



Local Forward-Motion Panoramic Views for Localization and Lesion Detection for Multi-Camera Wireless Capsule Endoscopy Videos

Marina Oliveira^{1,2} ^a and Helder Araujo^{1,2} ^b

¹*Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal*

²*Department of Electrotechnical Engineering, University of Coimbra, Coimbra, Portugal*

Keywords: Computer Vision, Forward-Motion Panorama, Optical Flow, Wireless Capsule Endoscopy.

Abstract: Understanding Wireless Capsule Endoscopy videos is a challenging process since it demands a substantial amount of time and expertise to recognise and accurately interpret findings. The low lesion detection rate with this technology is mainly attributed to the poor image quality of the retrieved frames, the large sets of image data information to process and the time constraints. To overcome these limitations, in this paper, we explore a methodology for constructing local forward-motion panoramic overviews to condense valuable information for lesion detection and localization procedures.

1 INTRODUCTION

Each average eight-hour Wireless Capsule Endoscopy (WCE) video comprises approximately fifty thousand frames (Spyrou et al., 2013). The viewing time with its specific software can take up to several hours of undivided attention for the clinician to accurately detect, distinguish and localize large varieties of gastrointestinal (GI) lesions along the GI tract.

Another main limitation is the poor image quality of the retrieved frames, which results in a detection rate as low as 40% (Spyrou et al., 2013).


Some approaches for reducing reading time and increasing detection rates have already been proposed throughout the years and are mainly focused on the selection of the most representative frames (MRF) for video summarization. Although these solutions are limited, the RAPID Reader Software, for example, allows the view of multiple consecutive frames for clinicians to examine more than one frame in one sitting. The QuickView algorithm also provides a fast-highlighted preview of WCE videos by presenting them with high frame rates in stable image sequences and lower frame rates in regions where sudden changes occur (Spyrou et al., 2013). Another approach called epitomized summarization is able to downsize the number of frames up to 10% by creating epitomes from a compilation of classified frames


based on ground truth data of abnormal tissue and/or artefacts (Iakovidis et al., 2010).

Another main limitation of this technology is the limited field of view. The viewing angle of a camera is determined by the amount of light that crosses its lens (Třebický et al., 2016). A small viewing angle provides the clinician with a small viewing domain for the examination, which also lowers the detection rate (Swain, 2003). The viewing angles of current commercial camera capsules vary from 140 to 170 degrees (Brown and Jayatissa, 2020). Recent capsules intend to overcome this limitation by offering more than one camera in one capsule, for example, two cameras at opposing ends.

In order to offer a broader field of view and an increased area of analysis without hardware alterations, the construction of a panoramic image of the full trajectory or a panoramic image of local regions of interest could be a solution. Ultimately, by taking advantage of the fact that the multiple cameras of the same capsule register the same tissue structures, a more robust panoramic result could be achieved. A panorama of the full trajectory or multiple local panoramic views in regions of interest could provide an overview of the GI tract, lower the viewing time and optimize lesion detection rates.

The construction of a full panorama involves the reconstruction of the specific surface geometry and the corresponding motion estimation (Yoshimoto et al., 2020). The chosen approach for the procedure may vary depending on how the video or the

^a  <https://orcid.org/0000-0001-9271-0357>

^b  <https://orcid.org/0000-0002-9544-424X>

set of sequential images are obtained since it determines the geometry of the problem (Cao et al., 2018). In most cases of panoramic view construction, the dataset video is recorded with a single camera that undergoes pure rotation motion around its optical axis. In other cases, the data is a set of images with overlapping domains acquired from multiple cameras that are then stitched into a wide-view panorama. In cases such as the ones with endoscopic capsules, where the videos are acquired by one or more cameras from a capsule that moves ideally with pure translation along the optical axis, the goal is to identify and extract the overlapping radial domains from each frame and stitch them to obtain a forward-motion panorama. (Cao et al., 2018).

In this paper, we explore a methodology to overcome the above-mentioned limitations through the construction of a local panoramic view. We present the steps towards a robust solution taking into account the geometry of this problem and the specificity that comes with working with WCE videos.

1.1 Paper Organization

In the related work section, we survey several attempts to construct panoramas given a forward-motion camera inside a tubular structure, both in the medical imaging field and other fields with a similar problem geometry. In the experimental dataset section, we discuss the initial set of multi-camera WCE videos and the reasons behind the need to create a synthetic model for this problem. In the methodology section, we present the principles that our procedure was based on, describe the concrete steps we took to build a local forward-motion panorama and we present the evaluation metrics chosen to assess the result. Then, in the results and discussion section, we present the output of each step of the previously described methodology and consider the alternative paths we could have chosen. The conclusion provides a brief overview of the work developed in this paper and, most importantly, the future work that can be developed from the first steps taken by our approach.

2 RELATED WORK

Previous approaches for the generation of panorama images of tubular-shaped organs focus on 2D images from the oesophagus (Kim et al., 1995)(Seibel et al., 2008). These methods use a tubular model such as cylindrical projection because the capsule moves along the optical axis. The 2D frames are unwrapped around a previously computed centre of projection.

Then, given the camera motion estimation between sequential frames with an optical flow approach, the projections were mapped into a cylindrical surface.

Behrens et al. developed an image mosaicing algorithm for local panorama construction from bladder video sequences in fluorescence conventional endoscopy. The image information was extracted and an affine parameter model with iterative optimization was adopted to determine the best image transform given mean squared error measurements. Some visual artefacts were inevitably produced by non-homogeneous lighting and were compensated in the stitching step with a mutual linear interpolation function (Behrens, 2008). Five years later, *Spyrou et al.* proposed an approach that presents an automatically assembled visual summary using WCE videos based on the idea of pipe projection proposed by *Rouso et al.* The frames are geometrically transformed with feature matching techniques and stitched together to construct a panoramic image. Ideally, the construction of a panoramic image enables the viewing of multiple frames simultaneously and provides a broader field of view without information loss (Spyrou et al., 2013).

Given the similarity in geometry, some geological engineering approaches to study structural characteristics and spatial distribution patterns of fissures in rock masses may be valuable to consider. As with capsule endoscopy, the feature extraction process in these images is challenging since the datasets, Axial View Panoramic Borehole Televiewer (APBT) videos, also have poor image quality. The probe used for the forward-motion along the tube generates a slight rotation so its trajectory is not strictly translational along the central axis of the borehole. Cao et al. proposed, in 2018, an approach for the construction of an unfolded image of a borehole from APBT videos. Firstly, an algorithm for the automatic location of the centre is based on the circularity of annular borehole images, then the annular image sequences are unfolded with Daugman's rubber sheet model (RSM) and an interpolation algorithm. The unfolded image sequences are then fused to generate an unfolded panoramic image with a projection registration algorithm (Cao et al., 2018).

K. Yoshimoto et al. developed a prototype stereo endoscopy with a compound eye system named Thin Observation Module by Bound Optics (TOMBO) which allows for the depth mapping of each point from the 2D frames to produce 3D data. Later, *K. Yoshimoto et al.* proposed a procedure to acquire 3D panorama images of the oesophagus from conventional endoscopy. The methodology comprised the acquisition of a sequential set of frames from the GI

tract with an endoscope, the reconstruction of the corresponding 3D surfaces, and the estimation of its position using scene flow and surface merging. This approach improves the quality of the frames by reducing the number of missing points from low resolution and stereo matching failures. The method was initially validated with a phantom for the size estimation of the texture and the moving distance and later with a pig oesophagus (Yoshimoto et al., 2020).

3 EXPERIMENTAL DATASET

The goal is to obtain the local panoramic views from the consecutive frames of WCE videos from patients with Crohn’s disease obtained specifically from the multi-camera capsule PillCam Colon2.

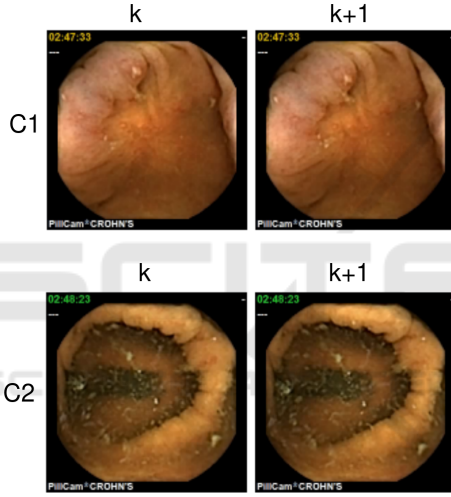


Figure 1: Example sample of sequential frames (k and $k+1$) from a patient’s exam video, obtained with the front and back camera (C1 and C2) of a PillCam Colon2 capsule.

At this stage, a synthetic dataset for an initial proof of concept was constructed in order to provide ground truth information regarding motion. For this reason, a colon-like texture tubular model was created with the Blender Software given a few restrictions.

The model which consisted of a hollow tubular structure was created with the projection of textured portions of the images from the above-mentioned videos on the inside. A straight path was also created passing through the inside of the tube for the construction of the animation. Two cameras were coupled but placed facing opposing directions and added at one end of the tube. Each camera was coupled with a light source as to follow the pre-defined path. In this way, as with capsule endoscopy videos, the animation of the purely translational displacement of each camera

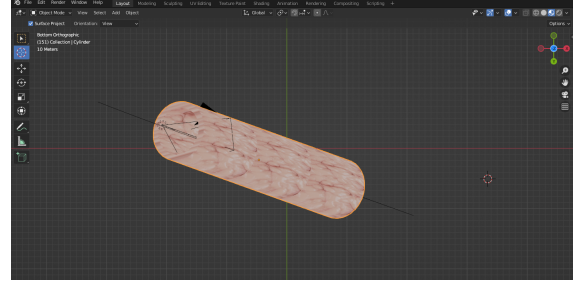


Figure 2: Tubular model, created with the Blender Software, with colon-like textured walls, two opposing cameras coupled with light sources animated along a pure translation trajectory.

is rendered with lighting condition variations as a result of a light source that accompanies the motion of the camera. This model allowed us to divide the entire procedure of building a panoramic view into smaller steps.

4 METHODOLOGY

Although we can simplify the geometry of the problem by thinking of the interior of the colon as a cylindrical tube, unlike the conventional endoscopic or geological exploration probes mentioned in the related work section, the capsule does not move in a purely translational way along the optical axis. As we can see in Figure 3, when the capsule is in pure translation motion along its optical axis, the displacement of the pixels in sequential frames is radial, pointing outwards during forward motion and inwards during backward motion. On the other hand, when the capsule undergoes pure rotation around its optical axis, the displacement of pixels from sequential frames can be described as clockwise/counterclockwise motion.

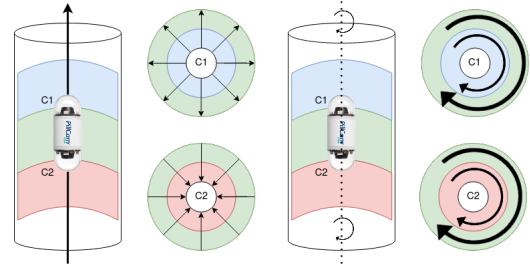


Figure 3: Relationship between the motion of the camera along the tubular-like organ (pure translation represented on the left and pure rotation on the right) and the corresponding pixel displacement between sequential frames for both front and back camera.

Since the movement of the capsule inside the colon is the result of the peristaltic movements of

the tissue, its displacement will be composed of periods of pure translation along the optical axis, periods of pure rotation around its optical axis and periods when it translates and rotates at the same time. Given the geometry of the problem, since the sequential frames associated with the pure rotation of the capsule around its optical axis do not offer additional information for the panoramic view, we are only interested in consecutive frames associated with pure translation motion.

Since forward motion and zoom can be handled well with the generalized pipe representation proposed by *Rousso et al.* (Rousso et al., 1998), shown in Figure 4, our approach is also based on those core principles with a few constraints and adaptations given the specificity of our datasets.

4.1 Generalized Pipe Representation

To transform the representation from the radial displacement of the image pixels into parallel displacement, it is possible to project the 2D planar image onto a 3D cylinder, with pipe projection.

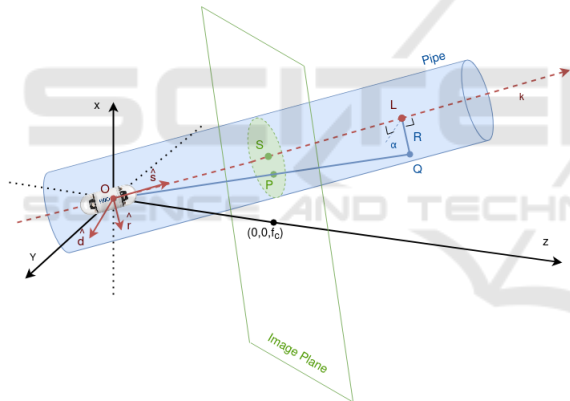


Figure 4: Diagram illustrating the projection of a 2D planar image (image plane) onto a 3D cylinder (pipe): $s = S/|S|$ is the axis of the pipe; R is the radius; $O = (0, 0, 0)$ is the optical center; $S = (s_x, s_y, f_c)$ is the focus of expansion; f_c is the focal length and Q is the projection on the pipe of each point $P = (x, y, f_c)$ from the plane.

The axis of the pipe $s = S/|S|$ is chosen so that it passes through the optical center $O = (0, 0, 0)$ and the focus of expansion (FOE) $S = (s_x, s_y, f_c)$, with f_c as the focal length (Rousso et al., 1998).

Each point Q is the projection of each original point $P = (x, y, f_c)$, distanced from the axis s by the radius R of the pipe, and collinear with both P and O .

Given k as the position along the axis \hat{s} , with \hat{d} and \hat{r} as unit vectors chosen to form a cartesian coordinate system together with \hat{s} and α as the angle from \hat{d} , the 3D position of a point Q on the pipe is expressed in

Equation (1) (Rousso et al., 1998).

$$Q = (Q_x, Q_y, Q_z) = k\hat{s} + R\cos(\alpha)\hat{d} + R\sin(\alpha)\hat{r} \quad (1)$$

Since the pixels whose original distance from s is less than R become magnified and the pixels with greater distance than R shrink once projected on the pipe, selecting the radius as $R = \sqrt{f_c^2 + (\frac{w}{2})^2 + (\frac{h}{2})^2}$, where w is the width and h is the height of the image, preserves the geometry and resolution of the image. The resolution decreases as $|Q_z - f_c|$, so it is best preserved around the intersection of the pipe with the image plane ($Q_z = f_c$) (Rousso et al., 1998).

4.2 Pipe Mosaicing

Given two corresponding points $P_k = (x_k, y_k)$ in image I_k and $P_{k+1} = (x_{k+1}, y_{k+1})$ in images I_{k+1} , the flow vector (u, v) is a function of the position (x_k, y_k) . The scanning broom chosen for the mosaicing process must be a curve $F(x, y) = 0$ perpendicular to the optical flow and as close as possible to the centre of the image in order to minimize lens distortion.

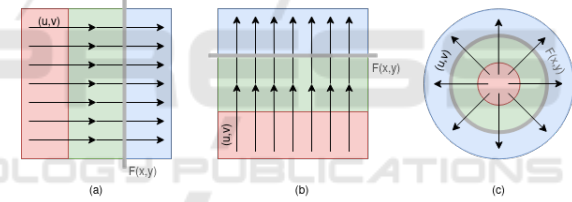


Figure 5: Given OF as a function of the position, the scanning broom $F(x, y) = 0$ chosen for the mosaicing process must be perpendicular to (u, v) : (a) $F(x, y)$ is a vertical straight line for uniform horizontal OF; (b) $F(x, y)$ is a horizontal straight line for uniform vertical OF; (c) $F(x, y)$ is a circumference centred around the FOE for radial OF.

If the optical flow is horizontal, $F(x, y)$ is a vertical straight line, as shown in Figure 5a); with vertical optical flow, $F(x, y)$ is a horizontal straight line, as shown in Figure 5b); and if the optical flow represents zoom or forward motion, $F(x, y)$ is a circumference around the centre of the FOE, as shown in Figure 5c) (Rousso et al., 1998).

4.3 Coordinate System Transformation

Since straight optical flow and straight scanning brooms are simpler to operate during the mosaicing process, the sequential frames were converted from cartesian coordinates to polar coordinates, as shown in Figure 6. Given a known FOE and the relationship between cartesian and polar coordinates, shown in Equation (2), a rectangular image can be converted

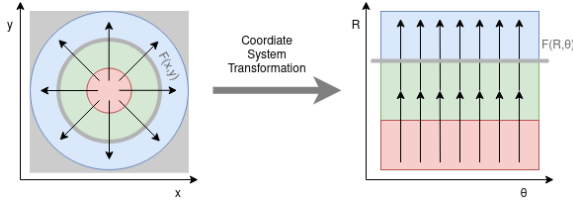


Figure 6: Coordinate system conversion from cartesian to polar image coordinates (with a central FOE) in order to obtain frames with a straight optical flow pattern and apply a straight scanning broom in the mosaicing process.

into a polar image with radius r and angular coordinate θ .

$$\begin{cases} x = r \cos(\theta) \\ y = r \sin(\theta) \end{cases} \quad (2)$$

For this case, each channel of each RGB frame is converted from cartesian to polar coordinates with a bilinear interpolation algorithm to interpolate between points that do not exactly lay in the image. Then, each channel converted to polar coordinates is coupled to obtain the final RGB polar frame. This process is repeated for each sequential frame, as shown in Figure 7.

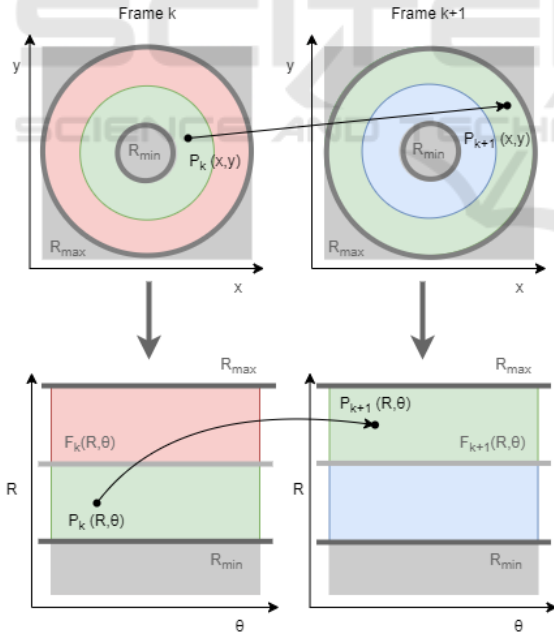


Figure 7: Diagram illustration of the difference between a point correspondence between two sequential cartesian frames and the corresponding two sequential polar frames after the coordinate system conversion (with a central FOE).

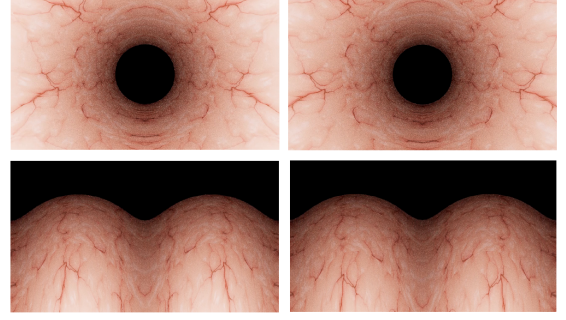


Figure 8: Sample example of two consecutive frames rendered from the colon-like synthetic blender model and the corresponding two sequential polar frames after the coordinate system conversion with a central FOE.

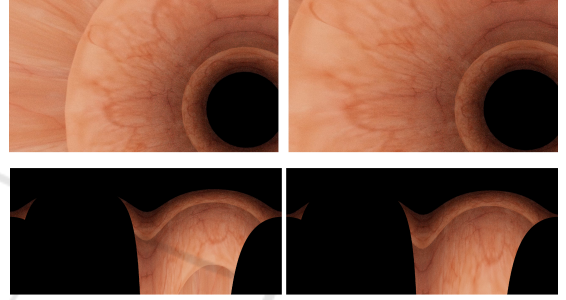


Figure 9: Sample example of two consecutive frames rendered from the colon-like synthetic blender model and the corresponding two sequential polar frames after the coordinate system conversion with a non-central FOE.

4.4 Image Registration from Optical Flow with the RAFT Network

In our previous assessment (Oliveira et al., 2021), from the approaches explored to find robust correspondences between consecutive frames of WCE videos, we found that the best results were obtained with matches computed from the optical flow results estimated with deep learning. For this paper, we used the RAFT network (Teed and Deng, 2021) to compute the flow vectors to determine the matches for the image registration process between sequential frames of the video.

Optical flow is the result of the per-pixel motion estimation between video frames. Given a pair of sequential frames, which in this case is a pair of image frames previously converted into polar images, a dense displacement field maps each pixel in one frame to the corresponding coordinates in another.

This energy minimization problem sets a trade-off between data and regularization terms. Existing solutions still offer limitations regarding occlusions, low-textured surfaces, fast-moving objects and motion blur. Unlike traditional approaches, with deep

learning, features and motion priors are learned instead of handcrafted. This problem is not yet closed since the design of architectures with faster and easier training procedures, better performances and adequate generalization capabilities is still a necessity.

The Recurrent All-Pairs Field Transforms (RAFT) is an optimization-based deep network architecture created to obtain more robust optical flow estimates. The three major components of the RAFT network are the feature encoder, which extracts per-pixel features per pixel using a convolution network, a correlation layer, which calculates pixel similarity to produce correlation volumes for all pairs of pixels, and a recurrent update operator, which is essentially an iterative optimization algorithm that selects values from those volumes and updates the flow field. (Teed and Deng, 2021) For this case, from each pair of consecutive po-

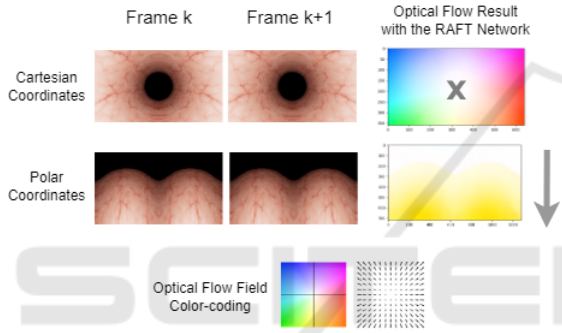


Figure 10: Sample example of two consecutive frames rendered from the colon-like synthetic blender model and the corresponding two sequential polar frames after the coordinate system conversion with a non-central FOE.

lar frames, given the subset of points P_k^i associated with each pixel i from frame k , given the the optical flow output (u_k^i, v_k^i) from the RAFT Network, the corresponding set of P_{k+1}^i points coordinates in the consecutive frame were computed. This correspondence of points is performed using a linear interpolation algorithm, in order to be used as matching points for the final panorama mosaicing.

4.5 Homography Matrix

Projective geometry studies the properties of a projective plane IP^2 given a set of invertible linear transformations of homogeneous coordinates that map lines to lines. Under the mapping $h: IP^2 \rightarrow IP^2$, if three points x_1, x_2 and x_3 lie on the same line, then $h(x_1), h(x_2)$ and $h(x_3)$ lie on the same line, thus preserving collinearity (Hartley and Zisserman, 2003). A mapping is a projectivity if and only if there exists a non-singular 3×3 matrix H such that for any point in

IP^2 represented by a vector x it is true that

$$h(x) = Hx \quad (3)$$

where H is the homography matrix. From the nine elements of H , only eight ratios are independent, so this transformation has eight degrees of freedom (Hartley and Zisserman, 2003).

$$H_{3 \times 3} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \quad (4)$$

For this paper, for each pair of purely translational sequential polar frames, given the correspondences (P_k, P_{k+1}) from the optical flow estimates with the RAFT Network, a global homography H is obtained. The computation is performed with Singular Value Decomposition (SVD) and the random sample consensus (RANSAC) algorithm for outlier removal.

$$P_{k+1} = HP_k \quad (5)$$

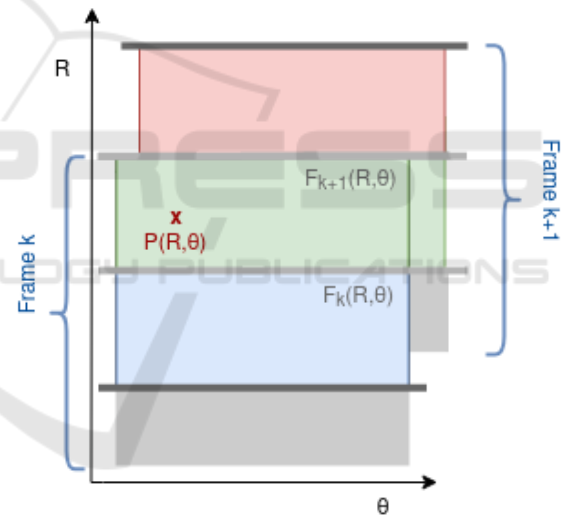


Figure 11: Ideal mosaicing result given a perfect overlay between the two corresponding points in both frame k and $k+1$.

The final mosaic can be obtained by warping the set of strips retrieved from the sequential images given the point correspondences (Rousso et al., 1998). Each strip must be warped to match the boundaries of the previous stitched strips.

In our case, from the frame k , the strip bounded by the two curves $F_k(r, \theta) = 0$ and $F_{k+1}'(r, \theta) = 0$, as shown in Figure 11, ensuring the continuity and non-redundancy of the information, as the orthogonality of the optical flow is assured. (Rousso et al., 1998).

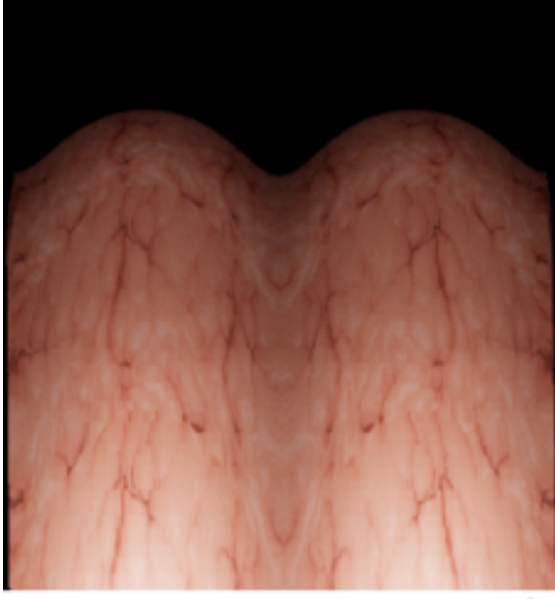


Figure 12: Final local forward-motion panorama obtained from the mosaicing of all 30 pairs of consecutive polar frames given one video rendered from the tubular Blender model.

4.6 Image Registration Evaluation

1. **Sum of Squared Differences (SSD):** Given the assumption that the only difference between two registered images is Gaussian noise, the accuracy of the registration method can be evaluated by the intensity difference of the registered image pair, for example, with the sum of squared differences. SSD is sensitive to smaller samples with large intensity differences (Song, 2017). In the case of a forward-motion panorama, each image is registered to its previous one. For a transformation $h_{i,i-1}$ that maps a point in image I_i to point x in I_{i-1} , where $R = [x_1, x_2, \dots, x_N]$ is a subset of points in I_{i-1} , SSD can be computed, over the region R , as shown in Equation 6. For an ideal set of registrations, SSD will equal zero.

$$SSD_{i,R} = \sum_{x=x_1}^{x_N} (I_i(h_{i,i-1}(x)) - I_{i-1}(x))^2 \quad (6)$$

2. **Intensity Variance (IV):** When a registration method performs well, the registered image is as close as possible to the target image and its average intensity image is the sharpest (Song, 2017). To measure the sharpness of the average intensity image is to compute the intensity variance of the registered images. Given a transformation $h_{i,i-1}$ that maps a point in image I_i to point x in I_{i-1} , where $R = [x_1, x_2, \dots, x_N]$ is a subset of points

in I_{i-1} , the IV of image I_i registered to image I_{i-1} , over the region R , is computed as expressed in Equations 7 and 8. For an ideal registration, IV will be equal to zero.

$$IV_{i,R}(x) = \sum_{x=x_1}^{x_N} (I_{i-1}(h_{i-1,i}(x)) - ave_i(x))^2 \quad (7)$$

$$ave_i(x) = \frac{1}{N} \sum_{x=x_1}^{x_N} I_i(h_{i-1,i}(x)) \quad (8)$$

3. **Correlation Coefficient (CC):** Assuming that the intensity relationship between two registered images is linear, the correlation coefficient measures its linear dependence (Song, 2017). With a transformation $h_{i,i-1}$ that maps a point in image I_i to point x in I_{i-1} , where $R = [x_1, x_2, \dots, x_N]$ is a subset of points in I_{i-1} , the CC of an image I_i registered to image I_j , over the region R , can be computed with Equations 9-11. The ideal CC, given a pair of perfectly registered images, is equal to one.

$$d_{i-1}(x) = I_{i-1}(h_{i-1,i}(x)) - \overline{I_{i-1}} \quad (9)$$

$$d_i(x) = I_i(x) - \overline{I_i} \quad (10)$$

$$CC_{i,R} = \frac{\sum_{x=x_1}^{x_N} d_{i-1}(x)d_i(x)}{\sqrt{\sum_{x=x_1}^{x_N} d_{i-1}(x)^2 \sum_{x=x_1}^{x_N} d_i(x)^2}} \quad (11)$$

5 IMAGE QUALITY ASSESSMENT

Image Quality Assessment (IQA) aims to quantify the quality of an image in terms of human perception. IQA algorithms are commonly used for the quality assessment of compressed images. No-Reference (NR) or blind IQA, focus on the estimation of the quality of a degraded or newly generated image given no ground truth reference image or even the type of processing the image is subjected to (Madhusudana et al., 2022), which is the case with our local forward-motion panorama.

A few CNN-based NR-IQA models have already been developed in the past years, such as CN-NIQA (Madhusudana et al., 2022), CONTRastive Image Quality Evaluator (CONTRIQUE) (Kang et al., 2014), and VIDGIQA (Guan et al., 2017).

For all three above-mentioned models, the higher the value, the higher the quality of the image. In order to have a quantitative evaluation of our final local forward-motion panorama obtained after the polar image stitching process, these NR-IQA models were used and their values were computed for our image.

6 MOTION SEGMENTATION OF MULTI-CAMERA WCE VIDEOS

To simplify the complex motion pattern of the capsule in a real scenario and start with a simpler problem, we will assume that rotation and translation do not occur simultaneously. In the specific case of our patient dataset from the multi-camera capsule PillCam Colon2, we have two cameras at opposing ends, so when one of the cameras is performing a pure translational forward-motion, the other camera is undergoing a pure translational backward motion.

In sequential frames, given the same time interval, pixels will move radially for both cameras but their displacement vectors will point inwards for one of the cameras and outwards for the other. For this reason, by analyzing the pixel displacement between sequential frames from both cameras at the same instant it is possible to extract from the full video the frames that correspond to the pure translation motion that allows us to obtain a local panorama.

For this analysis, the optical flow between sequential frames from both cameras of the same capsule was obtained with the RAFT network in order to choose pure translational motion segments (Teed and Deng, 2021).

7 RESULTS AND DISCUSSION

Figure 8 shows a sample pair of consecutive synthetic frames obtained from the video rendered from the Blender model and each corresponding polar representation after the coordinate system conversion, presented in Equation 2, assuming a central FOE. Figure 9 shows another pair of two consecutive frames from another video rendered from the Blender model with a non-central but known FOE. In order to apply this pipeline to a real-case scenario with the multi-camera capsule images, a methodology needs to be developed for the computation of the FOE in cases that are non-central.

The optical flow result obtained with the RAFT network given the polar representation of the consecutive frames is shown in Figure 10. The output colour is all yellow, corresponding to vertical flow field as expected since the displacement of the capsule is designed to be purely translational in our Blender model. This step provides the image registration estimation since it allows the retrieval of the pixel-by-pixel point correspondences with each optical flow vector associated with each pixel. In addition, it also provides a conceptual validation of our initial idea for the ideal case scenario where the endoscopic capsule

only moves with pure translation, since all displacement vectors point downwards.

Figures 13-15 show the values of the metrics SSD, IV and CC between each image I_{k+1} and the previous one I_k for each one of the intensities of each RGB channel and its grayscale, for all 30 pairs of consecutive frames assessed from the rendered video.

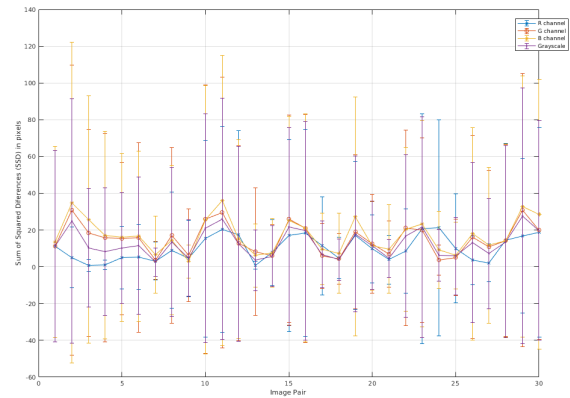


Figure 13: SSD between each image I_{k+1} and its previous one I_k for the intensities of each RGB channel and for grayscale, for all 30 pairs of consecutive frames from the video rendered with the Blender Software.

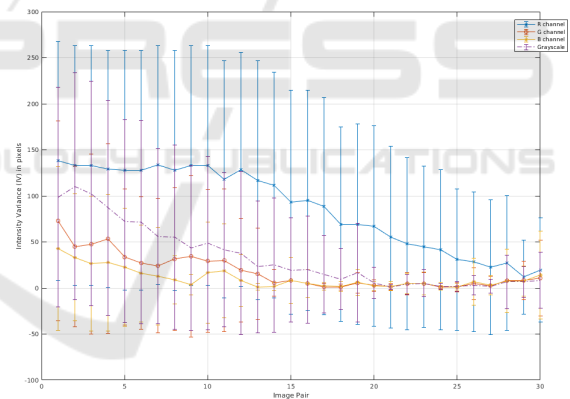


Figure 14: IV between each image I_{k+1} and its previous one I_k for the intensities of each RGB channel and for grayscale, for all 30 pairs of consecutive frames from the video rendered with the Blender Software.

Figure 11 shows the final panorama obtained after the mosaicing process of all 30 pairs of consecutive frames from the forward-motion video with the blender model.

Table 1 shows the normalized NR-IQA values for the final local panorama from Figure 11 given the three above-mentioned CNN-based models.

Since the registration results are satisfactory but the IQA results from the final panorama are far from ideal, there needs to be further exploration of a more robust methodology for the mosaicing process, for ex-

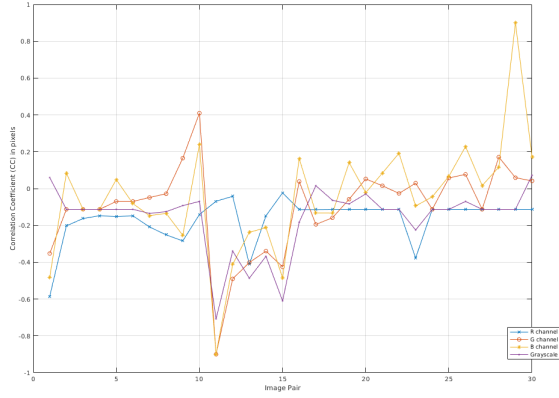


Figure 15: CC between each image I_{k+1} and its previous one I_k for the intensities of each RGB channel and for grayscale, for all 30 pairs of consecutive frames from the video rendered with the Blender Software.

Table 1: NR-IQA normalized values ([0,1]) of the final local forward-motion panorama using the CNNIQA (Madhusudana et al., 2022), CONTRIQUE (Kang et al., 2014) and VIDGIQA (Guan et al., 2017) models.

CNNIQA	CONTRIQUE	VIDGIQA
0.267	0.478	0.114

ample, with the computation of several local homographies instead of a global one or by exploring non-classical deep learning tools for the iterative stitching. Further research work is also required in order to deal with the discontinuities where the boundaries of each strip are visible, creating artefacts that lower the image quality.

Figure 16 shows a sequential pair of frames retrieved from a WCE video from a patient with the PillCam Colon2. Both frames are from the two cameras on opposite sides of the capsule (C1 and C2), in cartesian and polar coordinates, and the corresponding optical flow estimation from the polar representation with the RAFT network. A robust metric for the comparative analysis of the OF vector field of consecutive frames from the opposing end cameras (C1 and C2) is also needed to use as a segmentation criterion for the motion segmentation process. Following this methodological line of work, given the fact that the purely translational frames correspond to backward and forward motion from both back and front cameras, which are rigidly connected and that both correspond essentially to a vertical OF vector field in polar coordinates, as shown in Figure 16, our future goal is the computation of these local panoramas with the patient videos.

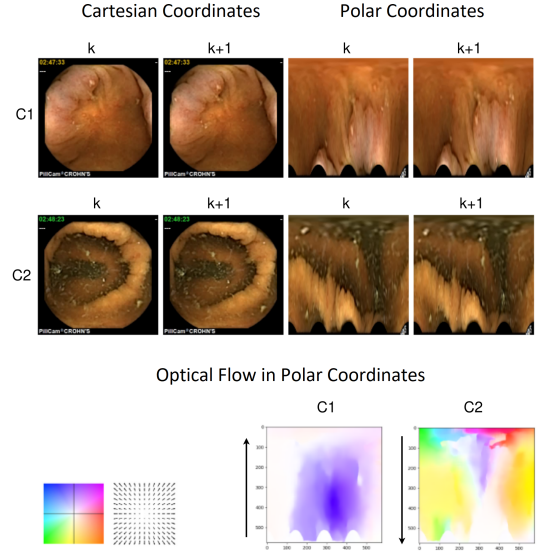


Figure 16: Example of a sequential pair of frames from the cameras on both ends of the capsule (C1 and C2) and the optical flow estimation given the polar representation with the RAFT network (Teed and Deng, 2021).

8 CONCLUSIONS

The work developed in this paper constitutes a crucial step for the development of local endoscopic panoramas to reduce the viewing time of clinicians and increase lesion detection rates.

Although the methodology can still be improved, as described above, a proof of concept for the construction of a motion-forward panorama was successfully carried out given the colon-like textured tubular model frames. Given this starting point, after a motion segmentation process, local endoscopic panoramas could be obtained with patient videos, in regions of interest, where the capsule moves in pure translation, as proposed. With this multi-camera capsule patient video, by constructing a local panoramic view from each camera (C1 and C2), since they travel the same path and observe the same structures, it may also be possible, in future work, to condense both local panoramas into a more robust one.

In the future, if there is also the need to create a global panoramic overview of the entire GI tract, in order to avoid discontinuities in regions where no transformation can be computed between consecutive frames, the generation of intermediate views for the mosaicing step can also be explored.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of project PTDC/EMD-EMD/28960/2017, entitled "Multi-Cam Capsule Endoscopy Imagery: 3D Capsule Location and Detection of Abnormalities", funded by FCT, the PhD Scholarship 2020.06592.BD funded by FCT, and the Institute of Systems and Robotics - University of Coimbra, under project UIDB/0048/2020, funded by FCT.

REFERENCES

- Behrens, A. (2008). Creating panoramic images for bladder fluorescence endoscopy. *Acta Polytechnica*, 48.
- Brown, A. P. and Jayatissa, A. H. (2020). Analysis of current and future technologies of capsule endoscopy: A mini review. *Archives of Preventive Medicine*, 5(1):031–034.
- Cao, M., Deng, Z., Rai, L., Teng, S., Zhao, M., and Collier, M. (2018). Generating panoramic unfolded image from borehole video acquired through APBT. *Multi-media Tools and Applications*, 77(19):25149–25179.
- Guan, J., Yi, S., Zeng, X., Cham, W. K., and Wang, X. (2017). Visual Importance and Distortion Guided Deep Image Quality Assessment Framework. *IEEE Transactions on Multimedia*, 19(11):2505–2520.
- Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press.
- Iakovidis, D., Tsevas, S., and Polydorou, A. (2010). Reduction of capsule endoscopy reading times by unsupervised image mining. *Computerized Medical Imaging and Graphics*, 34(6):471–478. Biomedical Image Technologies and Methods - BIBE 2008.
- Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1733–1740.
- Kim, R., Baggott, B. B., Rose, S., Shar, A. O., Mallory, D. L., Lasky, S. S., Kressloff, M., Faccenda, L. Y., and Reynolds, J. C. (1995). Quantitative endoscopy: Precise computerized measurement of metaplastic epithelial surface area in barrett's esophagus. *Gastroenterology*, 108(2):360–366.
- Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. (2022). Image Quality Assessment Using Contrastive Learning. *IEEE Transactions on Image Processing*, 31:4149–4161.
- Oliveira, M., Araujo, H., Figueiredo, I. N., Pinto, L., Curto, E., and Perdigoto, L. (2021). Registration of consecutive frames from wireless capsule endoscopy for 3d motion estimation. *IEEE Access*, 9:119533–119545.
- Rousso, B., Peleg, S., Finci, I., and Rav-Acha, A. (1998). Universal mosaicing using pipe projection. *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952.
- Seibel, E. J., Carroll, R. E., Dominitz, J. A., Johnston, R. S., Melville, C. D., Lee, C. M., Seitz, S. M., and Kimmey, M. B. (2008). Tethered capsule endoscopy, a low-cost and high-performance alternative technology for the screening of esophageal cancer and Barrett's esophagus. *IEEE Transactions on Biomedical Engineering*, 55(3):1032–1042.
- Song, J. H. (2017). *Methods for evaluating image registration*. The University of Iowa.
- Spyrou, E., Diamantis, D., and Iakovidis, D. K. (2013). Panoramic visual summaries for efficient reading of capsule endoscopy videos. pages 41–46.
- Swain, P. (2003). Wireless capsule endoscopy. *Gut*, 52(suppl 4):iv48–iv50.
- Třebický, V., Fialová, J., Kleisner, K., and Havlíček, J. (2016). Focal length affects depicted shape and perception of facial images. *PLOS ONE*, 11(2):1–14.
- Teed, Z. and Deng, J. (2021). RAFT: Recurrent All-Pairs Field Transforms for Optical Flow (Extended Abstract). *IJCAI International Joint Conference on Artificial Intelligence*, pages 4839–4843.
- Yoshimoto, K., Watabe, K., Tani, M., Fujinaga, T., Iijima, H., Tsujii, M., Takahashi, H., Takehara, T., and Yamada, K. (2020). Three-dimensional panorama image of tubular structure using stereo endoscopy. *International Journal of Innovative Computing, Information and Control*, 16(3).