Automática. 861-866. https://doi.org/10.17979/spudc.9788497498609.861

**Resumen**

La colonoscopia es la técnica de referencia en la detección del cáncer colorrectal. Sin embargo, los métodos asistidos por ordenador no se usan mucho en estos procedimientos. Este trabajo se enmarca dentro del proyecto EndoMapper, que tiene como objetivo crear reconstrucciones del colon que puedan ser utilizadas para ayudar a los médicos o para llevar a cabo cirugía robotizada. La asociación de datos es un elemento clave en estos sistemas, realizando la asociación entre píxeles de la imagen para permitir posteriores reconstrucciones 3D. En este trabajo, evaluamos un método de estimación de flujo óptico denso sobre imágenes de colonoscopia, adaptándolo al dominio del colon mediante aprendizaje no supervisado. Diseñamos un conjunto de datos para *training*, *validation* y *test* a partir de las secuencias reales de colonoscopia del conjunto de datos de Endomapper. El modelo se ha re-entrenado obteniendo una versión adaptada al dominio del colon. La validación experimental muestra cómo el modelo entrenado puede estimar el flujo de manera robusta bajo cambios de iluminación. También muestra una capacidad excepcional para estimar el flujo entre imágenes de colonoscopia muy separadas con grandes rotaciones.

*Palabras clave:* endoscopia, flujo óptico, aprendizaje no supervisado

**Dense optical flow estimation in colonoscopy images using an unsupervised learning approach**

**Abstract**

Colonoscopy is the gold standard in colorectal cancer screening. However, computer-assisted interventional methods are not widely used in these procedures. The Endomapper project, in which this work is embedded, aims to create reconstructions of the colon that can be used to assist doctors or for robotic surgery. Data association is a key element in these systems, performing the association between image pixels to enable subsequent 3D reconstructions. In this work, we evaluated a dense optical flow method on colonoscopy images, adapting it to the colon domain by using unsupervised learning. We built a dataset for *training*, *validation* and *test* from the real colonoscopy sequences of the Endomapper dataset. The model has been re-trained obtaining a version adapted to the colon domain. Experimental validation shows how the trained model is able to estimate flow robustly under illumination changes. It also shows an exceptional ability to estimate flow between widely separated colonoscopy images with large rotations.

*Keywords:* Endoscopy, Optical Flow, unsupervised learning

## 1. Introducción

El proyecto Endomapper tiene como objetivo conseguir un mapa en tiempo real del colon a partir de secuencias de colonoscopias, Este mapa puede ser de gran ayuda tanto para médicos como para llevar a cabo cirugía robotizada. Algunos trabajos previos capaces de obtener duchas reconstrucciones son el SLAM (Simultaneous Localization And Mapping) denso de Ma et al. (2021), que obtiene un mapa denso y la posición de la cámara en el colon, usando una red de profundidad y un tracking fotométrico o el trabajo de Rodríguez et al. (2022), que considera entornos deformables y hace un tracking basado en Lucas-Kanade Baker and Matthews (2004). Uno de los procesos de vital relevancia en las tecnologías SLAM es la asociación de datos que, en el dominio de las imágenes, supone ser capaces de relacionar píxeles entre dos imágenes que corresponden al mismo lugar.

Tradicionalmente, la asociación de datos se ha hecho con características clásicas como ORB Rublee et al. (2011) o SIFT Lowe (2004), que fallan en el colon debido a las bajas texturas y los cambios de iluminación. Otra familia de métodos son los de flujo óptico clásico Fortun et al. (2015), que tratan de obtener el movimiento de todos los píxeles de una imagen a otra. Estos métodos suelen depender de la consistencia fotométrica, es decir, que la intensidad de dos píxeles correspondientes en

*[797115,jmorlana,josemari]@unizar.es

dos imágenes sea constante. Esta restricción no se cumple en las imágenes del colon, haciendo que estos métodos tampoco funcionen bien.

Los métodos de flujo más modernos utilizan métodos de *deep learning*, y han demostrado ser capaces de superar algunas de las limitaciones de los métodos de flujo clásicos. Una ventaja de estos métodos de flujo es que consideran toda la información de la imagen, en lugar de utilizar información únicamente local como en el caso de las características locales. La desventaja de estos métodos es que confían en tener una supervisión ground truth a la hora de entrenar los modelos de estimación de flujo, que no suele estar disponible en el caso de las colonoscopias. Ilg et al. (2017); Teed and Deng (2020) usaban supervisión a través de simulaciones, ya que en la práctica es complejo hallar supervisión fiable para todos los píxeles. En ocasiones, el salto de dominio entre simulaciones e imágenes reales puede provocar que los modelos no generalicen a entornos reales.

Debido a estas restricciones, se ha optado por un modelo de estimación de flujo que una utiliza una técnica de entrenamiento no supervisado, *Warp Consistency* Truong et al. (2021), requiriendo únicamente pares de imágenes que observen el mismo lugar para realizar el entrenamiento de la red. *Warp Consistency* fue diseñada para la estimación de flujo óptico en imágenes del Megadepth Dataset Li and Snavely (2018), compuesto por imágenes de edificios y monumentos. Sin embargo, se ha comprobado que posee un buen desempeño en el ámbito de las colonoscopias Azagra et al. (2022). *Warp Consistency* es una técnica aplicable a cualquier red de flujo. En este trabajo, se ha utilizado GLU-Net Truong et al. (2020), la misma red que emplean los autores originales.

La elección de este sistema se sustenta en las dos cuestiones mencionadas previamente. Por un lado, el uso de un modelo de flujo basado en *deep learning* permite calcular una transformación coherente para todos los píxeles de la imagen, al utilizar toda la información contenida en la imagen al mismo tiempo. Esto, a diferencia de los métodos de características locales, permite estimar una transformación entre imágenes aun cuando la textura local en la imagen sea baja. Por otro lado, *warp consistency* emplea un algoritmo no supervisado, lo cual es vital para el dominio del colon, donde no se dispone de supervisión confiable.

La validación experimental muestra la mejoría al entrenar el modelo con las imágenes del Endomapper Dataset Azagra et al. (2022). Otros trabajos como Morlana et al. (2021) y Morlana et al. (2023) ya habían demostrado la eficacia de entrenar con imágenes covisibles en colonoscopia, en su caso para el reconocimiento de lugares.

Las contribuciones principales de este trabajo son:

- Entrenamiento no supervisado de GLU-Net en imágenes reales de colonoscopia
- Construcción del dataset a partir de secuencias del Endomapper, obteniendo la covisibilidad mediante SfM Schonberger and Frahm (2016)

## 2. Aprendizaje no supervisado del flujo óptico

El flujo óptico entre dos imágenes se puede definir como el movimiento que hay que aplicar a cada píxel de la primera imagen para que coincida con su correspondiente píxel en la segunda imagen. De esta forma, si a cada píxel de la primera imagen se le aplica su correspondiente valor de flujo, se obtiene una aproximación de la segunda imagen (Figura 1).

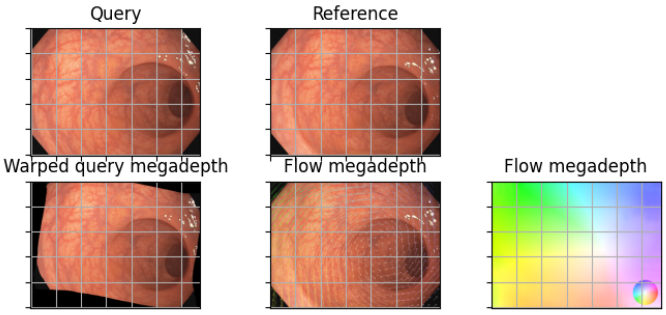


Figura 1: Estimación de flujo óptico. Se muestran las imágenes de *query* y *referencia*, así como la *warped query* (izda), resultado de aplicar el flujo a la *query* para convertirla en la imagen de referencia. También se muestra el flujo muestreado (centro) y el denso (dcha).

### 2.1. Flujo óptico no supervisado mediante warp consistency

*Warp Consistency* (WarpC) es una técnica para supervisar el cálculo del flujo óptico denso. Su funcionamiento se basa en el empleo de tripletas de imágenes. Para supervisar el flujo óptico $F$ entre 2 imágenes $I$ y $J$, se crea en primer lugar la imagen $I'$, que es el resultado de aplicar una transformación conocida $W$ a la imagen $I$. Estas transformaciones pueden ser homografías, transformaciones afines o transformaciones *thin plate spline* (tps).

Con las 3 imágenes, se calcula por un lado el flujo desde la imagen $I$ a la $J$ y por otro el de la $I'$ a la $J$. Como se conoce el flujo $W$ de $I$ a $I'$ y se tiene tanto el de $I'$ a $J$ como el de $I$ a $J$, se puede evaluar la calidad de la estimación del flujo de $I$ a $J$, porque debería ser la misma que de $I$ a $J$ pasando por $I'$, a partir de esta evaluación de la calidad de la estimación de flujo se puede establecer la función de loss y la supervisión.

Utilizando una notación más formal, se puede definir $W$ como el flujo de $I'$ a $J$ más el warping $I'$ a $J$ del flujo de $J$ a $I$:

$$W = F_{I'\to J} + \Phi_{F_{I'\to J}}(F_{J\to I}) \tag{1}$$

donde el warping $\Phi_F$ de una función T se define como el flujo F que cumple: $\Phi_F(T)(x) = T(x + F(x))$.

### 2.2. GLUNet

En este trabajo se ha aplicado WarpC a la red de estimación de flujo GLUNet Truong et al. (2020) siguiendo la propuesta de Truong et al. (2021).

La arquitectura de GLUNet está basada en capas de correlación global y rama local para estimar el flujo óptico denso entre imágenes. La red se divide en dos ramas: una rama global que captura características de alto nivel y una local que se enfoca en detalles más específicos. Estas ramas se fusionan mediante un módulo de fusión adaptativo que ajusta la resolución de la entrada global para adaptarse a la escala del detalle local. Además, con esta arquitectura es posible calcular con gran precisión y fiabilidad desplazamientos a largas distancias, incluso cuando hay cambios en la apariencia o el ángulo de observación.

### 2.3. *Loss function*

La función de pérdida (loss) utilizada en WarpC se basa en la consistencia de flujo (*warp consistency*), que establece que el flujo óptico estimado entre dos imágenes debe ser consistente con el flujo óptico estimado entre una imagen y su versión deformada. Esto se detalla en la Subsección 2.1. Teniendo esto en cuenta, la función de loss de WarpC se define como la suma ponderada de dos términos. Por un lado, un término de consistencia de flujo global, que mide la coherencia del flujo óptico en toda la imagen y, por otro lado, un término de flujo local que tiene en cuenta la consistencia del flujo en regiones pequeñas.

Con la misma notación que la utilizada en Subsección 2.1, los términos que componen la función de loss de WarpC son la loss *W-bipath* (Ecuación 2) y la loss *warp* (Ecuación 3). La primera se obtiene a partir de la consistencia del flujo entre $I'$ y $J$ más el warp de $J$ a $I$ con $W$ (que es el flujo sintético de $I'$ a $I$). Es decir, esta loss se minimiza cuando el flujo de $I'$ a $J$ más el warp de $J$ a $I$ es similar a $W$. Por otra parte, la segunda función de loss se obtiene de la restricción de consistencia del flujo de $I'$ a $I$ con respecto a $W$. Como $W$ es el flujo de $I$ a $I'$, esta función se minizará si ambos flujos tienen un valor similar. Estas 2 funciones de loss se agregan en (Ecuación 4) junto a un parámetro de regularización $\lambda$. Este parámetro $\lambda$ se ajusta automáticamente tras cada iteración del entrenamiento según se indica en la Ecuación 5.

$$L_W = \left\| \widehat{F}_{I'\to J} + \Phi_{\widehat{F}_{I'\to J}}(\widehat{F}_{J\to I}) - W \right\| \quad (2)$$

$$L_{warp} = \left\| \widehat{F}_{I'\to I} - W \right\| \quad (3)$$

$$L = L_W + \lambda L_{warp} \quad (4)$$

$$\lambda = L_W / L_{warp} \quad (5)$$

$$(6)$$

## 3. Adaptación al dominio de las colonoscopias

A continuación se detallan las modificaciones realizadas para adaptar el modelo WarpC+GLUNet al dominio de las colonoscopias.

### 3.1. *Datos de entrenamiento y validación*

Los datos de entrenamiento consisten en 8 secuencias del Endomapper Dataset Azagra et al. (2022). Estas imágenes han sido procesadas con COLMAP Schonberger and Frahm (2016), obteniendo una serie de clústers de ellas. Un clúster es un conjunto de imágenes covisibles, es decir, que observan el mismo lugar. Se define que dos imágenes son covisibles si observan puntos en común. Cada secuencia contiene 24.63 clústers de media (obtenidos mediante Schonberger and Frahm (2016)), teniendo cada clúster una media de 125.15 imágenes.

Los datos de validación se componen de 2 secuencias distintas a las 8 de entrenamiento pero que siguen su misma estructura. Además, con todas las imágenes de cada clúster se puede construir su matriz de solapamiento. Esta es la matriz que representa el nivel de covisibilidad de cada par de imágenes del clúster. Cada uno de los elementos de dicha matriz se obtiene a partir de cada par de imágenes, más concretamente del resultado de dividir el número de puntos que tienen en común con el número de puntos de cada una de las imágenes. En la Figura 2 se pueden ver ejemplos de un par de imágenes que tienen un alto nivel de covisibilidad (0.61) y otro cuyo nivel de covisibilidad es bajo (0.0017).

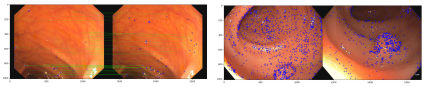

Figura 2: Covisibilidad. Alta covisibilidad (izda), 0.61, dos frames de diferencia. Baja covisibilidad (dcha), 0.0017, 200 frames de diferencia.

### 3.2. *Datos de test*

De cara a evaluar el desempeño del modelo WarpC+GLUNet, se ha diseñado un sistema de evaluación basado en flujo óptico *sparse*, en el que no se dispone del valor de flujo para todos los píxeles, sino para un subconjunto de ellos. A partir de una secuencia diferente de las de entrenamiento y validación, se han obtenido un conjunto de clústers siguiendo el mismo sistema que la Subsección 3.1. De estos clústers, se han obtenido pares de imágenes con distinto nivel de covisibilidad (Figura 2) y se han ordenado en función de su flujo medio ground-truth. El flujo ground-truth utilizado se trata de un flujo *sparse* debido a que sólo se conoce su valor para un grupo reducido de píxeles de cada par de imágenes. De esta forma, se han generado un conjunto de 79 pares de imágenes. En la Figura 3 se muestran tres tests (el más fácil, el más difícil y el que se encuentra en la mitad) del conjunto total de tests junto con su flujo ground-truth. Estos son los tests 1, 39 y 79 y tienen un valor de flujo medio de 6.40 px, 112.99 px y 412.77 px respectivamente.

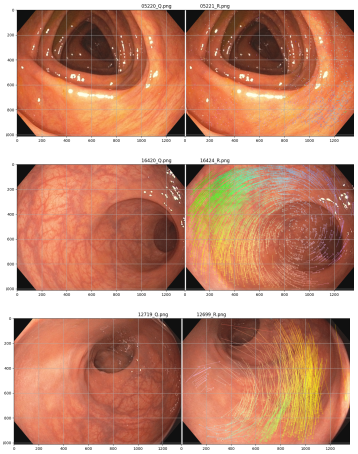

Figura 3: Pares de imágenes utilizados para test. La primera columna muestra la imagen de query y la segunda la de referencia. Sobre las imágenes de referencia se muestra el flujo *sparse* ground-truth.

### 3.3. *Metodología de entrenamiento*

Para entrenar el modelo, primero se seleccionan los pares de entrenamiento y validación a utilizar. Después, se filtran dichos pares utilizando dos umbrales mínimo y máximo para los valores de la matriz de solapamiento (Subsubsección 4.3.3). Una vez filtrados los pares, se aplica un redimensionado y un recorte a todas las imágenes (Subsubsección 4.3.1). Una vez que las imágenes tienen el tamaño adecuado, se inicia el entrenamiento. El entrenamiento está configurado para ejecutarse durante

80 épocas. Para cada una de ellas, se eligen 300 pares de forma aleatoria con los que se entrena GLUNet mediante WarpC y 25 pares, también escogidos de forma aleatoria, para la validación.

## 4. Experimentos

El objetivo de esta sección es evaluar el desempeño del entrenamiento auto-supervisado mediante WarpC+GLUNet Truong et al. (2021). Por una parte, se evalúa el sistema out-of-the box, esto es, entrenado en Megadepth Li and Snavely (2018) y con test en EndoMapper Azagra et al. (2022). Posteriormente, se evalúa como mejora el desempeño al entrenar con el EndoMapper dataset Azagra et al. (2022). Finalmente, se muestran ejemplos de la capacidad del WarpC+GLUNet para calcular correspondencias entre imágenes reales del colon.

### *4.1. Métricas de evaluación*

Para evaluar los distintos modelos que se van a utilizar, se va a hacer uso de dos métricas: el AEPE y el histograma de error acumulado. El EPE (End Point Error) es el error existente entre la posición estimada de un píxel de una imagen y su posición real según el ground truth. De esta forma, el AEPE (Average End Point Error) es la media de los EPE de una imagen. Por su parte, el histograma de error acumulado muestra la distribución de los errores de todos los puntos medidos en cada par de imágenes.

### *4.2. Sistema out-of-the-box*

En esta sección se van a describir las pruebas realizadas con el modelo WarpC+GLUNet propuesto por Truong et al. (2021). Este modelo se utilizará con el nombre *megadepth* de cara a los diferentes tests. El modelo *megadepth* ha sido entrenado utilizando el Megadepth dataset Li and Snavely (2018) que consta de imágenes de edificios de 196 localizaciones.

#### *4.2.1. Toleracia a cambios de iluminación*

En los pares de imágenes del Endomapper Dataset Azagra et al. (2022), la iluminación puede llegar a tener un alto nivel de variabilidad. Con el objetivo de comprobar si esta variación afectaba a la eficacia de los modelos anteriores, se diseñó un test en el que se aplicó una variación artificial de la iluminación mediante la modificación del brillo de las imágenes.

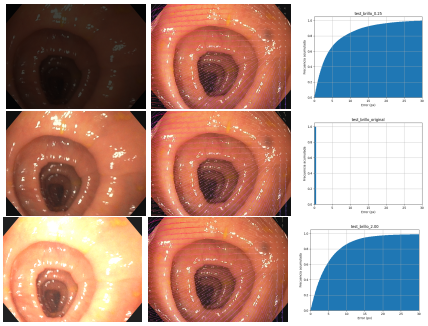

Figura 4: Test de robustez frente al cambio de brillo. Cada fila representa los cambios de brillo de factor 0.25, original, y 2. En cada columna se pueden ver la imagen de query con el nivel de brillo modificado, la imagen de referencia con el flujo dibujado con flechas y el histograma acumulado de la diferencia de flujo en valor absoluto respecto del flujo del par original.

En concreto, se han realizado un conjunto de 4 tests alterando los valores de brillo con factores 0.25, 0.5, 1.5, 2.0. En todos los tests de brillo se utilizaron las imágenes de query y de referencia presentes en la fila central de la Figura 4. En la figura, cerca del 80 % de los errores son inferiores a 10 píxeles y el error en la mediana está próximo a los 2.5 píxeles. También puede apreciarse como el patrón general del flujo se mantiene. Se muestra que el modelo escogido, WarpC+GLUNet, es capaz de seguir produciendo flujos válidos ante cambios bruscos de iluminación, mostrando así su robustez.

### *4.3. Sistema entrenado en EndoMapper*

En el entrenamiento es posible modificar una serie de parámetros que influyen en el modelo generado. A continuación, se describen aquellos que han sido modificados para obtener el modelo final, que se utilizará con el nombre *endomapper*.

#### *4.3.1. Factores de reescalado y recorte*

A las imágenes de entrenamiento se realiza un reescalado a $750 \times 750$ para, posteriormente, realizar un recorte de la zona central de tamaño $520 \times 520$. De esta forma, se eliminan los posibles bordes negros de la imagen. Sin embargo, dado que las imágenes de Azagra et al. (2022) son rectangulares de tamaño $1350 \times 1012$, este procedimiento causa un cambio del píxel aspect ratio en dichas imágenes que se traduce en una menor capacidad para estimar correctamente el flujo. Para solucionar este inconveniente, se modificó el factor de reescalado para que fuera rectangular, conservando la relación de aspecto de las imágenes. El recorte por su parte, se configuró para capturar el mayor área posible de cada imagen, de forma que, la forma de la imagen recortada siguiera siendo cuadrada como se puede ver en la Figura 5.

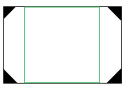

Figura 5: Recorte de forma cuadrada que maximiza el área capturada de las imágenenes del Endomapper Dataset Azagra et al. (2022).

#### *4.3.2. Factor de giro*

Este parámetro controla el valor del máximo ángulo que podían tener las rotaciones generadas por el generador de warps para la aumentación de los datos durante el entrenamiento de la red. Originalmente, este valor estaba establecido en $\pi/12$ rad (que se corresponden con 15°). No obstante, debido a que en los pares de imágenes utilizados existen rotaciones mayores, se ha incrementado este valor a $\pi/4$ rad (correspondiente con 45°).

#### *4.3.3. Umbrales de la matriz de solapamiento*

Este parámetro controla los valores mínimo y máximo a la hora de seleccionar las imágenes en función de su nivel de covisibilidad (Subsección 3.1). De esta forma, los pares de imágenes que tengan un valor inferior al umbral mínimo o superior al umbral máximo serán descartados y no se tendrán en cuenta para el entrenamiento. Estos valores eran inicialmente 0.3 y 1.0, pero existía un número nada despreciable de pares que podrían utilizarse a pesar de su bajo grado de covisibilidad. Por ello, se decidió reducir el umbral mínimo a 0.1. Para evitar utilizar pares demasiado sencillos, el umbral máximo se fijó a 0.99.

#### 4.3.4. *Entrenamiento del extractor de características*

En el modelo original de WarpC+GLUNet se optó por utilizar los pesos preentrenados de ImageNet Krizhevsky et al. (2017) puesto que los resultados que se obtienen tanto si se entrena toda la red como si no, son muy parecidos. En el caso del Endomapper Dataset Azagra et al. (2022) el dominio de las imágenes utilizadas (colonoscopias) difiere respecto de ImageNet, por lo que se ha elegido entrenar la red completa.

### 4.4. *Evaluación experimental*

De cara a comparar el modelo *megadepth* y el *endomapper* Azagra et al. (2022), se han generado 79 pares de imágenes según la Subsección 3.2 y se han ordenado en función de su flujo medio ground-truth. Después, se han separado en tres categorías de forma equitativa: fáciles, medios y difíciles. Para cada test, se ha estimado su flujo mediante el modelo descrito en la Subsección 4.3 y se han calculado sus EPE (Subsección 4.1) a partir de los puntos de los cuales se conoce el flujo sparse ground-truth. A partir de los EPE de los tests, se han calculado los histogramas acumulativos del EPE por cada categoría. Estos pueden ser visualizados en la Figura 6. Como se puede ver, el modelo *endomapper* supera ligeramente al modelo *megadepth* tanto en los tests fáciles como en los intermedios. En los tests difíciles, *endomapper* está por encima de *megadepth* excepto entre los EPEs 75 y 180.

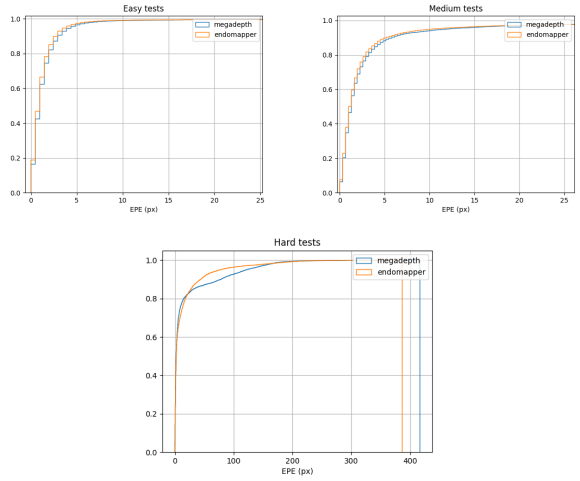


Figura 6: Curvas de los histogramas acumulados de cada categoría de test. En cada uno se muestra para cada valor de EPE el correspondiente porcentaje de puntos que tienen un EPE menor o igual que él.

Por otra parte, se han elegido varios tests de diferentes niveles de dificultad para mostrar el funcionamiento del sistema. Se ha elegido el test que se encuentra en la mediana de cada categoría (fácil mediano, medio mediano y difícil mediano) para evaluar cada uno de los dos modelos. También se ha elegido como test el más fácil y el más difícil. Estos se muestran ordenados en la Figura 7. Se observa que en todos los tests, exceptuando el medio mediano,*endomapper* tiene un mejor desempeño pues el histograma acumulado está siempre por encima del *megadepth*. Se ha conseguido incluso captar aspectos como el giro de las imágenes. Además, se aprecia que los errores máximos son bastante elevados. Esto estará causado por la presencia de espurios en los clústers generados para test (Subsección 3.2).

### 4.5. *Coste computacional*

Se han utilizado computadores diferentes para el entrenamiento del modelo "endomapper":

- CPU: Intel i7-9700K, 8 cores. 32 GB RAM. GPU: 1x NVIDIA TITAN V 12 GB GRAM
- DGX server. CPU: Intel(R) Xeon(R) CPU E5-2698v4, 40 cores, 80 hilos. 500 GB RAM. GPU 8x Tesla V100-SXM2 32 GB GRAM

El tiempo de ejecución del entrenamiento del modelo "endomapper" fue de 3 días en el computador de una única GPU y 2.5 días en DGX (utilizando 2 de sus 8 GPUs). El cálculo del flujo entre 2 imágenes de 1350x1012 tarda 2 segundos en cargar el modelo entrenado y otros 2 segundos en calcular el flujo. Sin embargo, si se calculan más flujos de forma sucesiva, el tiempo de estimación de los mismos será de 1 segundo. Al reducir las imágenes a la mitad de tamaño (675x506 píxeles), el tiempo de carga del modelo no se vio afectado. Sin embargo, el tiempo de cálculo del primer flujo se redujo a 1 segundo y los tiempos de cálculo de flujos consecutivos se redujeron a 0.5 segundos.

## 5. Conclusiones

Este trabajo supone una de las primeras aproximaciones a la aplicación de técnicas de flujo por aprendizaje no supervisado en el dominio de las colonoscopias. Se ha evaluado el método de entrenamiento WarpC Truong et al. (2021) y la red de estimación de flujo GLUNet Truong et al. (2020) en secuencias del Endomapper Dataset Azagra et al. (2022), mostrando un rendimiento sorprendente para relacionar imágenes mediante flujo denso.

El método tiene gran potencial para encontrar correspondencias en colonoscopias, donde otros métodos fallan por completo. Primero se ha evaluado el modelo que había sido entrenado previamente en megadepth (Subsección 4.2), mostrando un desempeño excelente para flujos pequeños y medios. En el trabajo, se ha propuesto un entrenamiento que, usando datos de colonoscopias y una serie de modificaciones, permiten mejorar el funcionamiento ligeramente, especialmente en el caso de que el flujo entre las imágenes sea elevado, donde la mejora es más significativa.

## Agradecimientos

Financiado por el proyecto EU H2020 EndoMapper grant:863146, beca de colaboración Ministerio de Educación y Formación Profesional 22CO1/006926, Proyecto PID2021-127685NB-I00 del Ministerio de Ciencia e Innovación, y por el Gobierno de Aragón mediante el proyecto DGA_T45-17R y el contrato predoctoral de Javier Morlana.

## Referencias

Azagra, P., Sostres, C., Ferrandez, Á., Riazuelo, L., Tomasini, C., Barbed, O. L., Morlana, J., Recasens, D., Batlle, V. M., Gómez-Rodríguez, J. J., Elvira, R., López, J., Oriol, C., Civera, J., Tardós, J. D., Murillo, A. C., Lanas, A., Montiel, J. M. M., apr 2022. EndoMapper dataset of complete calibrated endoscopy procedures.
URL: https://arxiv.org/abs/2204.14240v1

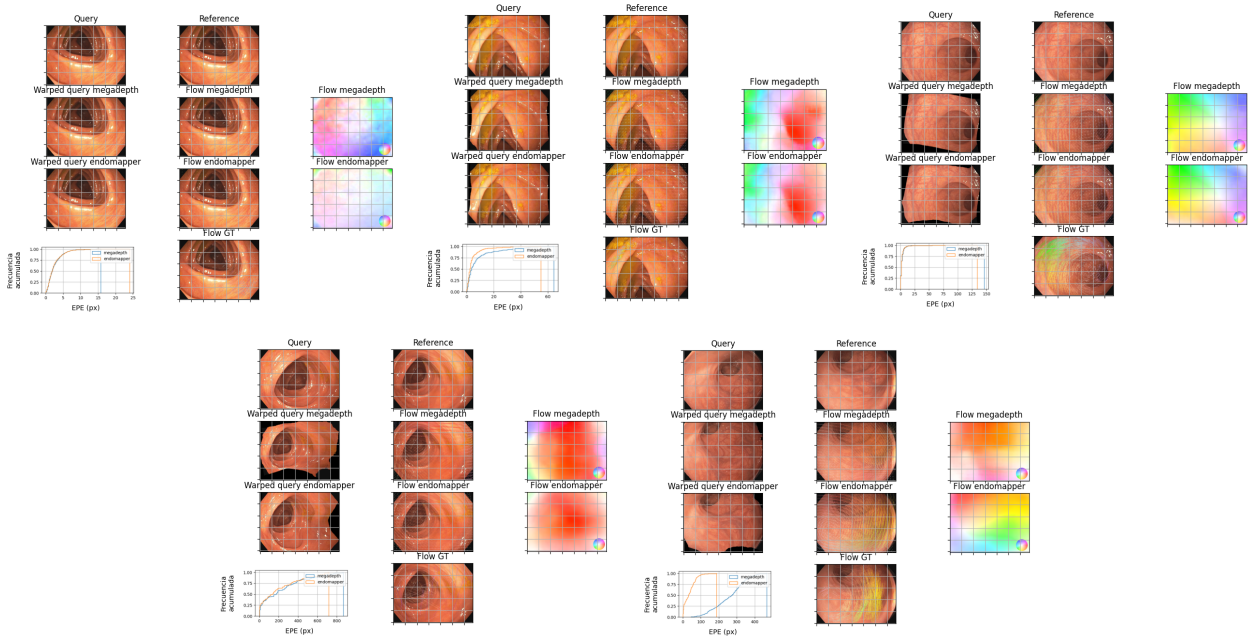


Figura 7: Tests realizados para comparar los modelos *megadepth* y *endomapper*. Los tests son los siguientes (de izquierda a derecha y de arriba a abajo, empezando por el de arriba a la izquierda): el más fácil, el fácil mediano, el medio mediano, el difícil mediano y el más difícil. Para cada test, se muestra en la primera fila, su imagen de query y su imagen de referencia; en la segunda fila, la warped query tanto, el flujo con flechas y el flujo con mapa de color) tanto para el modelo *megadepth* como con el modelo *endomapper*; y en la última fila, se muestran la curva del histograma acumulado del EPE de cada modelo y el flujo sparse ground-truth.

Baker, S., Matthews, I., feb 2004. Lucas-Kanade 20 years on: A unifying framework. International Journal of Computer Vision 56 (3), 221–255.
URL: https://link.springer.com/article/10.1023/B:VISI.0000011205.11775.fd
DOI: 10.1023/B:VISI.0000011205.11775.FD/METRICS

Fortun, D., Bouthemy, P., Kervrann, C., may 2015. Optical flow modeling and computation: A survey. Computer Vision and Image Understanding 134, 1–21.
DOI: 10.1016/J.CVIU.2015.02.008

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., nov 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-January, 1647–1655.
DOI: 10.1109/CVPR.2017.179

Krizhevsky, A., Sutskever, I., Hinton, G. E., may 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60 (6), 84–90.
URL: https://dl.acm.org/doi/10.1145/3065386
DOI: 10.1145/3065386

Li, Z., Snavely, N., 2018. MegaDepth: Learning Single-View Depth Prediction From Internet Photos.
URL: http://www.cs.cornell.edu/projects/

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60, 91–110.

Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S. K., Rosenman, J., Frahm, J.-M., 2021. Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. Medical image analysis 72, 102100.

Morlana, J., Millán, P. A., Civera, J., Montiel, J. M., 2021. Self-supervised visual place recognition for colonoscopy sequences. In: Medical Imaging with Deep Learning.

Morlana, J., Tardós, J. D., Montiel, J., 2023. Colonmapper: topological mapping and localization for colonoscopy. arXiv preprint arXiv:2305.05546.

Rodríguez, J. J. G., Montiel, J. M., Tardós, J. D., 2022. Tracking monocular camera pose and deformation for SLAM inside the human body. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5278–5285.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision. Ieee, pp. 2564–2571.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113.

Teed, Z., Deng, J., 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Lecture Notes in Computer Science 12347 LNCS, 402–419.

Truong, P., Danelljan, M., Timofte, R., 2020. GLU-Net: Global-Local Universal Network for Dense Flow and Correspondences.
URL: https://github.com/PruneTruong/GLU-Net.

Truong, P., Danelljan, M., Yu, F., Van Gool, L., apr 2021. Warp Consistency for Unsupervised Learning of Dense Correspondences. Proceedings of the IEEE International Conference on Computer Vision, 10326–10336.
URL: https://arxiv.org/abs/2104.03308v3
DOI: 10.1109/ICCV48922.2021.01018