Deep Endoscopic Visual Measurements

Dimitris K. Iakovidis, *Senior Member, IEEE*, George Dimas, Alexandros Karargyris, Federico Bianchi, *Member, IEEE*, Gastone Ciuti, *Member, IEEE*, and Anastasios Koulaouzidis

***Abstract*—Robotic endoscopic systems offer a minimally invasive approach to the examination of internal body structures, and their application is rapidly extending to cover the increasing needs for accurate therapeutic interventions. In this context, it is essential for such systems to be able to perform measurements, such as measuring the distance travelled by a wireless capsule endoscope, so as to determine the location of a lesion in the gastrointestinal (GI) tract, or to measure the size of lesions for diagnostic purposes. In this paper, we investigate the feasibility of performing contactless measurements using a computer vision approach based on neural networks. The proposed system integrates a deep convolutional image registration approach and a multilayer feed-forward neural network in a novel architecture. The main advantage of this system, with respect to the state-of-the-art ones, is that it is more generic in the sense that it is: *i)* unconstrained by specific models, *ii)* more robust to non-rigid deformations, and *iii)* adaptable to most of the endoscopic systems and environments, while enabling measurements of enhanced accuracy. The performance of this system is evaluated in *ex-vivo* conditions using a phantom experimental model and a robotically-assisted test bench. The results obtained promise a wider applicability and impact in endoscopy in the era of big data.**

***Index Terms*—Endoscopy, neural networks, deep learning, deep matching, measurements.**

I. Introduction

SINCE 2000, with the advent of wireless capsule endoscopy (WCE) [1], the assessment of the gastrointestinal (GI) tract has undergone a major transformation with the application of robotic technology and disruptive solutions. Nowadays, WCE is considered an attractive alternative to traditional "cabled" GI endoscopy, due to its capability to minimize patient discomfort, eliminate risk of perforation, and sedation-related complications; therefore, it could be a useful solution in increasing the take-up of population-based GI screening programs. Currently, WCE is primarily used in the minimally-invasive examination of the small bowel, with lesser application in the oesophagus, stomach, and colon [2].

To date, several companies, research institutes and universities have explored the field of WCE by developing new principles and solutions for actuation, localization, sensing, and data telemetry [3]. In this context, a variety of robotic WCE systems has been proposed [4]. These include a capsule device which is equipped with three miniature legs, each carrying a wheel [5]; a legged walking tele-operated robotic capsule [6]; magnetically-driven robotic capsules [7], [8] and; robotic capsules equipped with reservoirs and mechanical parts for drug delivery [9][10]. However, despite major forward strides, there are still several technological limitations, mainly with respect to: *i)* accurate localization [11]; *ii)* accurate navigation of the capsule inside the bowel lumen; *iii)* ability to provide treatment [3]; *iv)* identification and characterization of bowel pathology [12]; and *v)* ability to accurately measure lesion size due to the two-dimensional nature of the WCE images. In light of the above, size measurement and localization capabilities are considered important prerequisites to overcome the rest of the aforementioned limitations. The vast majority of decision-making in GI endoscopy integrates lesion size and localization information. Internal localization of the endoscope, derived by measuring its displacement from landmark anatomical structures (*e.g.*, distance from the rectum, incisor teeth or pylorus), is essential for lesion localization, accurate navigation, and treatment delivery. Size measurement is also essential for management, *e.g.*, larger polyps have a higher likelihood of malignancy.

In this paper, we investigate a novel computer vision approach to enable contactless *in-vivo* travel distance and size measurements estimation in endoscopy. Going beyond the state-of-the-art [12][13][14][15][16], we propose a system that is entirely based on an artificially intelligent structure enabling adaptability to different conditions, such as the camera parameters and the environment under inspection. This provides a cost-efficient means of enhancing current endoscopic systems with localization and measurement capabilities, since it does not rely on any additional hardware. Also, it contributes to the longer vision of developing robust endoscopic systems with perception of the environment [17].

The rest of this paper is organized in four main parts: Section II provides an overview of the relevant state-of-the-art visual measurement systems. Section III presents the architecture of the proposed intelligent visual measurement system and its application framework. The performed experiments and the obtained results are presented in Section

This paper was submitted for review on June, 30, 2018. The work described in this paper was partially supported by the European Commission within the framework of the "Endoscopic versatile robotic guidance, diagnosis and therapy of magnetic driven soft-tethered endoluminal robots" Project, H2020-ICT-24-2015 (EU Project-G.A. number: 688592). The authors thank all the collaborators of the EU project.

D. K. Iakovidis and G. Dimas, are with the Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece (e-mail: diakovidis@uth.gr, gdimas@uth.gr).

A. Karargyris, is with the IBM Research , San Jose, CA, USA (e-mail: akarargyris@gmail.com).

G. Ciuti and F. Bianchi, are with The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy (e-mail: gastone.ciuti@santannapisa.it, federico.bianchi@santannapisa.it).

A. Koulaouzidis, is with the Endoscopy Unit, Royal Infirmary of Edinburgh, Edinburgh, UK (e-mail: akoulaouzidis@hotmail.com).

IV. Discussion and conclusions, derived from this study, are summarized in the last section, Section V.

II. Visual Measurement Systems in GI Endoscopy

A. Previous Work

In the current literature, the most popular method for camera calibration is the one presented by Zhang [18]. The setting requires a planar pattern (*i.e.* chessboard) and the camera records video from at least two different orientations. Among different orientations, either the camera or the plane is moving. The method does not require knowledge of camera motion. The radial lens distortion is also modelled in. Other works, such as Hu *et al.* [19] and Kannala-Brandt [20], have tried to address lens distortion but they require knowledge of focal length and optical field of view.

In the domain of WCE, Iakovidis *et al.* [21] has shown that a Visual Odometry (VO) approach, applied to the GI tract, is feasible as the error in scaling parameters can remain low enough to be considered practical. In later works, Spyrou *et al.* [22][23] investigated various methods to extract key-points and features on WCE video frames and showed that the scale-invariant feature transform (SIFT) approach provides a more accurate localization performance. The reported errors were of the order of 10^{-3} on average; however, the measurements performed were only relative, not in physical units, and CE motion was simulated by image rotation and scaling.

Works such as those of Bao *et al.* [24] and Mi *et al.* [25] used simulated experiments to estimate and model the locomotion of a Capsule Endoscope (CE) using a virtual camera and tube. Bao *et al.* [24] showed that they can estimate the speed of the endoscopic capsule with an accuracy of 93%, while the CE localization accuracy was less than 2.71cm on average. However, the main concern with both studies is that these models are exclusively based on simulations.

We recently proposed a novel VO approach to estimate the displacement of a CE in physical length units [12]. This methodology relied on the camera calibration method of Kannala-Brandt [20], helping to address the lack of knowledge of the WCE intrinsic parameters. Based on real CE data, we achieved a maximum value of Mean Absolute Error (MAE) of the actual distance covered by the CE equal to 7.2±1.4cm.

Artificial Neural Networks (ANNs) have been extensively used to address camera calibration. Ahmed *et al.* [26] used a 3-layer (1-hidden layer) Feed-Forward Neural Networks (MFNNs) with sigmoidal activation functions. This network was designed to map 3D world coordinates of a 2D image plane. The weights of its hidden layer corresponded to the extrinsic camera parameters, while the output layer to the intrinsic ones. The root mean squared calibration error of that approach was 0.092. In [27] a 3-layer MFNN was utilized to solve a similar coordination mapping problem by performing 3D reconstruction instead of camera calibration. In that study it was observed that the error was following a linear increase as distance between the object and the camera was increasing beyond the range of the training distances. Besdok [28] proposed a Radial Basis Function (RBF) network to train a genetic algorithm approach for camera calibration.

In the context of depth estimation with machine learning techniques, Nadeem and Kaufman [29] proposed a machine learning algorithm to create a depth map image from an image captured during traditional or capsule colonoscopy (direct inspection of the endoscopic inner surfaces). The depth map was used to detect and depict the boundaries of a polyp in the given image. However, their framework focuses on the relative depth values.

Because ANNs provide good performance in camera calibration, their use has been extended to the domain of VO in WCE. Dimas *et al.* [13] used a relatively large network with 24 inputs, 1000 hidden neurons and 3 output neurons (denoted as a 24-1000-3 architecture) to estimate scaling between 2 points in consecutive CE video frames using pixel intensity as an additional feature to corresponding point coordinates. In [16], this work was further extended by investigating how different color elements and matching algorithms affect the performance of the network. The Kanade-Lucas-Tomasi (KLT) – RANdom SAmple Consensus (RANSAC) scheme for the point matching, using a 30-5-3 network architecture with CIE-*Lab* color components resulted in best accuracy. This method outperformed our previous geometric localization approach [12], with an average localization error of 2.70±1.62cm.

Applications of ANNs have also been proposed in the context of depth map estimation from regular images. Garg *et al*. [30] presented a Convolutional Neural Network (CNN), which is a deep ANN architecture, for monocular depth estimation. In their work they also proposed an unsupervised way for the training of the network by using images captured with a stereo camera rig. The minimum relative error reported in their paper was 0.169. In endoscopy, CNNs have been employed for depth estimation by Mahmood and Durr [31]. They combined a CNN with Conditional Random Fields (CRF) for the estimation of depth maps and topographical reconstruction of monocular endoscopic images. This joint architecture does not require any assumption on the geometric model of the camera. Their model was trained on a set of synthetically-generated data and tested on both real and virtual endoscopy data. The reported relative errors for virtual and real endoscopy data were 0.183 and 0.242, respectively. More recently, Mahmood and Durr [32] proposed another deep learning-based method that does not require any hand-crafted features. For the training of this a model, a large amount of augmented data was required. To cope with this problem, they generated training images, using a synthetic, texture-free colon phantom. The relative depth estimation error for the phantom test data was reported to be 0.164.

Quite a few works have been published describing methods to measure the dimensions of lesions in the GI tract. A recent work by Vakil *et al*. [33] performed a study which compared 3 types of measurements: *i)* empirical measurements conducted by doctors, *ii)* measurements based on open-biopsy forceps, and *iii)* measurements based on image processing. The camera was calibrated by establishing a relation between the known dimensions of the forceps' open jaws and the physical dimension of lesions measured in pixels. The percentage MAE on measurements on modelled data for open-biopsy forceps and image processing based technique was 41.8±23.3% and 1.8±2.2%, respectively. Same data in *in-vivo* condition are 26.5±5.7% and 2.8±3.2%. More recently, Zhou *et al.* [34] studied polyp detection and radius measurement in small

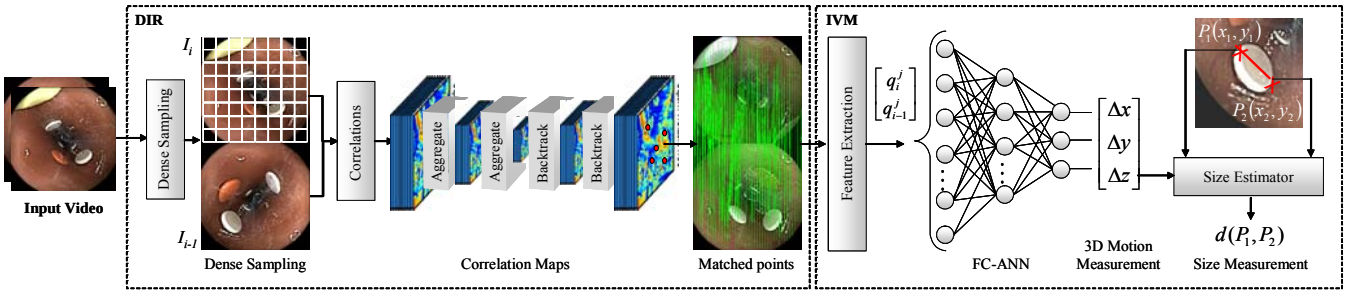

Fig. 1. Deep visual measurement system architecture. Pairs of consecutive video frames I_i and I_{i-1} , i=1,2,… are used as input. In the DIR module I_i is densely sampled and the correlation of each sample (patch) with I' is estimated via convolution to form respective correlation maps. Several aggregation layers are used for the construction of a multi-level pyramid of correlation maps of different scales. By backtracking the local maxima of the correlation maps, at each level, a set of matching points between the two input video frames are estimated. In IVM coordinate and color features are extracted from the matched points, which are subsequently used to estimate the 3D CE motion and the size of objects indicated by the user, *e.g.*, by drawing a line segment P_1P_2.

intestine for CE videos. Their methodology relied on establishing a relationship between the visual sharpness of the target object and its dimensions. The reported error ratio of radius calculation was 9.77%. Park *et al.* [35] proposed measuring the size of gastrointestinal lesions by a relation similar to the one used in [33]. The MAE reported was of 0.26±0.21mm. Lastly, a novel device was proposed by Goldstein *et al.* [36] to measure polyps' dimensions. They used a virtual tape and not a physical object as reference. The upper limit of the 90% of the absolute difference of the estimation using the virtual tape and the reference was of 0.55±0.31mm. Another method for polyps' measurement, presented by Visentini-Scarzanella *et al.* [37], was based on structured light.

The literature review performed in this section reveals that current visual CE localization and *in-vivo* size measurement methods are still at an early stage. Still open research issues limiting the real-world applicability of the current visual localization methods mainly include further robustness and accuracy enhancement and the development of more realistic experimental setups enabling measurement validation with ground truth data. With respect to the size measurements in physical units, all the current methods are based on external references, such as forceps or virtual tapes.

B. Contributions of this Work

Considering the identified open research issues and the effectiveness of the machine-learning-based measurement approaches reviewed in the previous sub-section, contributions of this paper beyond the state-of-the art include:

- We propose an ANN-based system architecture that unlike state-of-the-art systems enables both distance and size measurements.
- We investigate an unsupervised deep convolutional approach that is robust to non-rigid deformations for more accurate and less parametric CE motion estimation.
- We propose a size measurement methodology that exploits motion estimation over a video frame sequence so as to avoid the use of external references.

III. Deep Visual Measurement System

The proposed visual measurement system is based on deep image analysis and ANNs (Fig. 1). It consists of two modules: *i)* the Deep Image Registration (DIR) module, and *ii)* the Intelligent Visual Measurement (IVM) module. The first one determines matching points between consecutive video frames, and the second one establishes a mapping between the 2D motion of the matched points and the 3D motion of the endoscope within the intestine, to perform travel distance and size measurements.

A. Deep Image Registration Module

An important step for accurate visual measurements is the detection of points of interest in pairs of consecutive video frames and their registration by finding correspondences (matches) between the interest points.

The DIR module is based on Deep Matching (DM) [38]. It is inspired by the Deep Convolutional Neural Networks (DCNNs) but it does not require any training. It is a fully unsupervised method aiming to discover correspondences between images. The use of the DM in DIR is motivated by: a) its improved matching performance over state-of-the-art matching algorithms on benchmark datasets [38]; b) the fact that it is non-parametric and that it does not depend on any model; c) it can handle non-rigid deformations and efficiently determine dense correspondences in the presence of significant changes between images. The latter is important for endoscopy applications since there are a lot of tissue deformations and floating objects, *e.g.*, debris, contributing to false motion estimations among consecutive images.

The methodology of DM relies on a hierarchical, multilayer, correlation architecture [38]. Given two images I and I' of size $W\times H$, image I is split into non-overlapping atomic patches (samples) $I_{N,p}$ of size $N\times N$ ($N=2^l$, l=2,3,…) centred at $\boldsymbol{p}\in G_N$, where G_N represents a grid of points, between pixels, corresponding to the centres of the patches, *e.g.*, $G_4=\{2,6,10,\ldots W\text{-}2\}\times\{2,6,10,\ldots H\text{-}2\}$ (Fig. 2) with the first centre to be at the grid point with coordinates $(2,2)^{\mathrm{T}}$, which is located in between of four pixels $I(2,2)$, $I(2,3)$, $I(3,2)$ and $I(3,3)$. The correlation between each patch of I at every location of I' is computed to obtain a corresponding correlation map by convolution. This convolution can be expressed as $C_{N,p}=I^F_{N,p}*I'$, where F denotes a horizontal and vertical flip. The resulting correlation maps $C_{N,p}$ are subsequently used to build upper-level, correlation maps, and iteratively develop a multi-level correlation pyramid in a bottom-up way. The upper-level correlation maps are smaller

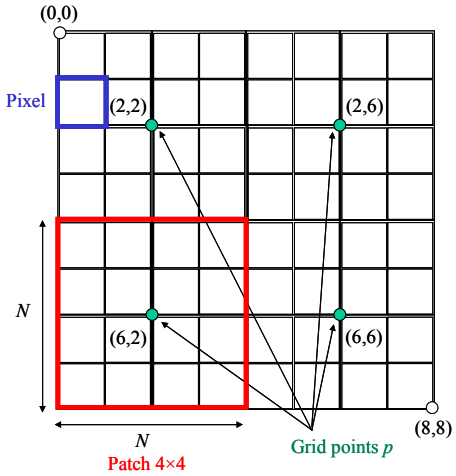

Fig. 2. Example 8×8-pixel image (W=H=8), with the small squares to represent its pixels. The green points represent the G_4={2,6}×{2,6} grid points for this image, as defined in our manuscript. The respective patches are of 4×4 pixels (N=4).

but they are obtained from larger image patches. These patches are formed by concatenation of quadruplets of patches centred at each $\boldsymbol{p}$. Thus, for N=2^l, l>2, patches of size $N/2\times N/2$ are being concatenated to form patches $I_{N,\boldsymbol{p}}$ of size $N\times N$. The iterative concatenation process of the patches and the correlation maps $C'_{N,p,i}$, i=0..3, corresponding to these patches can be described by the following equations:

$$I_{N,p} = \left[I_{\frac{N}{2},p+s_{N,i}} \right]_{i=0..3}, \; C'_{N,p,i} = C_{\frac{N}{2},p+s_{N,i}} \qquad (1)$$

where $s_{N,i} = No_i/4$ describes the positional shift of a children patch $i\in[0, 3]$ relatively to its parent patch, with $o_0 = [-1,-1]^{\mathrm{T}}$, $o_1 = [-1,1]^{\mathrm{T}}$, $o_2 = [1,-1]^{\mathrm{T}}$, $o_3 = [1,1]^{\mathrm{T}}$. The upper-level correlation maps of the patches $I_{N,\boldsymbol{p}}$ are computed from their children's correlation maps $C'_{N,p,i}$, i=0..3, as follows:

$$C_{N,p} = R_\lambda\left(\frac{1}{4}\sum_{i=0}^{3}\left(T_{o_i}\circ P\right)\left(C'_{N,p,i}\right)\right) \qquad (2)$$

where R_λ denotes a power transform known as rectification [39], and $T_{o_i}\circ P$ denotes the combination of a translation operator T_{o_i} with a max-pooling operator P. More specifically, each correlation map $C'_{N,p,i}$, i=0..3, is translated by one pixel towards the direction indicated by o_i, and undergoes max-pooling using a 3×3 filter with stride 2, which dyadically downscales it. Equation (2) estimates the average of these maps and rectifies it to obtain $C_{N,p}$. This process is repeated while $N<\max(W,H)$.

A score $S = C_{N,p}(p')$, calculated by (2) in the multi-level correlation pyramid, represents a deformation-tolerant similarity of two patches $I_{N,p}$ and $I'_{N,p'}$. Considering this as an entry point in the pyramid, atomic matches between I and I' can be obtained by backtracking local maxima in the correlation maps and undoing the steps used to aggregate them during the pyramid construction. The atomic matches, obtained from different entry points in the pyramid, are subsequently merged and the matches with the lower similarity scores are discarded.

Similarly with our previous methods [12], [13], [15], [16], [21–23], on the sets of matched points found using DM, RANSAC can be applied in order to keep only the matches which follow a certain geometrical pattern, and in that sense remove outliers. However, the architecture illustrated in Fig. 1, does not include RANSAC because our experimental results (Section IV) show that RANSAC can be omitted. In the context of this work, images I and I' represent a pair of consecutive endoscopic video frames I_i and I_{i-1}, i=1,2,… .

B. Intelligent Visual Measurement Module (IVM)

The IVM module is based on a Fully Connected feed-forward ANN (FC-ANN) enabling the measurement of the 3D motion of the endoscope, which is used for the estimation of the travel distance of the endoscope within the GI tract. In addition, this information is exploited to perform size measurements of clinical findings with a novel methodology.

1) 3D Motion Measurement

Motion measurement is performed with a computationally simple, non-linear, and adaptive to any camera model, sub-system. It is based on a 3-layer FC-ANN architecture (with one hidden layer), which is well-known for its universal approximation capacity [40].

The FC-ANN receives as *input* a set of features extracted from the matching points detected by the DIR module in pairs of consecutive endoscopic frames. Given a minimum set of m matched points per frame, 5 features are extracted per point forming a feature vector q: *i)* its coordinates (u, v), and *ii)* its color values expressed in CIE-*Lab* color space (L, a, b). The use of color information has been proved useful in modelling the appearance of the tissue as a function of its distance from the endoscope [16]. The dimensionality of the input layer of the FC-ANN is determined by the concatenation of the extracted features from the two frames into a single feature vector $N = 5\times2\times m$. The minimum number of matching points m is determined as the least number of matches per frame pair found in the training set. If the number of matching points in a frame pair is larger than m, the extra matches are divided into groups of m matches, forming respective groups of N-dimensional input vectors q^j, j=1,2… up to the largest integral multiple of m, *e.g.*, if the number of matches found is 10, and m=3, a total of three N-dimensional input vectors will be formed, and the remaining matches are discarded. The FC-ANN inputs should be provided in a consistent way preserving the relative spatial information between the consecutive frames, *e.g.*, in our study we have considered that the first half of the inputs are obtained from the points of frame I_i and the second half from the respective points of the next frame I_{i-1}.

The hidden along with the output layer of the FC-ANN perform a mapping of the spatial and color components of the corresponding points, provided in its input, to the 3D relative displacement of the CE. The aggregation of the inputs performed by the FC-ANN focus on this mapping, disregarding the within-frame spatial relations of the points, which for our application are considered of lower relevance. The number of neurons in the hidden layer of the FC-ANN is an experimentally determined parameter of the system.

The dimensionality of the *output* layer of the FC-ANN is equal to that of the 3D real-world coordinate space of the

endoscope. It is composed of 3 neurons, each of which outputs a measurement for the motion of the endoscope along a different axis of the Cartesian coordinate system, *i.e.*, $\Delta x_i=x_i-x_{i-1}$, $\Delta y_i=y_i-y_{i-1}$ and $\Delta z_i=z_i-z_{i-1}$. These measurements are based on the 2D motion and appearance of the matching interest points found in the examined pair of consecutive input video frames I_i and I_{i-1}. The total distance travelled by the endoscope from a video frame I_i to a video frame I_{i+n} can be estimated by summing up the Euclidean distances between the intermediate consecutive pairs of frames:

$$D(I_i, I_{i+n}) = \sum_{k=i+1}^{i+n} \sqrt{(\Delta x_k)^2 + (\Delta y_k)^2 + (\Delta z_k)^2} \qquad (3)$$

Our approach, handles both forward and backward motion in the same way, taking into consideration only the absolute displacement and not the direction of the motion. The direction can be extracted from the differences in the scale of matched points. Let $P(x, y)$ a point in image I and $P'(x', y')$ the corresponding (matched point) in image I'. In order to determine if the CE is moving forward or backwards, we just need to calculate the L^2 (Euclidean) norm of each point. If the $L^2(P)/L^2(P') < 1$, then the CE is moving forward, if the $L^2(P)/L^2(P') > 1$, the CE is moving backwards.

2) Size Measurement

In order to determine the actual size of an object within an endoscopic image we need to establish a relation between its size measured in pixels and its size measured in physical units. By using the pinhole camera model, a world point (x, y, z) is projected to the image plane as follows [41]:

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \frac{f}{z} \begin{pmatrix} x \\ y \end{pmatrix} \qquad (4)$$

where f is the effective focal length of the camera. From the projection $(\tilde{u}, \tilde{v})^T$, we can obtain the corresponding image coordinates $(u, v)^T$ in pixel units with the respective transformation:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} D_u s_u \tilde{u} \\ D_v \tilde{v} \end{pmatrix} + \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \qquad (5)$$

where s_u is a scale factor, and D_u and D_v are coefficients needed to change the metric units to pixels. The vector $(u_0, v_0)^T$ represents the principal centre of the image. By replacing the projection $(\tilde{u}, \tilde{v})^T$ with respect to the world points (x, y, z) in (4) we have:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{f}{z} \begin{pmatrix} D_u s_u x \\ D_v y \end{pmatrix} + \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \qquad (6)$$

By solving the linear system with respect to x and y the following relations are obtained:

$$x = \frac{z(u-u_0)}{f D_u s_u}, \quad y = \frac{z(v-v_0)}{f D_v} \qquad (7)$$

In order to solve (7) for x and y we need to know the distance z between the camera and the world points, which are estimated using the ANN described in the first part of Section III.B. The quantities $f_x = fD_u s_u$, $f_y = fD_v$ and $(u_0, v_0)^T$ are estimated by following the camera calibration procedure described in [18]. Once we calculate the x and y values of two world points, denoted as $P_1=(x_1, y_1, z_1)$ and $P_2=(x_2, y_2, z_2)$, the length of the linear segment P_1P_2 can be measured by the Euclidean distance:

$$d(P_1, P_2) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2 + (z_2-z_1)^2} \qquad (8)$$

This distance is calculated for the size measurement of objects visible through the endoscope as follows: *i)* choose a video frame I_i where the object to be measured is clearly visible and select two points, $p_1 = (u_1, v_1)$ and $p_2 = (u_2, v_2)$, defining a linear segment on the object to be measured, *e.g.*, its diameter; *ii)* for each point p_l, l = 1, 2 perform a 3D motion measurement as described in the first part of Section III.B, from frame I_i to the first frame I_{i+n} where the point disappears, and set $z_l=\Delta z_i$, l = 1, 2; and *iii)* calculate x_l, y_l by substituting z_l in (6), and $d(P_1, P_2)$ by substituting the calculated x_l, y_l, z_l in (7). This methodology considers that lens distortion is corrected before the estimation of the (x_j, y_j, z_j); thus, it is directly applicable on commercial endoscopes.

IV. Experiments and Results

Experiments were performed to investigate the feasibility of performing travel distance and size measurements with the proposed methodology. Differently to previous experimental validation approaches, which were based on computer simulations, the experimental design of this study aims to provide a less ideal environment, with the necessary ground-truth information, to perform both motion and size measurements in real-world units.

A. Robotically-Assisted Experiment

The proposed experimental workbench aims to reproduce a real capsule-based endoscopic procedure but in a controlled and repeatable operating environment. A WCE, stably placed through a plastic rod on the end-effector of an accurate industrial robotic arm, was guided in a small bowel phantom. Several colored targets were located inside the bowel phantom, properly fixed to a custom-made support, to reproduce anatomical landmarks. The setup includes five main components (Fig. 3), discussed in detail below:

1) An accurate six degrees-of-freedom (DoFs) industrial robotic arm (RV-3SB robot, Mitsubishi, Tokyo, Japan) moves the WCE forward and backward inside the bowel phantom (Fig.3a). The robotic arm was programmed to move with controlled velocity and acceleration.
2) A rectilinear straight plastic rod attached to the end-effector of the industrial robotic arm. The rod, which houses the WCE in its final part, has a length (*i.e.* 330 mm) such as to cover the entire bowel phantom during locomotion (Fig. 3a).
3) A Pillcam® SB3 CE system (Medtronic, Minnesota, USA), including an image-receiving belt (Fig. 3b) and an endoscopic capsule (Figs. 3a-b) with the following

features: i) adaptive frame rate between 2 and 6 fps; ii) size of 11.4×26.2mm; iii) CMOS image sensor with a resolution of 320×320 pixels; iv) field of view of 156°; and v) depth of field between 0-30mm [2].

4) A 30cm-long double layer LifeLike bowel phantom (LifeLike, Biotissue Inc., Ontario, Canada) with an elliptical shape of approximately 22×30 mm (Fig. 3b). Inside the bowel phantom, twenty-four artificial colored pins were randomly arranged along four parallel lines with four different colors (Fig. 3c): 4 red, 13 white, 3 blue and 4 yellow. Each target, fixed with a plastic butterfly clutch in the external part of the phantom, presents circular shape with a head diameter of 0.95cm.
5) A custom-made support keeps the ends of bowel phantom, subsequently stretched in order to linearize it as much is possible while the CE is moved along its axis (Fig. 3b).

After properly placing the LifeLike bowel phantom on the custom-made support and establishing the starting point of the tests, all the distances between the starting point and the centre of each target were accurately measured with a digital calliper. The complete test involves forward and backward movements according to a number of pre-established incremental steps necessary to cover the entire length of the bowel phantom. Each single movement was performed at a constant velocity, while a constant time was elapsed between one-step and the next to allow the acquisition of a sufficient number of frames (approximately 10 frames) per position. The same procedure was repeated 6 times at different velocities: 0.5, 1 and 2mm/s. The correlation between the capsule position and collected images was calculated considering the timestamp of each frame and the motion velocity of the endoscopic capsule, moved through the external robotic arm.

B. 3D Motion Measurements

Several experiments were conducted for the evaluation of the proposed system with regards to its performance in measuring travel distances. The FC-ANN architecture of IVM was 30-5-3. The number of input neurons was N=30 since the least number of matches per frame pair was m=3 (Section III.B). The number of hidden neurons was selected as the least one minimizing the MAE for the particular input dimensionality [16].

A total of 12 video sequences were obtained using the experimental setup described in Section IV.A. A 5-fold Cross Validation (CV) experiment was performed using 8 video sequences out of the 12. The early stopping approach was used to avoid overfitting; thus, for each fold we split the dataset with proportions of 80% for training and 20% for validation according to the Pareto principle. There were not any overlaps of the frames used among the different folds and the subsets used for training and validation. The validation set included 4 videos where the CE was moving forward with velocities of 0.5 mm/s (two videos) and 1 mm/s (two videos), and 4 videos where the CE moving backwards with the same motion patterns. The remaining 4 out of the 12 videos were used for testing the performance of the proposed system on new data (unknown to the system). The velocity of the CE in these videos was 2 mm/s. The two of them were obtained using forward motion and the other two were obtained using backward motion. The performance of the proposed system was evaluated in terms of MAE, which is estimated from the output errors of FC-ANN, obtained over the total travel distance measured in the test sets against the ground truth. In order to avoid any dependence of the results from the color of the pins the correspondences between the images that were falling within the colored pins or on their edges, were excluded from all the experiments.

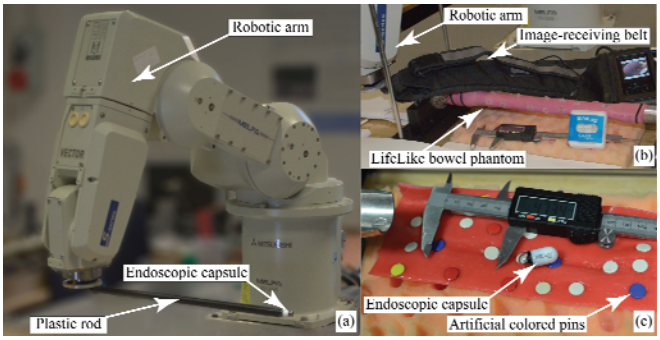

Fig. 3. (a) Detail of the robotic arm used during the experimental tests with the end-effector mounted on the plastic rod and the WCE at the end of the rod. (b) Overview of the robotically-assisted experimental setup including the robotic arm, plastic rod, image-receiving belt, endoscopic capsule, LifeLike bowel phantom held by a custom-made support. (c) Phantom bowel opened after the experiments with the colored pins on the four parallel lines in evidence.

We compared various DIR approaches including DM-RANSAC, DM without RANSAC, and SIFT-KLT-RANSAC [16]. The results on the test datasets of the 5-fold CV on average as estimated from the different folds, are presented in Table I. It can be noticed that DM-RANSAC provides the lowest average MAE over all DIR approaches. The results of the best performing FC-ANN determined by that CV process are summarized in Table II. This table presents the absolute errors obtained on the 4 testing videos that were not included in the CV process. In two of them the CE was moving forward (Front 1, Front 2) and in the other two the CE was moving backwards (Back 1, Back 2). In the last row of this table the MAE estimated per DIR approach from these 4 datasets is reported. The lowest MAE was obtained with the DM-based DIR approaches. Although DM without RANSAC produced a higher average MAE in the 5-fold CV evaluation, it resulted in the lowest MAE in Table II; however, its MAE can be considered comparable to that of DM-RANSAC. This is an indication that RANSAC could be omitted given a well-trained FC-ANN, aiming to the reduction of the overall computational complexity. Figure 4 illustrates the trajectories of the CE reconstructed by the proposed system (indicative results corresponding to the results of test video 'Front 1' of Table II). It can be noticed that both the compared methods produced very small errors (of the order of 10^{-3}cm) on X and Y axes, but the smaller one is achieved by DM.

C. Size Measurements

The performance of the proposed system in size measurement was evaluated by measuring the diameter of the pins attached to the interior of the lifelike bowel. The focal length of the CE was not known a-priori and it was determined by Zhang's [18] camera calibration procedure, as implemented

in Bouget's camera calibration toolbox [33], [43]. In pixel units, the focal length was estimated to be f_x=148, f_y=146 and the principal point pp_x=166 pp_y=152. After this estimation, we performed an optimization routine for the f_x and f_y values to achieve optimum results. The optimization was performed on the pins of the training set. The required ground truth of the z_l values of the camera towards points p_l defining the linear segment to be measured (Section III.B) was known. For the refinement of the f_x and f_y, we performed 20 measurements of the diameter of the pins. Instead of finding directly the x_l and y_l and regarding that the z_l and the diameter of the pin was known, we replaced x_l and y_l in (7) with (6). After iterating for different values of f_x and f_y they were settled at f_x=155, f_y=155. Experimentally, we observed that when the camera was measuring the diameter of the pins from a distance greater than ΔZ = 3cm, the error over the distance ratio was constant with a value of r = 0.1. This could be attributed to the limited image resolution. As the distance of the objects from the camera increases, they become smaller and the uncertainty in the size measurement of the objects (in pixels) increases. The effect of this increase was observed to be constant for the distances for which the pins were visible in the images. To cope with this systematic error, we set a correction factor defined as $c = 0.1 \cdot \Delta Z$ which is subtracted from the final result if ΔZ > 3cm. For the testing step, the DIR/IVM with and without the assistance of RANSAC, was used for the size estimation procedure. We chose these two approaches because they provided the most accurate 3D motion estimations.

In total we measured the diameter of 32 pins observed on the videos used for the testing of the DIR modules (their diameter was known and equal to 0.95cm). The distance from the point of observation of the camera, to the pins was unknown. For the estimation of z values we used and compared both DM and DM-RANSAC based approaches. The MAE for measuring the diameter of the pins using the DM approach to estimate the $z$values was of 0.19±0.18cm. By using the DM-RANSAC approach for measuring the same pins, we obtained a MAE of 0.23±0.18cm. Figure 5 illustrates the size measurement errors per pin, produced using these DIR approaches. The dashed lines indicate the respective MAEs of the pin diameter measurements. The black solid line represents a 30% error threshold, which is defined as a significance level in [44] (discussed in Section V). It can be noticed that in most cases both the compared DIR approaches result in an absolute error of less than 0.31cm (30% of the 0.95cm pin diameter).

TABLE I
5-FOLD CV MEAN ABSOLUTE ERRORS (IN CM)

Methods	**DM-RANSAC**	DM	SIFT-KLT-RANSAC [16]
MAE	**2.11±1.45**	4.49±1.55	3.72±2.19

TABLE II
ABSOLUTE ERRORS PER VIDEO OF THE TEST DATASET (IN CM)

Methods/ Datasets	DM-RANSAC	**DM**	SIFT-KLT-RANSAC [16]
Front 1	0.70	**0.13**	2.91
Front 2	**2.68**	3.04	4.83
Back 1	3.02	**1.35**	2.08
Back 2	**0.25**	0.32	0.98
MAE	**1.66±1.39**	**1.21±1.33**	2.70±1.62

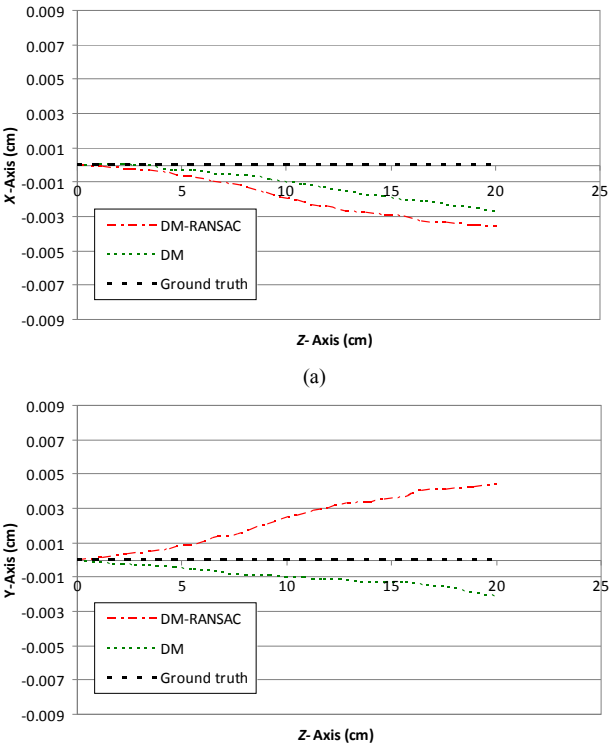

Fig. 4. Estimated trajectory of the camera using different DIR approaches based on video Front 1 (a). The trajectory of the endoscope with respect to Z and X axes. (b) The trajectory of the endoscope with respect to Z and Y axes. The error of all approaches on the X and Y axis is of the order of 10^{-3} cm.

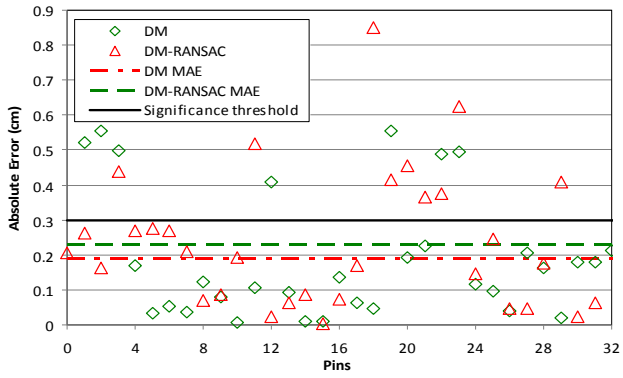

Fig. 5. Size measurement error per pin of the test dataset.

V. DISCUSSION AND CONCLUSIONS

We presented a novel system for *in-vivo* contactless measurements in endoscopy. Novel contributions include a system architecture entirely based on deep analysis of endoscopic video and neural networks, and a size measurement methodology that exploits 3D motion estimation over a video frame sequence. Advantages of the proposed system over the state-of-the-art includes *i)* enhanced motion measurement performance, *ii)* size measurement capability, and *iii)* enhanced domain adaptability by being independent from handcrafted features (such as SIFT) and geometric

models. The feasibility and the accuracy of the measurements were assessed with a set of experiments that include a robotically-assisted setup enabling validation of measurements in real-world units.

The experimental evaluation of the proposed system was performed with a CE, using a robotically-assisted experimental test bench. This test bench provided the necessary ground truth information for the validation of the performed measurements. However, the utility of the proposed system is not limited to WCE. The main reason for choosing a CE instead of a conventional, flexible, endoscope is that it is more challenging. Since CEs are wireless, the cues for estimating their travel distance within the GI lumen are limited. Current sensor-based CE localization approaches are useful in determining their position within the 3D abdominal space, which can be indirectly used for the approximation of the location of the CE within the GI lumen [11]. In our previous studies [13][15][16] we showed that the visual measurement of travel distances can enhance the localization accuracy of CE within the GI lumen. In this study we showed that the proposed system, which is more generic than the state-of-the-art systems of this kind can be even more accurate, and it can also be used for size measurements. This makes it a useful tool for medical decision support (*e.g.*, for size measurement of polyps, where size is a malignancy risk factor), which along with computer-aided abnormality detection and recognition methods [11] can be used for automated analysis of the large volumes of video frames produced by WCE procedures in the era of big data.

The advantages of the proposed system discussed in the beginning of this section are complemented by its enhanced accuracy over the recent relevant works [12][13][15][16]. As it can be derived from the results presented in Table II the MAE achieved by the proposed VO approach is almost half (45%) of that achieved by the most recent approach [16]. It is also notable that this significantly higher accuracy is achieved without the use of RANSAC, which introduces additional computational complexity. Furthermore, unlike studies using simulated data, such as [24],[25], our study was based on real images captured by the CE using an artificial bowel phantom. In other studies [14],[21–23] the motion of the CE was estimated in relative scales and not in physical units (cm) as in our study. Depth map estimation methods, such as [29],[30],[31],[32], could be used in the measurements context investigated in this paper, since the presented experiment addresses motion estimation along the Z-axis. However, the problem of CE motion estimation is 3D, and the proposed system enables the CE motion estimation in all (X, Y, Z) axes. Also, these depth map estimation methods currently provide relative scale estimations.

With respect to size measurements, in [44] of the sizes of different polyps were measured by two groups of medical experts; a group of endoscopists and a group of pathologists. Errors greater than 30% of the size of the objects, *i.e.*, the polyps, measured (Fig. 5), were considered significant enough to be characterized as inaccurate. Based on this consideration endoscopists, provided a 20% of inaccurate polyp measurements, whereas the respective percentage for the pathologists was only 4%. Considering the 30% threshold set in [44] over the size of the objects measured, which in our study are the colored pins, the inaccurate size measurements using DM-RANSAC and DM were 28.1% and 21.8% of the pins, respectively.

In comparison to the state-of-the-art techniques enabling size measurements in physical units [33],[35],[36] the size measurement approach implemented by the IVM module of the proposed system enables the estimation of the size of an object of interest without any external reference, such as forceps and virtual tapes. Furthermore, the proposed approach is more suitable for WCE, where forceps or other tools are not yet available (they are available only in concept CE models [3]). The error percentages with respect to the size measurement of the objects in these studies (Section II) were generally lower than the respective percentage achieved using DM (20±18.9%). However, their results are not directly comparable with the results in our study since they have been obtained with different datasets and experimental setups.

We have to recognize that the validation of the proposed VO estimation methodology was based on an experimental set up that had some limitations. Namely, the bowel phantom was placed in a straight line at a constant velocity each time. In real conditions, the GI tract is characterized by folds and the lumen is contracting, forcing the CE to move in a non-predictable way. Generally, the dominant motion of the capsule in-*vivo* is forward, whereas in contraction periods sometimes the CE is forced to move backwards. Also, the velocity of the CE is variable inside the GI tract. Regarding the size measurements, the fact that all the landmarks to be measured, had the same size and they were placed in the same visual perspective towards the CE, was simplifying the problem. Nevertheless, the results obtained open perspectives for further experimentation in more complex environments, aiming to the ultimate goal of *in-vivo* measurements.

To this end our next step is to evaluate the proposed intelligent system using a more realistic intestinal phantom model, *e.g.*, with turns and folds of various degrees; however, there are still several challenges to overcome with respect to establishing the ground-truth within such an environment. Future work includes the investigation of alternative image registration and measurement techniques, extending current depth map estimation approaches [29],[30],[31],[32], and deep learning methods for image registration, such as siamese networks [45].

The potentials of the proposed methodology are not limited to GI endoscopy. Considering that it is entirely adaptive it can also be used for *in-vivo* measurements in the context of other endoscopic procedures, such as colposcopy [46] and laparoscopy [47].

ACKNOWLEDGMENT

We would like to thank Prof. Ervin Toth, Department of Gastroenterology, Skåne University Hospital, Malmö, Lund University, Sweden, for providing the capsule endoscope used for the purposes of our study.

REFERENCES

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, p. 417, 2000.

[2] L. J. Sliker and G. Ciuti, "Flexible and capsule endoscopy for screening,

diagnosis and treatment," *Expert Review of Medical Devices*, vol. 11, no. 6, pp. 649–666, 2014.

[3] G. Ciuti, R. Caliò, D. Camboni, L. Neri, F. Bianchi, A. Arezzo, A. Koulaouzidis, S. Schostek, D. Stoyanov, C. Oddo, and others, "Frontiers of robotic endoscopic capsules: a review," *Journal of Micro-Bio Robotics*, vol. 11, no. 1–4, pp. 1–18, 2016.

[4] A. Koulaouzidis, D. K. Iakovidis, A. Karargyris, and E. Rondonotti, "Wireless endoscopy in 2020: Will it still be a capsule?," *World Journal of Gastroenterology*,, vol. 21, no. 17, p. 5119, 2015.

[5] A. Karargyris and A. Koulaouzidis, "OdoCapsule: next-generation wireless capsule endoscopy with accurate lesion localization and video stabilization capabilities," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 352–360, 2015.

[6] M. Quirini, A. Menciassi, S. Scapellato, P. Dario, F. Rieber, C.-N. Ho, S. Schostek, and M. O. Schurr, "Feasibility proof of a legged locomotion capsule for the GI tract," *Gastrointestinal Endoscopy*, vol. 67, no. 7, pp. 1153–1158, 2008.

[7] G. Ciuti, P. Valdastri, A. Menciassi, and P. Dario, "Robotic magnetic steering and locomotion of capsule endoscope for diagnostic and surgical endoluminal procedures," *Robotica*, vol. 28, no. 2, pp. 199–207, 2010.

[8] G. Ciuti, N. Pateromichelakis, M. Sfakiotakis, P. Valdastri, A. Menciassi, D. Tsakiris, and P. Dario, "A wireless module for vibratory motor control and inertial sensing in capsule endoscopy," *Sensors and Actuators A: Physical*, vol. 186, pp. 270–276, 2012.

[9] S. P. Woods, and T. G. Constandinou, "Wireless capsule endoscope for targeted drug delivery: mechanics and design considerations," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 945–953, 2013.

[10] A. Vikram Singh and M. Sitti, "Targeted drug delivery and imaging using mobile milli/microrobots: A promising future towards theranostic pharmaceutical design," *Current Pharmaceutical Design*, vol. 22, no. 11, pp. 1418–1428, 2016.

[11] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nature Reviews Gastroenterology and Hepatology*, vol. 12, no. 3, p. 172, 2015.

[12] D. K. Iakovidis, G. Dimas, A. Karargyris, G. Ciuti, F. Bianchi, A. Koulaouzidis, and E. Toth, "Robotic validation of visual odometry for wireless capsule endoscopy," in *IEEE Int. Conf. Imaging Systems and Techniques (IST),* 2016, pp. 83–87.

[13] G. Dimas, D. K. Iakovidis, G. Ciuti, A. Karargyris, and A. Koulaouzidis, "Visual Localization of Wireless Capsule Endoscopes Aided by Artificial Neural Networks," in *IEEE Int. Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 734–738.

[14] 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," *Biomedical Signal Processing and Control*, vol. 39, pp. 486–502, 2018.

[15] G. Dimas, D. K. Iakovidis, A. Karargyris, G. Ciuti, and A. Koulaouzidis, "An artificial neural network architecture for non-parametric visual odometry in wireless capsule endoscopy," *Measurement Science and Technology*, vol. 28, no. 9, p. 094005, 2017.

[16] G. Dimas, E. Spyrou, D. K. Iakovidis, and A. Koulaouzidis, "Intelligent visual localization of wireless capsule endoscopes enhanced by color information," *Comp. Biology & Medicine*, vol. 89, pp. 429–440, 2017.

[17] D. K. Iakovidis, R. Sarmiento, J. S. Silva, A. Histace, O. Romain, A. Koulaouzidis, C. Dehollain, A. Pinna, B. Granado, and X. Dray, "Towards intelligent capsules for robust wireless endoscopic imaging of the gut," in *IEEE Int. Conf. Imaging Systems and Techniques*, 2014, pp. 95–100.

[18] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *IEEE Int. Conf. Comp. Vision,* 1999, vol. 1, pp. 666–673.

[19] C. Hu, M. Meng, P. X. Liu, and X. Wang, "Image distortion correction for wireless capsule endoscope," in *IEEE Int. Conf. on Robotics and Automation. (ICRA)*, 2004, vol. 5, pp. 4718–4723.

[20] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. on Pat. Anal. Machine Intelligence*, vol. 28, no. 8, pp. 1335–1340, 2006.

[21] D. K. Iakovidis, E. Spyrou, D. Diamantis, and I. Tsiompanidis, "Capsule endoscope localization based on visual features," in *IEEE Int. Conf. Bioinformatics and Bioengineering (BIBE),* 2013, pp. 1–4.

[22] E. Spyrou and D. K. Iakovidis, "Video-based measurements for wireless capsule endoscope tracking," *Measurement Science and Technology*, vol. 25, no. 1, p. 015002, 2013.

[23] E. Spyrou, D. K. Iakovidis, S. Niafas, and A. Koulaouzidis, "Comparative assessment of feature extraction methods for visual odometry in wireless capsule endoscopy," *Comp. Biology & Medicine*, vol. 65, pp. 297–307, 2015.

[24] G. Bao, L. Mi, Y. Geng, M. Zhou, and K. Pahlavan, "A video-based speed estimation technique for localizing the wireless capsule endoscope inside gastrointestinal tract," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual Int. Conf. of the IEEE*, 2014, pp. 5615–5618.

[25] L. Mi, G. Bao, and K. Pahlavan, "Geometric estimation of intestinal contraction for motion tracking of video capsule endoscope," in *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2014, vol. 9036, p. 90360B.

[26] M. T. Ahmed, E. E. Hemayed, and A. A. Farag, "Neurocalibration: a neural network that can tell camera calibration parameters," in *IEEE Int. Conf. on Computer Vision, 1999,* vol. 1, pp. 463–468.

[27] Q. Memon and S. Khan, "Camera calibration and three-dimensional world reconstruction of stereo-vision using neural networks," *Int. Journal of Systems Science*, vol. 32, no. 9, pp. 1155–1159, 2001.

[28] E. Besdok, "3D Vision by using calibration pattern with inertial sensor and RBF Neural Networks," *Sensors*, vol. 9, no. 6, pp. 4572–4585, 2009.

[29] S. Nadeem and A. Kaufman, "Depth Reconstruction and Computer-Aided Polyp Detection in Optical Colonoscopy Video Frames," *arXiv preprint arXiv:1609.01329*, 2016.

[30] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, 2016, pp. 740–756.

[31] F. Mahmood and N. J. Durr, "Deep Learning and Conditional Random Fields-based Depth Estimation and Topographical Reconstruction from Conventional Endoscopy," *arXiv preprint arXiv:1710.11216*, 2017.

[32] F. Mahmood and N. J. Durr, "Deep learning-based depth estimation from a synthetic endoscopy image training set," in *Medical Imaging 2018: Image Processing*, 2018, vol. 10574, p. 1057421.

[33] N. Vakil, W. Smith, K. Bourgeois, E. C. Everbach, and K. Knyrim, "Endoscopic measurement of lesion size: improved accuracy with image processing," *Gastrointestinal Endoscopy*, vol. 40, no. 2, pp. 178–183, 1994.

[34] M. Zhou, G. Bao, Y. Geng, B. Alkandari, and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy," in *Biomedical Engineering and Informatics (BMEI), 2014 7th Int. Conf. on*, 2014, pp. 237–241.

[35] H. Park, J. Y. Ahn, H. Seo, G. Y. Pih, H. K. Na, J. H. Lee, K. W. Jung, D. H. Kim, K. D. Choi, H. J. Song, and others, "Validation of a novel endoscopic program for measuring the size of gastrointestinal lesions," *Surgical Endoscopy*, vol. 31, no. 11, pp. 4824–4830, 2017.

[36] O. Goldstein, O. Segol, P. D. Siersema, H. Jacob, and S. A. Gross, "Novel device for measuring polyp size: an ex vivo animal study," *Gut*, 2017, doi: 10.1136/gutjnl-2017-314829.

[37] M. Visentini-Scarzanella, T. Hanayama, R. Masutani, S. Yoshida, Y. Kominami, Y. Sanomura, S. Tanaka, R. Furukawa, and H. Kawasaki, "Tissue shape acquisition with a hybrid structured light and photometric stereo endoscopic system," in *Int. Workshop on Computer-Assisted and Robotic Endoscopy*, 2015, pp. 46–58.

[38] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Deepmatching: Hierarchical deformable dense matching," *Int. Journal of Computer Vision*, vol. 120, no. 3, pp. 300–323, 2016.

[39] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *JOSA A*, vol. 7, no. 5, pp. 923–932, 1990.

[40] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, no. 5, pp. 551–560, 1990.

[41] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. Computer Vision and Pattern Recognition, 1997*, pp. 1106–1112.

[42] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?," in *European Conference on Computer Vision*, 2016, pp. 809–824.

[43] J. Bouguet, "Camera calibration toolbox for Matlab," Caltech, Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[44] R. E. Schoen, L. D. Gerber, and C. Margulies, "The pathologic measurement of polyp size is preferable to the endoscopic estimate," *Gastrointestinal Endoscopy*, vol. 46, no. 6, pp. 492–496, 1997.

[45] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 10–18.

[46] B. L. Craine, E. R. Craine, C. J. O'Toole, and Q. Ji, "Digital imaging colposcopy: corrected area measurements using shape-from-shading," *IEEE Trans. on Medical Imaging*, vol. 17, no. 6, pp. 1003–1010, 1998.

[47] R. McKinlay, M. Shaw, and A. Park, "A technique for real-time digital measurements in laparoscopic surgery," *Surgical Endoscopy And Other Interventional Techniques*, vol. 18, no. 4, pp. 709–712, 2004.