IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 31, 2021, accepted August 13, 2021, date of publication August 27, 2021, date of current version September 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108234

Registration of Consecutive Frames From Wireless Capsule Endoscopy for 3D Motion Estimation

MARINA OLIVEIRA[1,2], HELDER ARAUJO[1,2], (Member, IEEE), ISABEL N. FIGUEIREDO[3], LUÍS PINTO[3], EVA CURTO[1,2], AND LUÍS PERDIGOTO[4]

[1]Department of Electrical and Computer Engineering (DEEC), Faculty of Sciences and Technology, University of Coimbra, 3004-531 Coimbra, Portugal
[2]Institute of Systems and Robotics, University of Coimbra, 3004-531 Coimbra, Portugal
[3]Department of Mathematics (CMUC), Faculty of Sciences and Technology, University of Coimbra, 3004-531 Coimbra, Portugal
[4]ESTG, Polytechnic Institute of Leiria, 2411-901 Leiria, Portugal

Corresponding author: Marina Oliveira (marina.oliveira@student.uc.pt)

This work was supported in part by the Research Project PTDC/EMD-EMD/28960/2017, entitled "Multi-Cam Capsule Endoscopy Imagery: 3D Capsule Location and Detection of Abnormalities" and the Ph.D. Scholarship 2020.06592.BD through the Fundação para a Ciência e a Tecnologia (FCT), Portugal, in part by the Institute of Systems and Robotics—University of Coimbra (ISR-Coimbra) under Grant UIDB/0048/2020, and in part by the Centre for Mathematics, University of Coimbra (CMUC), funded by the Portuguese Government through FCT/Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) under Grant UIDB/00324/2020.

ABSTRACT Wireless Capsule Endoscopy (WCE) is a non-invasive medical procedure devised for painless *in vivo* inspection of the gastrointestinal (GI) tract. It is especially valuable for the examination of the small intestine since it is difficult to reach by traditional endoscopic procedures. The setup includes a camera with an embedded light source and a circuit capable of acquiring and transmitting the video. The main challenge of this technology is the identification of the position and trajectory of the capsule as it travels through the GI tract, which is particularly relevant during the detection of anomalies in the tissue. Given only the information provided by the recorded images, it is possible to estimate the 3D motion of the camera capsule and provide a full trajectory reconstruction. A critical yet difficult step in this process is the image registration between sequential frames. Therefore, being able to determine accurate correspondences between points, regions or features in two consecutive frames is crucial for the computation of the relative rotation and translation of the capsule. This paper comprises a comparative assessment of methodologies to address this problem with a porcine colon dataset obtained with our experimental setup.

INDEX TERMS Capsule movement, deep leaning, feature extraction, image-based localization, optical flow, wireless capsule endoscopy.

I. INTRODUCTION

Endoscopic capsules are currently used for a variety of medical exams for the inspection of the full length of the GI tract and constitute a non-invasive approach without the risks involved with the sedation process and the risk of perforation from standard endoscopes. This is a very attractive gastroenterology alternative exam that is especially relevant for patients who require repeated inspection at regular intervals [1] and particularly for the examination of the small intestine which is not easily reached with other conventional endoscopic procedures [2]. During a standard WCE exam, the patient ingests a capsule that travels along the GI tract moved by peristalsis. Each capsule is equipped with light-emitting diodes and one or more cameras that acquire a sequential set of frames that is then transmitted to a recorder. The frames are low-resolution images affected by significant geometric and radiometric distortion due to the small-sized lenses and poor lighting conditions. These images are inspected by a clinician to identify and locate possible lesions such as polyps and ulcers [2].

One of the difficulties of the WCE procedure is the length of reading and video reporting time. The average video report time is 30 to 60 minutes depending on the trajectory of the capsule along the GI tract and pathology in question. An extra challenge is faced with the risk of missing particular pathologies including indiscreet mucosal bulges given the complex nature of lesions and the fatigability of the human

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

eye. Another complication regarding WCE is the possibility of retention which requires endoscopic or surgical intervention to retrieve the capsule. The retention rate is 1.2–2.1% in patients with suspected small bowel bleeding; 2.35% for suspected small bowel Crohn's disease; 4.63% for established small bowel Crohn's disease; 2.2% for patients with abdominal pain and diarrhoea; and 2.1% in patients with neoplastic lesions [1]. Lastly, one of the main drawbacks of the WCE technique is the lack of information regarding both the position and orientation of the endoscopic capsule as it moves throughout the GI tract [3]. Information regarding the localization and motion of the capsule is particularly valuable when an abnormality is detected in the tissue. The development of automatic methods to overcome this limitation is therefore essential. These methods can be based on a variety of principles and techniques. Several computer vision techniques are based on image analysis alone, using the images acquired by the capsules to estimate motion and displacement throughout time. Recently, artificial intelligence has produced notable progress in this field [1].

A. PAPER ORGANIZATION

In this paper, after discussing the related work, a novel experimental setup is presented in the data acquisition section. The approach chosen for the identification and localization of the capsule that is presented in the methodology section is based on image analysis alone. The image registration step of the localization process is addressed within the framework of WCE video frames by exploring different image registration approaches. The main goal is to estimate robust correspondences between overlapping regions of closely-spaced frames from the acquired data. Lastly, for a quantitative evaluation of the results, the computation of a residual error using the fundamental matrix and the computation of the corresponding rotation and translation errors using the essential matrix is presented. The results obtained are then presented and discussed.

II. RELATED WORK

WCE localization systems are broadly classified into three types according to the sensing method: magnetic-field-strength methods, electromagnetic wave and field-based methods, and image-based methods [4], [5], [8], [9]. Magnetic localization techniques can be implemented using an internal permanent magnet in the capsule and a sensing module outside the capsule. An alternative to this approach is to use a magnet outside and a sensing module inside the capsule [15]. Other methods combine magnetic localization with magnetic actuation. The PillCam capsule from Medtronic, for example, can perform localization using a set of 8 receivers located on the patient's abdomen. The intensity of the Radio Frequency (RF) signals is used for the estimation of the location of the capsule, an approach that does not require any additional equipment [16]. RF-based approaches can use various principles for capsule localization: radio frequency identification (RFID); time of arrival (TOA); direction of arrival (DOA); time difference of arrival (TDOA); angle of arrival (AOA); and received signal strength indicator (RSSI) [4], [6], [7], [17]. There are also approaches combining RF localization and computer vision to determine 3D motion [18] and capsule orientation [19] that are complementary to magnetic techniques.

A. COMPUTER VISION METHODOLOGIES

Since endoscopy capsules are equipped with cameras and light-emitting diodes, other approaches are based on image analysis and computer vision [20]. Computer vision methodologies measure the displacement of the capsule inside the GI tract as the rigid motion of the capsule [21] by retrieving, for example, visual features or image intensities changes between video frames [22]. In the first part of this process, after the extraction of points of interest, features or visual cues, the image registration between video frames is performed. Next, the 3D rigid motion of the camera capsule between frames is estimated to allow for odometry estimation and capsule localization estimation. This computation is obtained relative to the capsule itself given internal landmarks and taking into account the luminal geometry [21], [23].

The estimation of the 3D motion of the capsule relies on the computation of the Essential matrix [24]. Given two images A and B acquired by a calibrated camera, with I_A and I_B representing the homogeneous coordinates of the pixels of the images, the following relationship applies:

$$I_A^T E\, I_B = 0 \tag{1}$$

where E, the Essential matrix, is a $3 * 3$ matrix of rank 2. Matrix E can be expressed as a function of the product of a 3D rotation matrix R and of a skew-symmetric matrix T made up with the elements of the translation vector $\vec{t} = (t_x, t_y.t_z)$

$$E = R\,T \tag{2}$$

where R and $\vec{t}$ describe the rotation and translation between the two camera positions of A and B. Given the matrix E, the 3D rotation R and translation $\vec{t}$ can be computed up to a scale factor.

Most of the localization methods mentioned require an external module to the capsule, which complicates the process. For this reason, computer vision approaches that only require the information provided by the recorded frames are quite promising. Still, the main difficulty imposed by these techniques relies on the search for sufficient and robust corresponding points, regions or features between frames in order to accurately compute the essential matrix.

B. IMAGE REGISTRATION

The estimation of matrix E requires that corresponding geometrical entities such as points, lines or regions are determined. Therefore the estimation of the 3D capsule displacement requires image registration. The registration process is an alignment problem, and it can be viewed as a spatial transformation of matching points between two sets of

data [25]. The registration process involves recovering the spatial transformation T that maps I_B to I_A:

$$T : I_B \rightarrow I_A \Leftrightarrow T(I_B) = I_A \tag{3}$$

As a result of Equation (1), the mapping depends on the depth of the 3D points. For capsule endoscopy, image registration depends both on the image changes due to capsule motion and to intestine motion. The intrinsic camera parameters and the distortion coefficients associated with the capsule's camera are extracted prior to the registration step by appropriate calibration in order to remove distortion in the endoscopy frames [25].

Image Registration methods can be grouped into direct (or pixel-based) strategies or feature-based strategies [11]–[13].

1) DIRECT (PIXEL-BASED) VS FEATURE-BASED METHODS

Strategies that determine a proper motion model to define the alignment between a pair of images, compute its parameters and shift or warp the images relative to each other and explore how much the pixels agree are called direct or pixel-based methods [13]. An error metric is chosen for the comparison and a search technique is also devised. The easiest technique is to do a full search and try all possible alignments, which can be computationally exhaustive. Alternatively, approaches that resort to Fourier transforms and hierarchical coarse-to-fine approaches based on image pyramids can be used to speed up the computation [14]. Some other approaches are based on the Taylor series expansion of the image function to get sub-pixel precision in the alignment [13].

The other main registration strategy opposed to the direct method is the feature-extraction method. In this technique, the algorithms first extract distinguishing features from both images, match the individual features and then determine a global correspondence in order to compute a robust geometric transformation between them [12].

Initially, in older feature-based methods, when the images were poorly textured the features ended up being unevenly distributed and the algorithms were not able to provide accurate matches for pairs that should have been aligned [12]. Additionally, in some of these feature-based methods, the matching relied solely on the cross-correlation between regions comprising the features which failed to produce a good alignment when the images were rotated. Contrarily, direct methods use all available information because of the contribution of every pixel. These methods also have a limited range of convergence. To overcome this challenge, coarse-to-fine techniques are generally used but the addition of more levels into the pyramid often ends up blurring important image details. Recent feature-based methods operate in scale-space and use orientation invariant descriptors to match images that differ in scale and orientation. These descriptors are designed for repeatability and the extracted features end up being well distributed which produces enough correspondences [12].

2) ARTIFICIAL INTELLIGENCE STRATEGIES

Artificial Intelligence (AI) methodologies devised for image registration are considered feature-based methods given the search technique for the correspondences. In machine learning strategies, image features are first extracted by the user and then an artificial neural network system is used in order to predict and/or classify the new data [1]. This learning process can be performed in a supervised or unsupervised manner, depending on whether or not ground truth information is available. Deep learning refers to a class of artificial neural network systems with several layers that have the advantage of automatically extracting features. In the medical image analysis field, the most used deep neural network structure is the convolutional neural network (CNN) [1].

III. DATA ACQUISITION

The data used for this assessment is the Mirocam dataset obtained with our novel experimental setup [10]. Since the colon is not *in-vivo* and the capsule cannot be moved through peristaltic movements, it was crucial to develop a different approach for the movement of the capsule along the *ex-vivo* porcine colon. The colon was cut longitudinally and fixed into a foam with a previously excavated path. This novel experimental setup allows for the recording of video frames along the entire length of a fixed *ex-vivo* porcine colon by a camera capsule that is moved by a robotic manipulator while it stores 3D motion information at each instant [10]. This setup provides valuable ground truth information regarding the sequence of camera poses at any given instant and consequently the camera's trajectory.

Hence, as presented in Figure 1, the experimental setup includes an ex-vivo porcine colon attached to a scaffold, a camera capsule, a capsule holder, a robotic arm, a data belt, a receiver and a receiver cradle. The capsule used was the MC1000 Mirocam Capsule from IntroMedic, which has a static frame rate of 3 FPS. The camera is attached to the gripper of a robotic arm with a two-piece capsule holder and moved through a preprogrammed path along a harvested *ex-vivo* porcine colon previously sutured into an excavated foam scaffold. In a normal exam, the patient wears the belt around the waist and the signal is transmitted from the capsule to the belt through the skin. In this case, since there is no patient, the signal cannot be transmitted through the skin so it reaches the belt with double ended alligator clamps.

Throughout the experiment, the robotic arm recorded the orientation and position of the gripper that holds the camera, along the predefined trajectory at regular time intervals, an information that is inaccessible in WCE exams. Hand-eye calibration was previously performed to estimate the rigid transformation between the gripper and the capsule camera.

A pinhole camera model with radial distortion was considered for the calibration of the capsule camera. The calibration parameters are presented in Table 1. Given that the frame rate of the capsule is fixed, it is possible to compare the orientation and location registered by the robot at each time

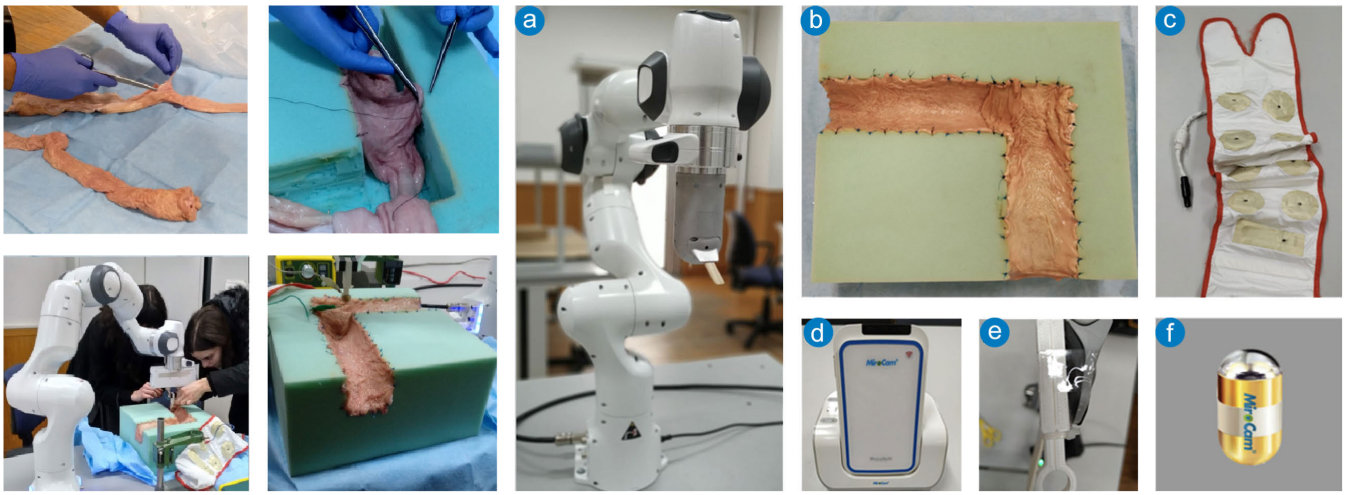

FIGURE 1. The experimental setup consisted of: (a) a robotic arm for the displacement of the capsule along the trajectory; (b) a pre-harvested and cleaned *ex-vivo* porcine colon sutured into an excavated foam; (c) a data belt for the signal transmission; (d) a receiver in a receiver cradle (that connects to the computer); (e) a special production two-piece capsule holder to fixate the capsule to the robotic arm; and (f) a MC1000 Mirocam Capsule from IntroMedic.

TABLE 1. Mirocam MC1000 capsule calibration parameters (in pixels): Principal point (c_x, c_y), Focal length (f_x, f_y), Skew (s) and radial distortion coefficients (k_1, k_2).

Principal Point		Focal Length		Skew	Radial Distortion	
c_x	c_y	f_x	f_y	s	k_1	k_2
178.28	186.47	149.29	148.29	0	-0.2441	0.0726

interval with the rotation and translation computed from the data extracted from the image sequence with each registration method, to identify the most suitable methodology for trajectory reconstruction of WCE frames.

IV. METHODOLOGY

An overview of the methodology used for this assessment is presented in Figure 4. Both direct (or pixel-based) and feature-based methods were explored. Some of these registration approaches are agnostic since they do not use a parametric model while others do. For the Direct (or pixel-based) method, a hybrid multi-scale elastic model with an affine pre-registration (MEIR/MPIR) developed especially for WCE video frames [32] was explored. The feature-based methods explored were chosen according to the results of the comparative assessment in [26] performed especially with images from wireless capsule endoscopy. These methods are Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Maximally Stable Extremal Regions (MSER) and Local Intensity Order Pattern (LIOP). Lastly, a commonly used deep-learning method for optical flow computation (PWC-Net) was also explored as a feature-based method for image registration.

In order to evaluate and compare the explored image registration methods, two approaches were used. One is based on the computation of the Fundamental matrix and the estimation of the distances between the corresponding features and the epipolar lines. The other approach is based on the estimation of the 3D motion that the endoscopic capsule undergoes. For that purpose the Essential matrix is computed and the trajectory data acquired by the robot manipulator responsible for the movement of capsule is used as ground-truth.

A. HYBRID MULTI-SCALE ELASTIC MODEL WITH AN AFFINE PRE-REGISTRATION (MEIR/MPIR)

This registration procedure is formulated as a minimization problem requiring a multiple scale description of the input frames that aim to reduce or eliminate possible local minima and to expedite the convergence of the method [32]. Since in a normal WCE exam, the capsule is driven by peristalsis, the model assumes that the overall movement is a combination of the rigid movement of the capsule itself and the non-rigid deformation of the small intestine, which is an elastic and deformable organ.

This method relies on the grey-scale version of the WCE video frames and defines the relationship between a pair of images (I_R, I_T), where I_R is the reference kept unchanged, I_T is the template and $x = (x_1, x_2)$ is an arbitrary pixel in the domain Ω. The aim is to find the geometric transformation ϕ, that minimizes the distance $D = D(I_R, I_T(\phi))$, defined in (4), involving the space of square integrable function $L^2(\Omega)$, between the transformed template image $I_T(\phi)$ and the reference image I_R [32].

The normalized dissimilarity measure (NDM) between images R and $T(\phi)$ is defined in Equation (5).

$$D = \frac{1}{2}\|T(\phi) - R\|^2_{L^2(\Omega)} = \frac{1}{2}\int_\Omega (T(\phi(x)) - R(x))^2 dx, \tag{4}$$

$$NDM := \frac{\|T(\phi) - R\|_{L^2(\Omega)}}{\|R\|_{L^2(\Omega)}}. \tag{5}$$

The multi-scale approach refers to a multi-scale representation of the data, reference R and template T images. R_{θ_i} and T_{θ_i} represent the interpolated reference and template images respectively, obtained with spline interpolation, for a pre-defined increasing sequence of scales, denoted by θ_i, with $i = 0, 1, \ldots, n$. At a coarse scale, only the most noticeable features in both images are preserved, while small details become more visible at finer scales.

Then, the multi-scale image registration (MEIR) defined in [32] consists of an affine pre-registration at the initial and coarse scale θ_0, defined by

$$\min_{\phi} \frac{1}{2} \|R_{\theta_0} - T_{\theta_0}(\phi)\|^2_{L^2(\Omega)}, \tag{6}$$

followed by a sequence of elastic image registration steps, at subsequent and increasingly finer scales θ_i, for $i = 1, \ldots, n$. To speed up the total optimization process and prevent possible local minima, the solution at scale θ_{i-1} is used as the starting point for the elastic registration at the finer scale θ_i and the unknown transformation ϕ is split into an identity part $I_{3\times3}$ and a deformation part u as shown in (7).

$$\min_{u} \frac{1}{2} \|R_{\theta_i} - T_{\theta_i}(I_{3\times3}d - u)\|^2_{L^2(\Omega)} + \alpha S(u) \tag{7}$$

$$S(u) := \int_{\Omega} \left(\frac{\lambda + \mu}{2} ||div\ u||^2 + \frac{1}{2} \sum_{i=1}^{2} ||\nabla u_i||^2\right) dx. \tag{8}$$

$S(u)$ is formulated as a function of the Lamé constants, λ and μ that characterize the elastic properties of the tissue. The regularization parameter α balances the impact of the similarity in the final cost function and the elastic regularization term $S(u)$, defined in (8), enables the optimization problem to be well-posed and restricts the solution to a linear elastic transformation.

A multi-scale affine image registration approach (MPIR) is also defined and can be thought of as a particular case of the MEIR approach with $\alpha = 0$ [32].

B. SCALE-INVARIANT FEATURE TRANSFORM (SIFT)

SIFT algorithm, described in Figure 3.1), implements a cascade filtering procedure for the identification of stable points in the scale space [27]. Each keypoint descriptor is extracted from a set of reference frames, stored in a database, compared with a new input frame and the points that minimize the euclidean distance between features vectors are then selected. The final subset of correspondences is assigned based on position, scale, and orientation [33]. Keypoint descriptors are created from local geometric deformations represented by blurred difference of Gaussians (DoG) image gradients in various orientation planes at multiple scales by determining both the magnitude and the orientation of the gradient around each position. Although SIFT is quite slow and it is not as effective for low powered devices [36], its features are partially invariant to illumination and distortion, are resistant to image noise and remain invariant to scaling, rotation and translation. [27], [28].

C. SPEEDED UP ROBUST FEATURES (SURF)

SURF algorithm, described in Figure 3.2), presents a fast point-extraction and description scheme that is proven to produce high robustness with changeable lighting conditions [29]. Although SURF is not very stable to rotation [36] and provides fewer key-points than SIFT, it is faster and more robust against different image transformations [29], [30].

D. MAXIMALLY STABLE EXTREMAL REGIONS (MSER)

Image sections that remain nearly unchanged along as extensive range of thresholds are designated Maximally Stable Extremal Regions (MSER)s. The MSER algorithm, described in Figure 3.3), achieves correspondences between frames from different viewpoints based on the extremal regions achieved with a local binarization technique by using predefined threshold values. These features are popular for fast blob detection and its description is rotation-invariant given that the information exploited is local [34]. MSER also has a limited performance on blurred and/or textured images, since blur can distort the shapes of the extracted MSERs [35].

E. LOCAL INTENSITY ORDER PATTERN (LIOP)

The Local Intensity Order Pattern (LIOP) algorithm, described in Figure 3.4), uses intensity order instead of raw intensities and exploits the fact that the relative order of pixel intensity is unchanged with monotonic variations. The feature descriptors described so far are sufficiently robust to multiple lighting and distortion variations but fail to produce the best results in a few particular cases with more complex lighting changes such as specular reflections and exposure time variations, which are very common in endoscopic datasets [31].

F. DEEP-LEARNING FOR OPTICAL FLOW USING PYRAMID, WARPING, AND COST VOLUME (PWC-NET)

PWC-Net is a compact CNN model for optical flow estimation designed according to pyramidal processing, warping and cost volume. The combination of deep learning and domain knowledge reduces model size and improves performance [37].

Firstly, since raw images are prone to variations in lighting conditions, this method uses learnable feature pyramids. In this architecture, a pyramid of feature representations is constructed given two images I_1 and I_2 with an L number of layers. The bottom level corresponds to the features of the input images and the upper lth level to the lth downsampling representation of the features. Secondly, the warping operation from traditional approaches is incorporated as a layer in the network to estimate large motion. Then, there is another layer to construct the cost volume, which is a more appropriate representation for optical flow estimation. These two layers for warping and cost volume have no learnable parameters in order to reduce model size. The optical flow estimator is a multi-layer CNN given the cost volume, the features from the first image and the upsampled optical flow as an input as shown in Figure 2. The output is the

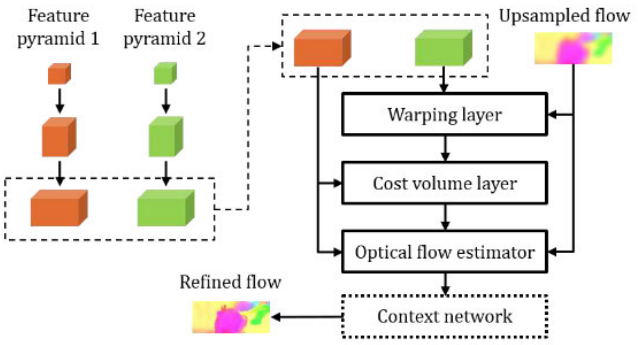

FIGURE 2. Using upsampled flow, the PWC-Net warps the features of the second image and then computes and processes the corresponding cost volume, ultimately to obtain a refined flow estimate. The arrows are indicative of the direction of the estimation of the optical flow. Pyramids are built in the opposite direction of flow estimation. (Adapted from [37]).

optical flow estimation in the lth level [37]. Finally, according to the post processing, the PWC-Net resorts to a context network to exploit the contextual information for optical flow refinement.

Other network architectures have also been designed using principles both from the stereo and optical flow information. These use image pyramids or three-level feature pyramids, while PWC-Net learns deeper feature pyramids to achieve better performance. Other architectures also warp the input images instead of the features, which hinders the information propagation. Thus, the PWC-Net is able to construct a multi-resolution cost volume and uses a low search range to reduce the computation [38].

Using the TensorFlow-based implementation tutorial of PWC-Net [37], available at [51] and selecting a pre-trained model (pwcnet-lg-6-2-multisteps-chairsthingsmix), an optical flow estimation was conducted for all pairs of sequential frames. The optical flow angle and magnitude values obtained with PWC-Net are stored in the RGB images presented in Figure 5.

G. ESTIMATION OF POINT MATCHES

For the MEIR and MPIR methods, with each pair of frames k and k + 1, given a subset of P_k points and the parameters computed (scale (s), rotation angle (θ), translation components (t_x and t_x)), were used to determine the rotation matrix R(θ), the translation vector $T(t_x, t_y)$ and the respective P'_{k+1} position of the transformed points P_k in frame k + 1.

$$P'_{k+1} = sR(\theta)P_k + T(t_x, t_y) \tag{9}$$

Regarding the PWC-Net, for each pair of consecutive frames, with the subset of points P_k from the initial frame, given the each angle and magnitude obtained for each pixel from the optical flow estimation, the corresponding P_{k+1} points coordinates in the consecutive frame were computed in order to be used as matching points.

For the feature-based registration methods described above (SIFT, SURF, MSER and LIOP algorithms), the set of extracted features from each pair of sequential frames k and k + 1 were used to compute the P_k and P_{k+1} point matches.

In addition to the correspondences determined with the registration methods, a preliminary manual annotation was also used in a subset of video frames to provide an additional reference benchmark to be used in the computation of the distances to the epipolar lines. This manual annotation is also useful to visually compare the quality of the matching points obtained from each approach. The 15 pairs of consecutive WCE images with the highest number of matches with non-zero displacement were chosen and manually annotated. All sets of point matches P_k and P_{k+1} from each registration procedure were corrected for lens distortion with the calibrated camera parameters.

All registration methods explored for the search of robust correspondences were implemented in Matlab 2019a, except for the pre-trained PWC-Net that was explored with Python 3.6, with a TensorFlow implementation. All registration results obtained were then compared with the computation of the fundamental and essential matrices also in Matlab 2019a. All tests were performed in a computer with a 3.4 GHz Intel Core i7 processor and 16 GB of RAM.

H. FUNDAMENTAL MATRIX ESTIMATION

The set of initial points P_k from frame k and points P_{k+1} from frame k + 1, obtained with each image registration method, are used to estimate the fundamental matrix F. For that purpose the normalized eight-point algorithm [24] was used. Given the epipolar lines in both frames, computed using Equation (10) and its dual (for the backward correspondence), the distances between the matched points and corresponding epipolar lines can be computed using Equations 11 and 12.

$$l_{k+1} = FP_k \tag{10}$$

With $l_{k+1} = [a_{k+1}b_{k+1}c_{k+1}]$ defining an epipolar line. If $P_{k+1} = [x_{k+1}y_{k+1}1]$, then:

$$a_{k+1}x_{k+1} + b_{k+1}y_{k+1} + c_{k+1} = 0 \tag{11}$$

which means that each point should belong to its corresponding epipolar line.

Ideally, for all i = 1, . . . ,n absolute epipolar distances d_i^F between each point P_{k+1} and each epipolar line l_{k+1} obtained with F, given by equation 12, should be equal to zero.

$$d_i^F = \frac{|a_{k+1}x_{k+1} + b_{k+1}y_{k+1} + c_{k+1}|}{\sqrt{a_{k+1}^2 + b_{k+1}^2}} \tag{12}$$

The distances $d_i^{F^T}$ in the opposite direction, from frame k + 1 to frame k, can be obtained with the same procedure but using the transpose of the fundamental matrix F^T instead of F and the corresponding epipolar lines $l_k = [a_k b_k c_k]$.

Consequently, the root mean squared distances associated with each pair of frames k and k + 1, would also be equal to zero. In order to determine the registration error between

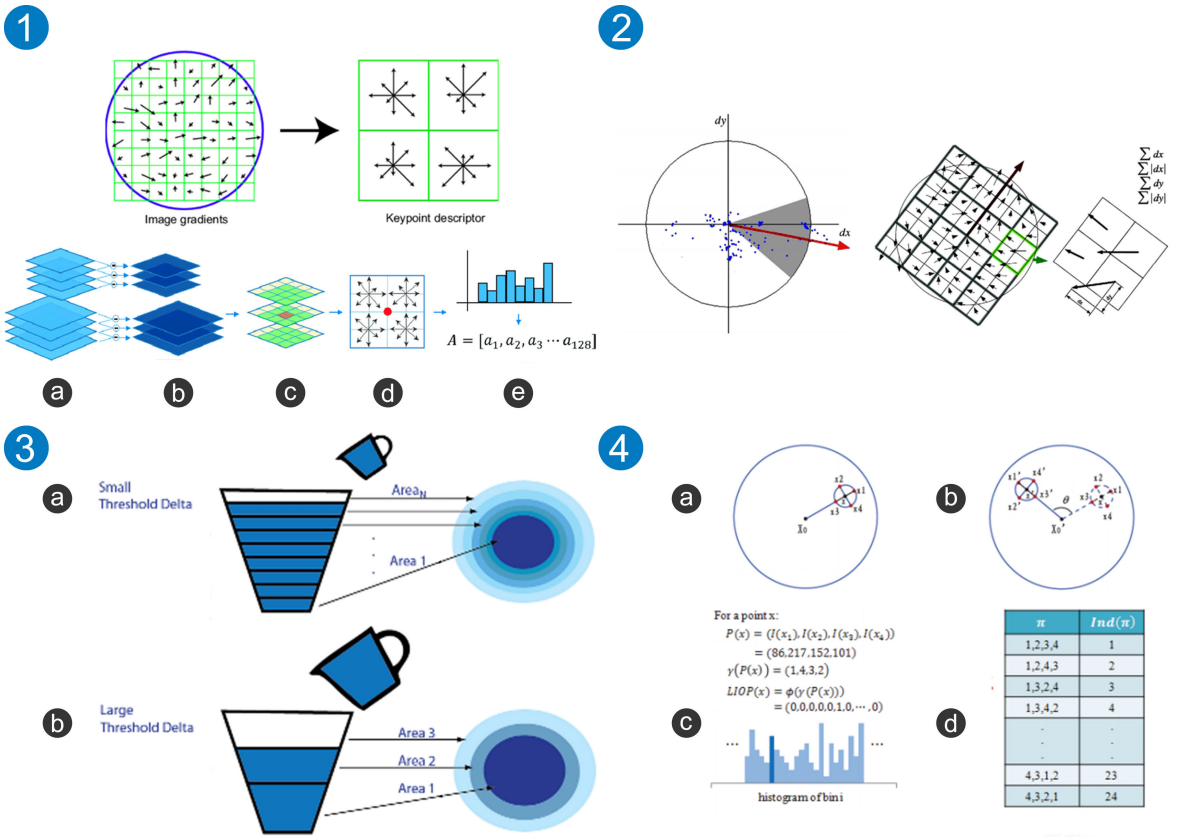

FIGURE 3. Compilation diagram of all feature-based methods explored in the comparative assessment. 1) SIFT: Scale-Invariant feature transform. On top, a key-point descriptor with arrow lengths representing the gradient magnitude sum within the region. (Adapted from [27]); In the bottom, the SIFT algorithm steps: a) Scale space generation; b) DOG image generation; c) Detection of local maximum and minimum; d) Gradient calculation; e) Histogram calculation and generation of dimensional vectors. (Adapted from [28]) 2) SURF: Speeded up robust features. Achieves the distribution intensity around each point of interest with the sum of the Haar wavelet responses around a circular neighbourhood. The responses are weighted by a Gaussian function and plotted in a two-dimensional space. The matches are selected when the maximum of the determinant of the Hessian matrix that characterizes the local changes is found. (Adapted from [29]) 3) MSER: Maximally stable extremal regions are produced when the local minima of the rate of change is identified with a threshold sequence. a) Small threshold delta; b) Large threshold data (Adapted from the mathworks documentation). 4) LIOP: Local intensity order pattern. a) Original patch construction; b) Patch rotation and noise removal; c) LIOP computation; d) Descriptor Construction by indexing. (Adapted from [31]).

frame k and k + 1 in both directions, the residual error is computed as suggested in [24] and presented in Equation (13).

$$e_{res} = \frac{1}{\sqrt{4n}}\left(\sum_{i=1}^{n}(d_i^F)^2 + \sum_{i=1}^{n}(d_i^{F^T})^2\right)^{\frac{1}{2}} \tag{13}$$

I. ESSENTIAL MATRIX ESTIMATION

The estimation of the essential matrices E allows for the recovery of the 3D rotation matrix and the translation vector T (up to a scale factor), throughout the trajectory [24]. The matrices were estimated using the M-estimator sample consensus (MSAC) algorithm [50] with bundle adjustment, for the set of point matches P_k and P_{k+1}, from all 15 pairs of frames, obtained with the image registration methods described (and the calibrated camera parameters from Table 1). The orientation and location of the calibrated camera relative to its previous pose were also obtained. For each pair of registered frames k and k + 1, the relative rotation matrices $R_{k,k+1}$ and the relative translation vectors $T_{k,k+1}$ were obtained. These matrices were compared against the relative rotation matrices $R^{robot}_{k,k+1}$ and the relative translation vectors $T^{robot}_{k,k+1}$ obtained from the robot data (ground-truth).

1) ROTATION MATRIX ERROR

Firstly, for the evaluation of the estimated rotation matrices, a rotation error matrix $R^{err}_{k,k+1}$ for each pair of k, k + 1 frames was computed as shown in Equation (14).

$$R^{err}_{k,k+1} = R^{T}_{k,k+1} * R^{robot}_{k,k+1} \tag{14}$$

This matrix is still a rotation matrix and can be represented using the axis-angle representation which parameterizes the rotation in a 3D Euclidean space by a vector corresponding to the axis of rotation and an angle of rotation [24]. For all pairs of consecutive frames, the rotation error matrices R_{err} obtained for each registration method (MEIR, MPIR, SIFT, SURF, MSER, LIOP and PWC-Net) were converted into the vector of the axis of rotation $v^{R}_{k,k+1}$ and the angle of

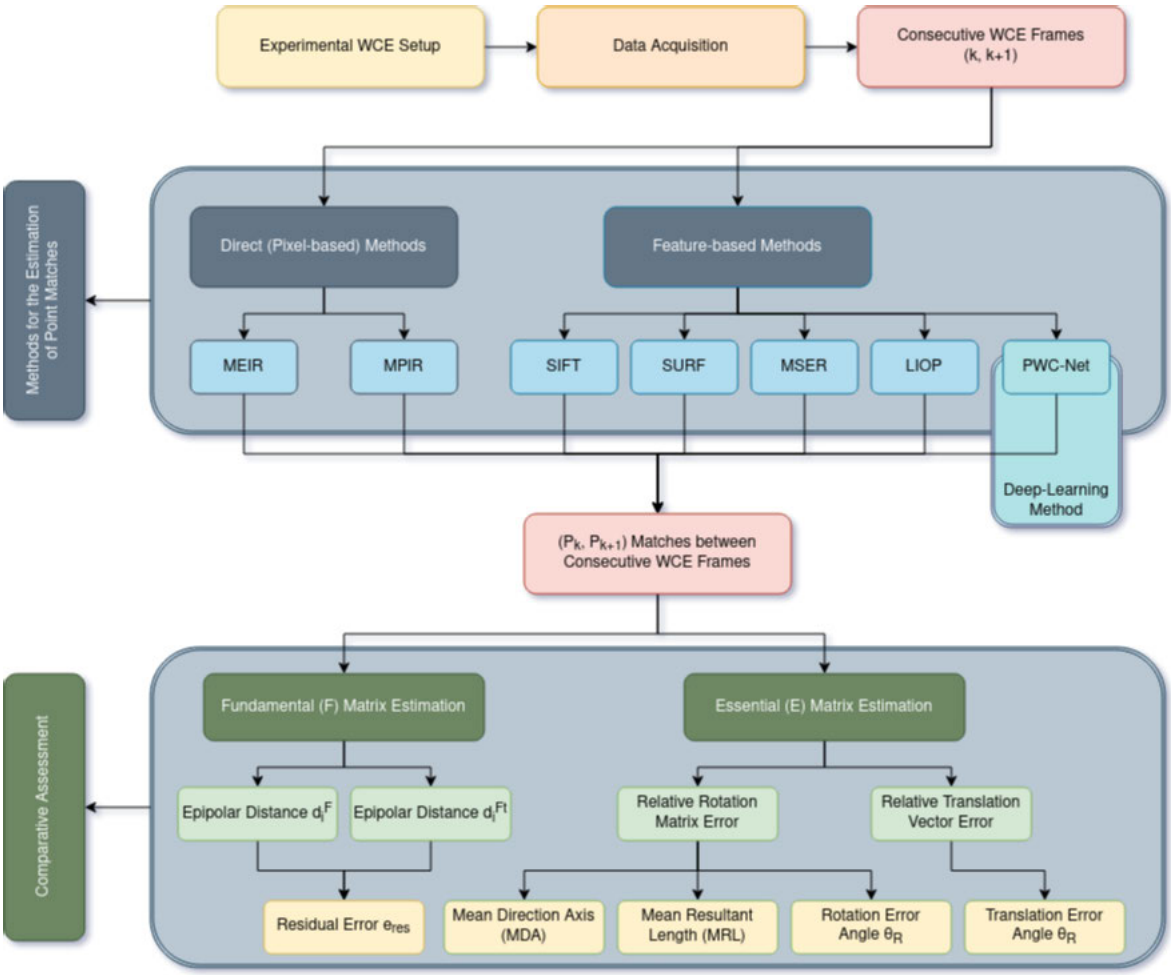

FIGURE 4. Diagram illustrating an overview of the methodology adopted in this study.

rotation $\theta^R_{k,k+1}$. The weighted average of each $v^R_{k,k+1}$ given all pairs of consecutive frames was estimated for each registration method and expressed in polar coordinates as described in [40] and represented in Equation (15), where $\overline{x_0}$ is a unit vector and $\overline{R} \geq 0$, so that $\overline{R} = ||\overline{x}||$ and $\overline{x_0} = ||\overline{x}||^{-1}\overline{x}$.

$$\overline{x} = \overline{R}\overline{x_0} \tag{15}$$

The vector $\overline{x_0}$ is called the Mean Direction Axis (MDA) and $\overline{R}$ is called the Mean Resultant Length (MRL) [40]. The MDA $(\overline{e}_x, \overline{e}_y, \overline{e}_z)$ for each registration method was estimated along with the standard mean deviation $(\sigma_{\overline{e}_x}, \sigma_{\overline{e}_y}, \sigma_{\overline{e}_z})$ for all its components. The mean error angle $\overline{\theta}_R$ was also obtained by computing the average of each $\theta^R_{k,k+1}$ obtained with each pair of consecutive frames along with the corresponding standard mean deviation $\sigma_{\overline{\theta}_R}$.

2) TRANSLATION VECTOR ERROR

For the comparison of the translation vectors, the cosine of the angle between the estimated translation vector $T_{k,k+1}$ and the translation vector registered by the robotic arm $T^{robot}_{k,k+1}$ was estimated for each pair of k, k + 1 frames as shown in Equation (16).

$$cos(\theta^T_{k,k+1}) = \frac{T_{k,k+1} \cdot T^{robot}_{k,k+1}}{||T_{k,k+1}||||T^{robot}_{k,k+1}||} \tag{16}$$

Ideally, the values for the cosine should be equal to 1 and the corresponding angle should be equal to zero. The translation angles $\theta^T_{k,k+1}$ between the two translation vectors were extracted from the cosine values. Finally, the weighted average of each $\theta^T_{k,k+1}$ along with the corresponding standard mean deviation $\sigma_{\overline{\theta}_R}$ were computed.

V. RESULTS

In Figure 5, a sample of the images from the experimental dataset is shown regarding the machine-learning PWC-Net strategy for the registration step. The input is made up of a subset of pairs of consecutive frames and the output are the components of the optical flow vectors, stored in the RGB channels. A few selected point matches obtained with each registration method (MEIR, MPIR, SIFT, SURF, MSER, and LIOP) in the first pair of consecutive frames from the experimental dataset, is compiled in Figure 6 along with

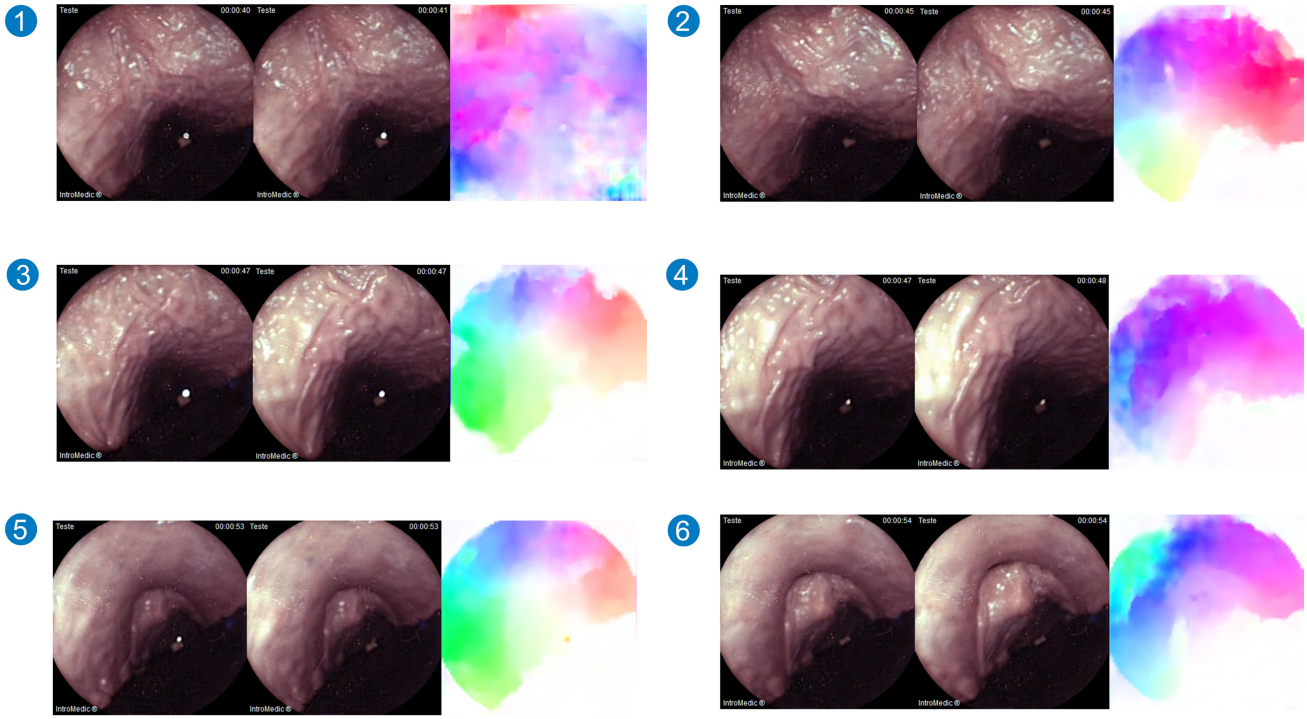

FIGURE 5. Sample images regarding the machine-learning strategy for the registration. Frame pairs 1) 190-191; 2) 214-215; 3) 222-223; 4) 225-226; 5) 251-252; 6) 255-256 from the experimental dataset. The network input is composed of a set of pairs of consecutive frames (left), and the output is the optical flow estimation achieved with the PWC-Net and stored in the RGB channels (right).

TABLE 2. Mean direction axis (MDA) of the rotation error matrix, Mean resultant length (MRL) of the rotation error matrix, Mean rotation angle ($\bar{\theta}_R$), Mean translation angle ($\bar{\theta}_T$) and elapsed time (ET) in seconds for each registration approach (MEIR, MPIR, SIFT, SURF, MSER, LIOP, PWC-Net)s. The sigmas are the corresponding the standard mean deviations.

	MDA ($\bar{e}_x$,$\bar{e}_y$,$\bar{e}_z$)	($\sigma_{\bar{e}_x}$, $\sigma_{\bar{e}_y}$, $\sigma_{\bar{e}_z}$)	MRL	σ_{MRL}	$\bar{\theta}_R$	$\sigma_{\bar{\theta}_R}$	$\bar{\theta}_T$	$\sigma_{\bar{\theta}_T}$	ET
MEIR	(0.0092, 0.0116, -0.0554)	(0.0479, 0.0472, 1.3272)	0.0472	1.3272	0.9574	0.7967	1.8700	0.6043	395.0586
MPIR	(0.0057, 0.0380, -0.0528)	(0.0766, 0.0954, 1.3246)	0.0954	1.3246	0.9587	0.7969	1.7248	0.6168	111.4123
SIFT	(0.0069, 0.0240, 0.0262)	(0.0784, 0.0708, 0.0564)	0.0708	0.0564	0.0485	0.0399	1.3882	0.7592	4.5755
SURF	(-0.1954, -0.0551, 0.1514)	(0.4748, 0.4183, 0.3772)	0.4183	0.3772	0.4709	0.9030	1.6468	0.6748	4.4651
MSER	(0.0017, 0.0110, -0.0007)	(0.0290, 0.0239, 0.0327)	0.0239	0.0327	0.8518	1.4213	1.5727	0.7482	1.1409
LIOP	(-0.0542, 0.0280, 0.0463)	(0.3353, 0.3320, 0.1881)	0.3320	0.1881	0.3474	0.7680	1.5031	0.6113	4.5529
PWC-Net	(0.0090, 0.0003, 0.0096)	(0.0306, 0.0394, 0.0283)	0.0394	0.0283	0.0268	0.0162	1.0481	0.6768	81.2709

the manually annotated matches for visual comparison. The residual errors obtained with the use of the fundamental matrix for each registration method for all 15 pairs of consecutive frames are shown in Figure 7. The mean axis of rotation, the mean resultant length and the mean angle of the rotation error matrix are presented in Table 2 along with the mean angle between translation vectors and the corresponding standard deviations for each registration procedure.

VI. DISCUSSION

The main contributions of this paper are the development of the experimental setup (that was assembled) and the approach for the estimation of the capsule's relative motion given only the common information provided by sequential images. This experimental setup is different because in an environment where the camera moves with the peristaltic movements of the GI tract it is not possible to obtain a ground truth regarding the position of the camera along the path. Since our goal is to determine a methodology that efficiently reconstructs the trajectory of the capsule at each instant of the route, it was necessary to develop a setup where this ground truth was accessible. In this case, the camera is moved by a robotic arm that can register the position of the camera at all times. The objective is to extract the rotation and translation of the camera between each pair of frames using only the information from the images.

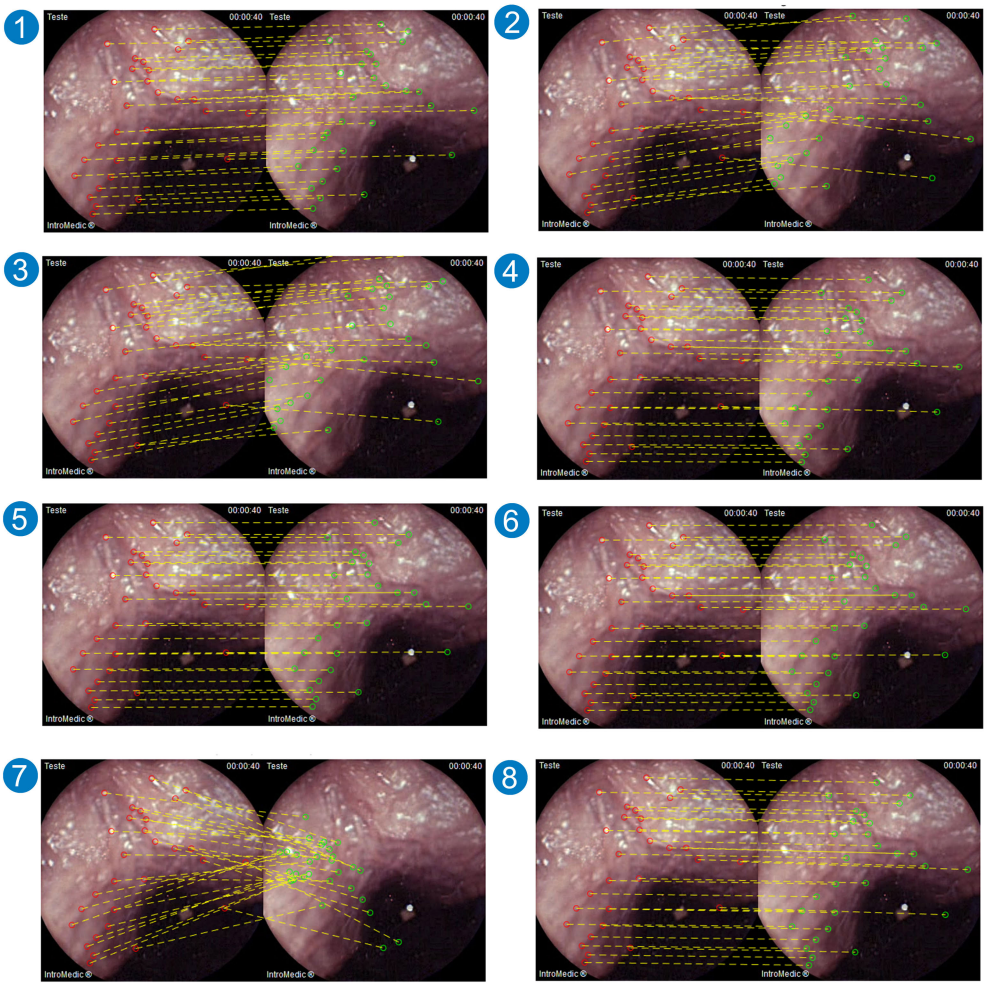

FIGURE 6. Sample consecutive frames from the experimental dataset and the corresponding matched points (P_k and P_{k+1}) obtained with each image registration approach. 1) Manual annotation (MA); Hybrid-based Strategy: 2) MEIR; 3) MPIR; Feature Extraction Strategy: 4) SIFT; 5) SURF; 6) MSER; 7) LIOP; and Machine-learning Strategy: 8) PWC-Net.

The matches obtained with LIOP, shown in Figure 6.7) are affected by significant errors. The residual errors obtained are high and inconsistent and the number of matches between frames is quite low, which is why in some image pairs it was not even possible to compute the essential matrix for the estimation of the rotation error and translation error. The results obtained with MEIR and MPIR are visually more plausible than those obtained with SIFT, SURF, and MSER. In the sample example shown in Figure 6.4);5);6) it can be seen that these methods do not produce suitable matches. Additionally, SURF does not produce enough matches to compute the fundamental matrix and consequently the residual error, in all pairs of frames, is significant. The MDA, MRL and $\overline{\theta_R}$ values for SURF, along with LIOP, are also high. Most likely the results obtained with these methods correspond to non-moving features. This can be concluded by comparing the point correspondences obtained with these methods and the manually annotated points in Figure 6. On the other hand, the fundamental matrices estimated using MEIR and MPIR were calculated with a small sample of matches, unlike SIFT, SURF and MSER which yielded random and more extensive point matches. For the purpose of image comparison, only the points with manually annotated matches were displayed in Figure 6. It is possible to assume that the residual errors, MDA, MRL and error angles would decrease if the estimates of the fundamental and essential matrices had been obtained with a larger set of matches. It could be expected that the method MEIR, that models deformation/elasticity, would yield better results with the porcine colon images, but that is not the case. MEIR and MPIR perform similarly with this dataset. Our experimental setup is different from real case WCE videos since, in our case, the colon is fixed and the capsule does not move as a result of the peristalsis of the small intestine. The capsule's movement is guided

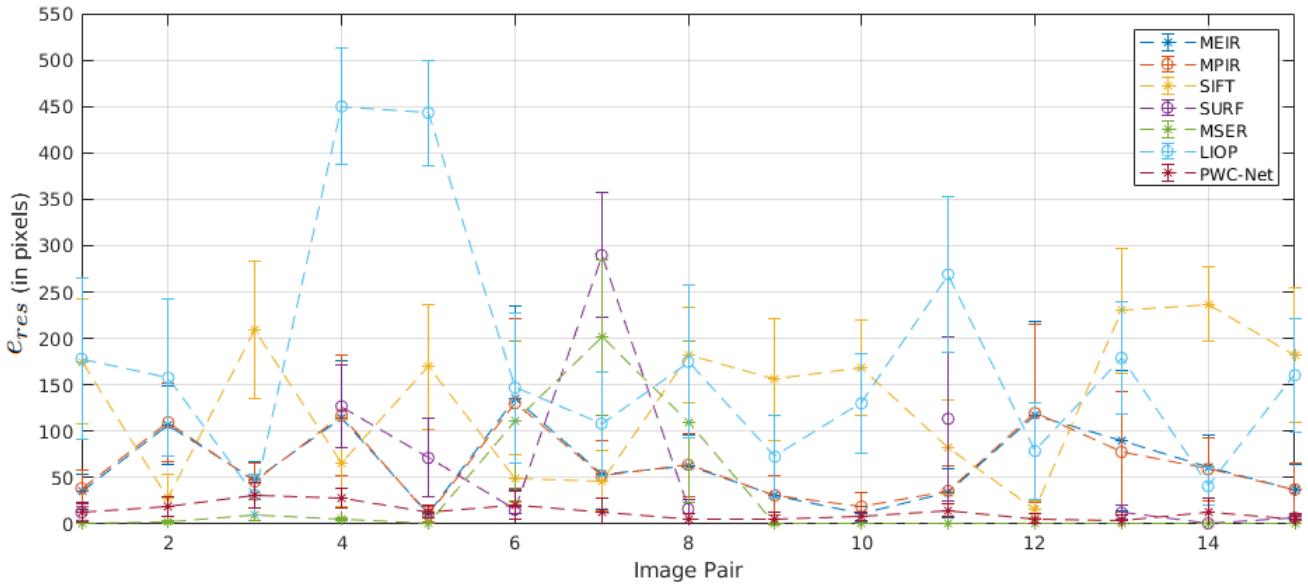

FIGURE 7. Residual error for each registration method (MEIR, MPIR, SIFT, SURF, MSER, LIOP, and PWC-Net) for all pairs of consecutive frames.

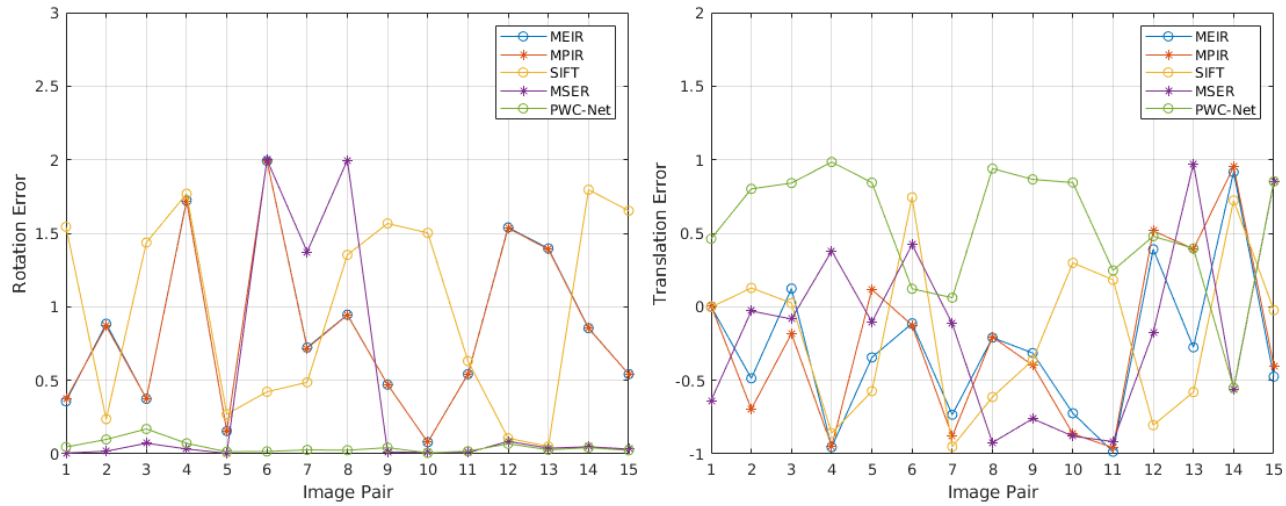

FIGURE 8. Rotation and translation error angle for each registration method (MEIR, MPIR, SIFT, MSER and PWC-Net) for all 15 pairs of consecutive frames.

by a robotic arm, so it is plausible that the method that accounts for elasticity does not yield better results than the one that disregards the existence of elasticity. MEIR and MPIR methods perform relatively well, considering the visual comparison of the matches in Figure 6.2);3). The pairs are closer to the annotated ones and the values for its residual errors are similar. The residual errors obtained with both MEIR and MPIR are lower and more consistent throughout the successive frames than the ones obtained with the feature-based methods. Additionally, the MDA, MRL and error angle results for both methods are also similar. The results obtained with the PWC-Net are the best in both the visual analysis regarding the accuracy of the matches and the quantitative analysis regarding the values for the residual errors, MDA, MRL, $\overline{\theta_R}$ and $\overline{\theta_T}$.

VII. CONCLUSION

One of the main challenges of WCE technology is the identification of the location and the trajectory of the capsule as it moves through the GI tract, which is especially relevant for the detection of anomalies. This paper explores a few image registration approaches between sequential frames as the first step into the process of overcoming this difficulty.

In this article, through comparison with ground truth information provided by the robot, it is proven that it is possible to determine the relative movement of the capsule between

frames, with the computation of the essential matrix, up to a scale factor, when the image registration method produces enough good matches between sequential frames given this difficult colon dataset. The handling of this dataset was a challenge due to the intrinsic difficulty of successfully achieving suitable point matches with any registration method. Using SIFT, SURF, MSER, and LIOP, in some cases, it was very difficult to extract enough features and matching points for the estimation of the fundamental and essential matrices. The results obtained with this evaluation allow the following conclusions: (1) the most common feature matching approaches used in computer vision are not adequate for these datasets; (2) MPIR and MEIR both perform similarly given that in this experimental procedure the capsule is moved by the robotic arm so there is no need to account for elasticity; (3) The best estimates of the capsule trajectory were obtained using the PWC-Net for the image registration of consecutive frames, which yielded the smallest residual errors, MDA, MRL and error angles.

In the future, we can move to conventional datasets, where the camera is moved by peristalsis, without the need for ground truth information, and adopt this procedure to reconstruct the full trajectory of the capsule. Additionally, further experiments will also be performed with datasets obtained with a variety of endoscopic capsules from different manufacturers.

REFERENCES

[1] X. Dray, D. Iakovidis, C. Houdeville, R. Jover, D. Diamantis, A. Histace, and A. Koulaouzidis, "Artificial intelligence in small bowel capsule endoscopy—Current status, challenges and future promise," *J. Gastroenterology Hepatology*, vol. 36, no. 1, pp. 12–19, Jan. 2021.

[2] I. N. Figueiredo, "Wireless capsule endoscope location and a robotic validation experiment," in *Proc. Medit. Conf. Med. Biol. Eng. Comput.* Cham, Switzerland: Springer, 2019, pp. 1361–1365.

[3] I. Umay, B. Fidan, and B. Barshan, "Localization and tracking of implantable biomedical sensors," *Sensors*, vol. 17, no. 3, p. 583, 2017.

[4] T. D. Than, G. Alici, H. Zhou, and W. Li, "A review of localization systems for robotic endoscopic capsules," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2387–2399, Sep. 2012.

[5] G. Ciuti, R. Caliò, D. Camboni, L. Neri, F. Bianchi, A. Arezzo, A. Koulaouzidis, S. Schostek, D. Stoyanov, C. M. Oddo, and B. Magnani, "Frontiers of robotic endoscopic capsules: A review," *J. Micro-Bio Robot.*, vol. 11, no. 1, pp. 1–18, 2016.

[6] N. Dey, A. S. Ashour, F. Shi, and R. S. Sherratt, "Wireless capsule gastrointestinal endoscopy: Direction-of-arrival estimation based localization survey," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 2–11, 2017.

[7] A. S. Ashour, N. Dey, W. S. Mohamed, J. G. Tromp, R. S. Sherratt, F. Shi, and L. Moraru, "Colored video analysis in wireless capsule endoscopy: A survey of state-of-the-art," *Current Med. Imag. Formerly Current Med. Imag. Rev.*, vol. 16, no. 9, pp. 1074–1084, Dec. 2020.

[8] F. Bianchi, A. Masaracchia, E. S. Barjuei, A. Menciassi, A. Arezzo, A. Koulaouzidis, D. Stoyanov, P. Dario, and G. Ciuti, "Localization strategies for robotic endoscopic capsules: A review," *Expert Rev. Med. Devices*, vol. 16, no. 5, pp. 381–403, May 2019.

[9] M. C. Hoang, E. Choi, B. Kang, J.-O. Park, and C.-S. Kim, "A miniaturized capsule endoscope equipped a marking module for intestinal tumor localization," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 3712–3715.

[10] K. B. Ozyoruk, G. I. Gokceler, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan, "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: Endo-SfMLearner," 2020, *arXiv:2006.16670*. [Online]. Available: http://arxiv.org/abs/2006.16670

[11] R. Szeliski, *Computer Vision: Algorithms and Applications*. Berlin, Germany: Springer, 2010.

[12] R. Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2006.

[13] N. Paragios, Y. Chen, and O. D. Faugeras, Eds., *Handbook of Mathematical Models in Computer Vision*. Berlin, Germany: Springer, 2006.

[14] J. Bergen, P. Anandan, K. Hanna, and R Hingorani, "Hierarchical model-based motion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 1992, pp. 237–252.

[15] S.-L. Liu, J. Kim, B. Kang, E. Choi, A. Hong, J.-O. Park, and C.-S. Kim, "Three-dimensional localization of a robotic capsule endoscope using magnetoquasistatic field," *IEEE Access*, vol. 8, pp. 141159–141169, 2020.

[16] *PillCam Capsule Endoscopy-User Manual, Rapid V.8.0*, IntroMedic, USA, 2015.

[17] Y. Geng and K. Pahlavan, "Design, implementation, and fundamental limits of image and RF based wireless capsule endoscopy hybrid localization," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 1951–1964, Aug. 2016.

[18] G. Bao, K. Pahlavan, and L. Mi, "Hybrid localization of microrobotic endoscopic capsule inside small intestine by data fusion of vision and RF sensors," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2669–2678, May 2015.

[19] L. Liu, C. Hu, W. Cai, and M. Q.-H. Meng, "Capsule endoscope localization based on computer vision technique," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 3711–3714.

[20] E. Spyrou and D. K. Iakovidis, "Homography-based orientation estimation for capsule endoscope tracking," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. Proc.*, Jul. 2012, pp. 101–105.

[21] W. Khan, S. M. L. Kabir, H. A. Khan, A. Al Helal, M. A. Mukit, and R. Mostafa, "A localization algorithm for capsule endoscopy based on feature point tracking," in *Proc. Int. Conf. Med. Eng., Health Informat. Technol. (MediTec)*, Dec. 2016, pp. 1–5.

[22] D. K. Iakovidis, E. Spyrou, D. Diamantis, and I. Tsiompanidis, "Capsule endoscope localization based on visual features," in *Proc. 13th IEEE Int. Conf. Bioinf. BioEng.*, Nov. 2013, pp. 1–4.

[23] M. Aghanouri, A. Ghaffari, and N. Dadashi, "Image-based localization of the active wireless capsule endoscope inside the stomach," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Feb. 2017, pp. 13–16.

[24] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[25] F. Deligianni, "Visual augmentation for virtual enviromnents in surgical training," Ph.D. dissertation, Dept. Comput., Imperial College London, London, U.K., 2006.

[26] E. Spyrou, D. K. Iakovidis, S. Niafas, and A. Koulaouzidis, "Comparative assessment of feature extraction methods for visual odometry in wireless capsule endoscopy," *Comput. Biol. Med.*, vol. 65, pp. 297–307, Oct. 2015.

[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

[28] H. Kartal, U. Alganci, and E. Sertel, "Automated orthorectification of VHR satellite images by SIFT-based RPC refinement," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 6, p. 229, Jun. 2018, doi: 10.3390/ijgi7060229.

[29] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understand.*, vol. 110, pp. 404–417, May 2006.

[31] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 603–610.

[32] I. N. Figueiredo, C. Leal, L. Pinto, P. N. Figueiredo, and R. Tsai, "Hybrid multiscale affine and elastic image registration approach towards wireless capsule endoscope localization," *Biomed. Signal Process. Control*, vol. 39, pp. 486–502, Jan. 2018.

[33] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[34] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[35] A. Śluzek, "Improving performances of MSER features in matching and retrieval tasks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 759–770.

[36] D. S. Aljutaili, "A speeded up robust scale-invariant feature transform currency recognition algorithm," *Int. J. Comput. Inf. Eng.* vol. 12, no. 6, pp. 365–370, 2018.

[37] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.
[38] D. Sun, X. Yang, and M. Liu, "Models matter, so does training: An empirical study of CNNs for optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1408–1423, Jun. 2019.
[39] E. S. Nadimi, M. M. Buijs, J. Herp, R. Kroijer, M. Kobaek-Larsen, E. Nielsen, C. D. Pedersen, V. Blanes-Vidal, and G. Baatrup, "Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy," *Comput. Electr. Eng.*, vol. 81, Jan. 2020, Art. no. 106531.
[40] K. Mardia and P. Jupp, *Directional Statistics*. Hoboken, NJ, USA: Wiley, 2000.
[41] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 930–943, Aug. 2003.
[42] S. Sornapudi, F. Meng, and S. Yi, "Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps," *Appl. Sci.*, vol. 9, no. 12, p. 2404, Jun. 2019.
[43] L. Barducci, J. C. Norton, S. Sarker, S. Mohammed, R. Jones, P. Valdastri, and B. S. Terry, "Fundamentals of the gut for capsule engineers," *Prog. Biomed. Eng.*, vol. 2, no. 4, Sep. 2020, Art. no. 042002.
[44] Y. Hwang, H. C. Lee Park, B. Tama, J. Kim, D. Cheung, W. Chung, Y.-S. Cho, K.-M. Lee, M.-G. Choi, S. Lee, and B.-I. Lee, "An improved classification and localization approach to small bowel capsule endoscopy using convolutional neural network," *Digestive Endoscopy*, vol. 33, no. 4, pp. 598–607, Jul. 2020.
[45] D. R. Cave, S. Hakimian, and K. Patel, "Current controversies concerning capsule endoscopy," *Digestive Diseases Sci.*, vol. 64, no. 11, pp. 3040–3047, Nov. 2019.
[46] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1438–1447, May 2020.
[47] E. Ferrante and N. Paragios, "Non-rigid 2D-3D medical image registration using Markov random fields," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 16. Berlin, Germany: Springer, 2013, pp. 163–170.
[48] U. Mitrović, B. Likar, F. Pernuš, and Ž. Špiclin, "3D–2D registration in endovascular image-guided surgery: Evaluation of state-of-the-art methods on cerebral angiograms," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 2, pp. 193–202, 2018.
[49] E. Spyrou and D. K. Iakovidis, "Video-based measurements for wireless capsule endoscope tracking," *Meas. Sci. Technol.*, vol. 25, no. 1, Jan. 2014, Art. no. 015002.
[50] Q. Zhu, J. Hu, W. Cai, and L. Henschen, "A new robot navigation algorithm for dynamic unknown environments based on dynamic path recomputation and an improved scout ant algorithm," *Appl. Soft Comput.*, vol. 11, no. 8, pp. 4667–4676, Dec. 2011, doi: 10.1016/j.asoc.2011.07.016.
[51] *Optical Flow Prediction With Tensorflow*. Accessed: Jun. 12, 2020. [Online]. Available: https://github.com/philferriere

MARINA OLIVEIRA received the M.Sc. degree in biomedical engineering from the Physics Department, in 2018. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Coimbra. Her research interests include 3-D computer vision, visual odometry, localization and mapping techniques, 3-D reconstruction, capsule endoscopy, and medical image analysis.

HELDER ARAUJO (Member, IEEE) is currently a Professor with the Department of Electrical and Computer Engineering, University of Coimbra. In the last few years, he has been working on non-central camera models, including aspects related to pose estimation and their applications. He has also developed work in active vision and on control of active vision systems. Recently, he has started work on the development of vision systems applied to medical endoscopy. His research interests include computer vision applied to robotics, robot navigation, and visual servoing.

ISABEL N. FIGUEIREDO received the Ph.D. degree in mathematics from University Pierre et Marie Curie, Paris, France, in 1989, with a focus on applied mathematics. She is currently a Professor with the Department of Mathematics, Faculty of Sciences and Technology, University of Coimbra, Portugal.

LUÍS PINTO received the B.S. and M.S. degrees in applied mathematics and the Ph.D. degree in mathematics from the University of Coimbra, Coimbra, Portugal. He studied numerical methods for solving integro-partial differential equations at the University of Coimbra, where he has been with the Center for Mathematics (CMUC), since 2014. He has been involved in several research projects in both image processing and numerical methods for partial differential equations.

EVA CURTO received the M.Sc. degree from the University of Coimbra, in 2018, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. Her research interests include 3-D computer vision, including medical endoscopy, pose estimation, and 3-D reconstruction of non-rigidly deforming objects.

LUÍS PERDIGOTO received the Ph.D. degree in electrical and computer engineering from the University of Coimbra, Portugal, in 2015. He is currently an Assistant Professor with the Department of Electrical Engineering, Polytechnic Institute of Leiria, Portugal. He is also a Researcher with the Institute for Systems and Robotics—Coimbra. His research interests include computer vision, robotics, and automation.