

Video-based measurements for wireless capsule endoscope tracking

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2014 Meas. Sci. Technol. 25 015002

(<http://iopscience.iop.org/0957-0233/25/1/015002>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 130.215.16.251

This content was downloaded on 04/04/2014 at 18:54

Please note that [terms and conditions apply](#).

Video-based measurements for wireless capsule endoscope tracking

Evaggelos Spyrou and Dimitris K Iakovidis

Department of Informatics and Computer Technology, Technological Educational Institute of Lamia, Lamia, Greece

E-mail: vspyrou@teilam.gr and iakovidis@ctr.teilam.gr

Received 3 April 2013, in final form 16 September 2013

Published 26 November 2013

Abstract

The wireless capsule endoscope is a swallowable medical device equipped with a miniature camera enabling the visual examination of the gastrointestinal (GI) tract. It wirelessly transmits thousands of images to an external video recording system, while its location and orientation are being tracked approximately by external sensor arrays. In this paper we investigate a video-based approach to tracking the capsule endoscope without requiring any external equipment. The proposed method involves extraction of speeded up robust features from video frames, registration of consecutive frames based on the random sample consensus algorithm, and estimation of the displacement and rotation of interest points within these frames. The results obtained by the application of this method on wireless capsule endoscopy videos indicate its effectiveness and improved performance over the state of the art. The findings of this research pave the way for a cost-effective localization and travel distance measurement of capsule endoscopes in the GI tract, which could contribute in the planning of more accurate surgical interventions.

Keywords: wireless capsule endoscopy, video, motion estimation, orientation, distance

(Some figures may appear in colour only in the online journal)

1. Introduction

Wireless capsule endoscopy (WCE) [1] has become a standard imaging technique for screening the gastrointestinal (GI) tract, especially the small intestine mucosa, which is not easily reached by the usual upper endoscopy and colonoscopy. It is performed by a swallowable capsule the size of a large vitamin pill that includes a miniature color video camera, a light, a battery and a video stream transmitter. The capsule is propelled by peristalsis through the gut and reaches the right colon typically in 5–8 h. During its journey in the human body it captures thousands of images and transmits them wirelessly to a video recorder. Energy consumption issues impose limitations in the image quality and the frame rate; however, wireless energy transmission has been proposed to cope with this issue [14]. The tracking of the capsule endoscope within the body is important for the localization of abnormal findings and planning of surgical intervention. Typically, it is performed by radio frequency (RF) sensor arrays, mounted externally to the human body, receiving the signals transmitted by the capsule endoscope. The average

position error reported for this technique is 37.7 mm, with a maximum error reaching 114 mm [2]. However, this technique has been criticized for its inability to definitively localize small bowel lesions, mainly because it is prone to inaccuracy due to differences in small bowel transit time or variant anatomy [15]. Anatomic landmarks such as the stomach and cecum are also used for lesion localization; however, this technique is useful mainly for only the proximal and distal small bowel.

In order to improve capsule tracking, techniques based on magnetic fields have been proposed as alternatives to the RF approach. For example, a significantly lower average position error of 3.3 mm and average orientation error of 3°, have been reported with the use of magnetic sensor arrays [2]. However, the overall accuracy of the magnetic tracking depends highly on the number of external sensors used [2], rotation information can be incomplete [3] and, as the RF tracking approach, it requires external equipment that can be discomforting for the patient. The use of magnetic sensors on board the capsule for tracking purposes has also been proposed in the context of robotic-aided active diagnostic examination, and the position errors reported include 14 mm

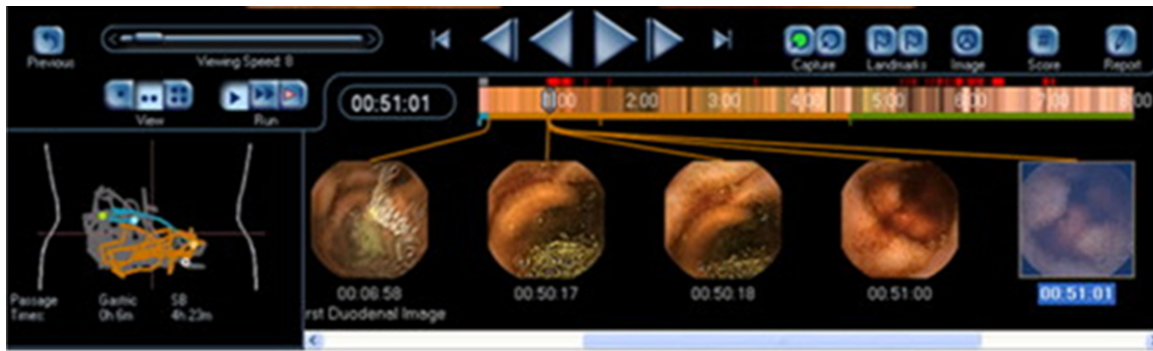


Figure 1. Control panel of the GUI of the Given Imaging Rapid WCE software. The location of the capsule is graphically indicated in the visualization tool in the bottom left corner of the GUI. Similar visualization tools are provided by software applications of other WCE companies [11].

along the X-axis, 11 mm along the Y-axis and 19 mm along the Z-axis [16]. State of the art video-based capsule localization methods, which could be used for capsule tracking, are based on the topographic segmentation of the WCE video into regions corresponding to different organs; therefore, they can be used only to infer the organ in which the capsule is located and not the position of the capsule within the organ [21], or to estimate the corresponding transit times [22]. Although several capsule tracking methods have been proposed, as reviewed in [21], none of them could offer a complete solution to address the challenging capsule localization problem. To date, a commercial WCE can only provide rough 2D position information, which is qualitatively represented by a 2D visualization in the graphical user interface (GUI) of the WCE software application (figure 1).

In this paper, we propose a method for the estimation of the displacement and the orientation of the capsule for its tracking, based only on information extracted from WCE video frames. The proposed method proceeds in three steps: (a) WCE image analysis for extraction of speeded up robust features (SURF) from video frames, (b) registration of consecutive WCE video frames based on the random sample consensus (RANSAC) algorithm, (c) estimation of the relative scaling and rotation of the frame contents, frame by frame as long as the same contents are visible in the consecutive frames; then, from the estimated scaling and rotation, the displacement and the rotation of the capsule endoscope in the GI tract can be derived. Both SURF and RANSAC approaches have been previously applied in the context of content-based image retrieval [4], whereas recently we applied them in the context of WCE for orientation estimation of the capsule [17]. Experiments indicated that the SURF–RANSAC approach is more accurate than the state of the art method proposed in [3], which is based on Kanade–Lucas–Tomasi optical flow computation (KLT). However, that study was only preliminary presenting experiments on a rather limited dataset. The promising results obtained, motivated us to significantly extend that approach by incorporating not only rotation but also scaling estimation along with a camera motion estimation model. This model enables the estimation of the displacement of the capsule endoscope, either forward or backward, within the GI tract, making feasible the estimation of the travel distance of the capsule as well. To the best of our knowledge this is the

first video-based approach for both WCE displacement and orientation estimation. Such an approach does not require any external equipment, and it could be considered as a supplementary approach to improve the accuracy of the sensor-based localization techniques. For example, a hybrid magnetic and vision technique, incorporating the orientation estimation method proposed in [3], has been utilized for 3D recovery of pathological tissues [18]. In the context of WCE, other applications of camera motion estimation models include WCE video summarization [19], 2D to cylindrical inverse projection of the WCE images [20], and a model of deformable rings (MDR) enabling a 2D map representation of the internal surface of the GI tract for WCE content interpretation based on intestinal motility patterns [28]. Considering the advancements in the field of WCE, this paper presents a novel application of a camera motion estimation model for *in vivo* measurements enabling capsule endoscope tracking. Its main contributions are: (a) the extension of the method we proposed in [17] for estimation of both the rotation and the displacement of the capsule endoscope, and (b) its experimental evaluation on a significantly larger, publicly available, database to investigate the feasibility of using such a method for capsule endoscope tracking and measurement of its travel distance.

The rest of this paper consists of five sections. Section 2 describes the image analysis process applied to each WCE video frame, section 3 describes the algorithm applied for the registration of consecutive video frames, and section 4 describes the motion estimation model considered. The experiments and the results obtained by the application of the proposed method on WCE videos are presented in section 5. The last section includes a discussion, summary of conclusions that can be derived from this study and realistic perspectives for future research.

2. WCE image analysis

The extraction of visual features, either global (from the whole image) or local (from image patches), has shown serious limitations in many vision related problems. Global features may change notably, due to e.g. small changes of viewpoint, zooming, changes of illumination, contrast, etc. While the same stands for local features, an additional problem lays in

the patch selection algorithms. Typically patches are extracted using segmentation or clustering techniques which appear to be very sensitive to image distortions such as the aforementioned. Thus, patches extracted from visually similar (to the human observer) images may be significantly different in terms of e.g. shape and size.

To overcome such issues, in this work, we adopt an approach based on a set of interest points. These points are locally extracted from images. More specifically, from each video frame we extract a set of interest (salient) points. Then, visual features are extracted from the surrounding area of each. For obvious reasons, given the problem under investigation, this set of points should be robust to affine image transformations, since they are typical to WCE video sequences.

To this end, we choose to extract the well-known and adopted in many problems in the field of computer vision, namely the SURF features [5], and use them in order to both capture and represent the visual properties of video frames. SURF have been very popular during the last few years in tasks such as pattern recognition, image registration, semantic image analysis, etc, as they have been proven to achieve high repeatability and distinctiveness. Moreover, their calculation is significantly faster as compared to other popular features of this kind such as scale invariant feature transform (SIFT) features [6] from which they have been inspired.

The SURF interest point extraction algorithm proceeds in two distinct steps. The first step detects local interest points, whereas the second step extracts a descriptor from the area surrounding each of them. In order to detect local interest points, the SURF algorithm adopts a fast approximation of the Hessian matrix that exploits integral images. The extracted interest points are then the local maxima of the Hessian matrix determinant. This blob response maxima process is carried out on several octaves of a Gaussian scale-space. Moreover, the correct scale of each point is automatically selected also using the Hessian determinant, as introduced in [7]. For exact point localization, an efficient non-maximum suppression algorithm is used at a $3 \times 3 \times 3$ intra-scale neighborhood [8].

The SURF descriptor captures the intensity content distribution around the points detected with the aforementioned process. The first-order Haar wavelet responses are computed with the use of integral images, resulting in a 64-dimensional feature vector. In order to achieve the desired property of rotation invariance, a dominant orientation is determined. This dominant orientation is the direction that maximizes the sum of the Haar wavelet responses in a sliding window of size $\pi/3$ around the neighborhood of each interest point.

For the computation of each descriptor, a square area with side equal to $20 \times s$ is used, around the corresponding interest point. This area is then divided into 4×4 blocks, with s denoting the interest point scale. This allows the descriptor to be scale invariant. For each one of the 16 blocks, four values are extracted, corresponding to the sum of the x , y , $|x|$ and $|y|$ first-order Haar wavelet responses in a 5×5 grid in the block. To make the descriptor robust to contrast changes, the computed descriptor vector is finally turned into a unit vector.

It is clear that the selection of the aforementioned low-level feature extraction scheme combines speed with

robustness to scale, rotation and contrast changes. The extraction of the SURF features is rather fast, thus the necessary amount of time is small, when compared to other similar approaches. This makes them effective for the analysis of WCE video sequences, even if speed is not a priority in the presented application. Moreover, their robustness to image variations ensures that the transformation between two images may always be calculated under certain visual changes, as in the case of the problem under investigation, which is the main reason we use them in our approach.

An example of the interest points from which the SURF features are extracted is illustrated in figure 2, where we use a typical WCE video frame. In this figure, scale is denoted by the yellow circle, while orientation, is denoted by the green line. The position of each of the interest points is the center of the yellow circle, i.e. the starting point of the green line. Figure 3 illustrates the video frame of figure 2, along with its consecutive frame in the video sequence, and the interest points determined by the SURF algorithm (scales and orientations have been omitted for clarity of presentation). We should note that for each point we only use its position and its description. Scale and orientation are discarded.

3. WCE video frames registration

After selecting a set of invariable interest points in order to capture the visual properties of a given video frame, the next step is to estimate a geometric transform between two consecutive frames. We expect that consecutive frames should typically contain similar visual features, as the capsule's camera viewpoint changes slowly. Such changes may often cause similar features to appear in a totally different spatial layout. However, one may observe easily, e.g. in figure 3, that many of the interest points detected should follow the same transformation. If this transformation is estimated, then we should be able to calculate parameters such as the scaling and the rotation of the WCE image content between two consecutive frames.

In order to detect similar features between consecutive video frames we have considered the application of an image registration approach between these frames, i.e. to find correspondences between them. To achieve this we adopted the well-known RANSAC algorithm [9]. Using RANSAC, we are able to determine the geometric transformation between two consecutive video frames given a set of tentative point-to-point correspondences. In general, these correspondences are not available, thus we need to calculate them each time. In this case, we determine a set of such correspondences based on the SURF descriptors, around the interest points, i.e. two points extracted from two consecutive video frames form a tentative correspondence if the distance of their visual descriptors falls below a given (user defined) threshold T_C . Thus, we calculate all the distances between all descriptor pairs, we filter them and obtain the set of tentative correspondences. In general, this threshold is not determined automatically. More specifically, we manually set its value, according to how strict we wish to be when we form a tentative correspondence between two interest points. Given a transformation between two video frames, we

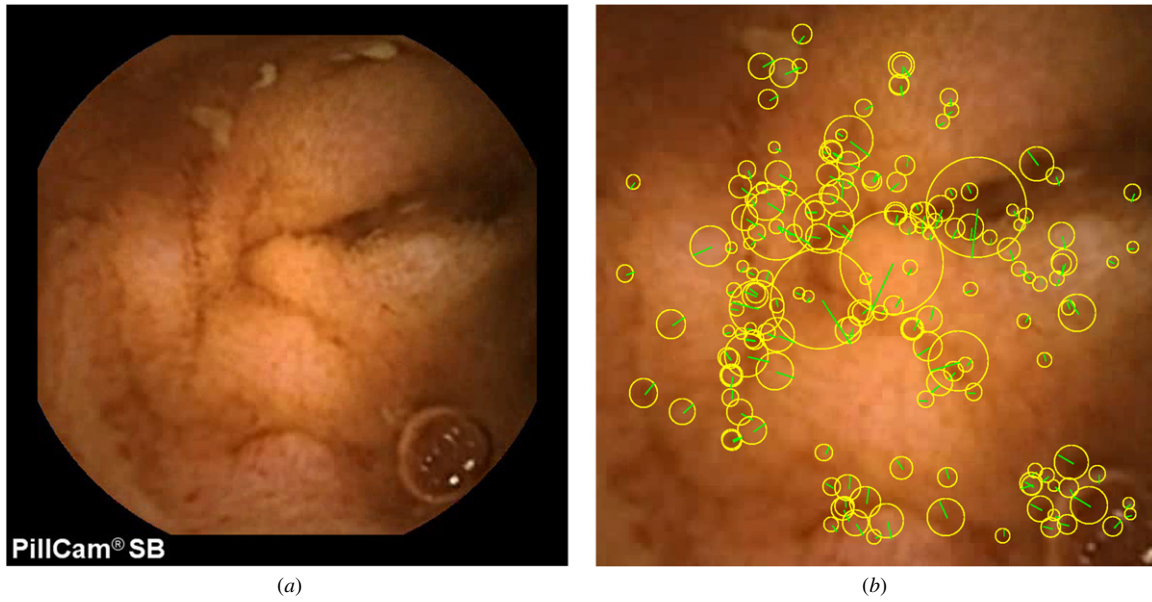


Figure 2. A typical WCE video frame. (a) Original video frame, as captured by the PillCam capsule. (b) Processed video frame, with interest points extracted from the frame illustrated in (a), using the SURF algorithm. The size of the yellow circle and the green line denote the scale and the dominant orientation, respectively.

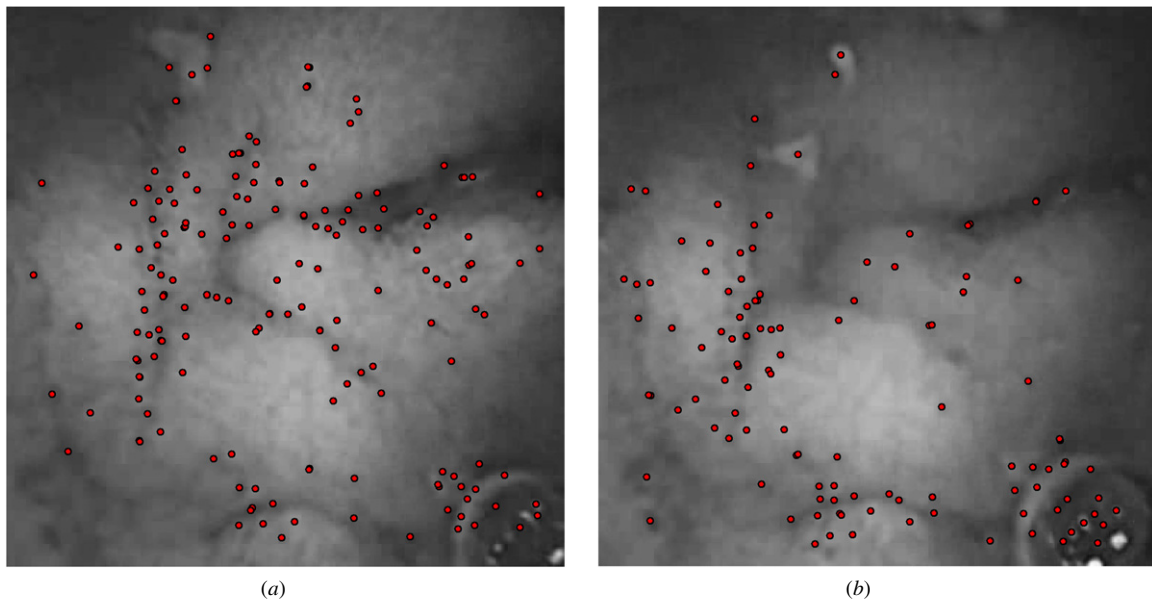


Figure 3. Extracted interest points from two consecutive frames, denoted by red dots. The first frame is depicted in color in figure 1; however, frames are presented in grayscale herein for illustrative purposes.

should expect that a subset of the tentative correspondences should support this model. We should emphasize that within this process, the position of interest points is not taken into account. Thus, a tentative correspondence pair can sometimes consist of points from significantly different parts of the corresponding consecutive video frames. Moreover, each point may belong to more than one tentative correspondence pair. When a transformation is estimated, based on a subset of these correspondences, this subset is denoted as the set of *inliers* for this transformation, while the rest, i.e. the false correspondences, denote the set of *outliers*.

The proposed method exploits the main advantage of RANSAC, which is its ability to determine a model in the

presence of many outliers, as in most practical cases. More specifically, RANSAC is used to estimate the transformation that maximizes the number of inliers, with a certain probability P_R , which is user defined. It is obvious that RANSAC relies a lot on the tentative correspondences among points, which will be provided at its initial step, and also the accuracy of the calculated model depends on the chosen probability. However, most research works choose a significantly high probability (e.g. $P_R \approx 99.8\%$), which practically leads to as accurate as possible results.

Figure 4(a) illustrates an example of the correspondences detected among the initially extracted interest points using SURF features for the two consecutive WCE video frames

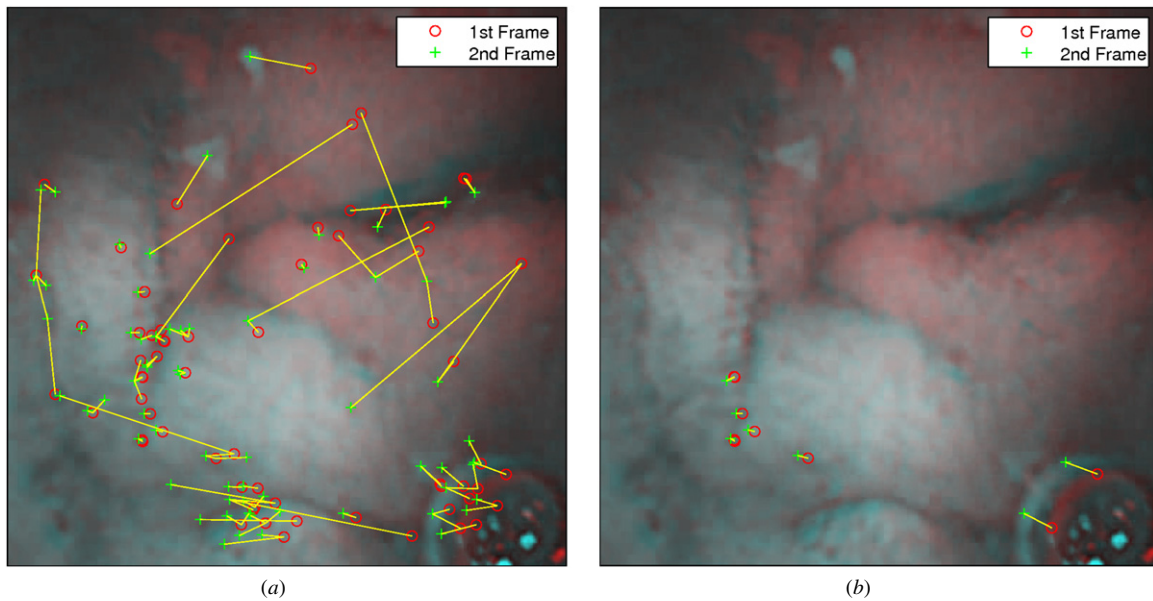


Figure 4. Correspondences between the overlaid consecutive video frames of 2. (a) Correspondences between all interest points of figure 2(a). (b) Inliers estimated using RANSAC.

of figure 3. For presentation purposes, the frames have been overlaid and pixel differences have been denoted by color composites. In other words, after subtracting corresponding pixel values of the second frame from those of the first one, a positive result leads to a red-tinted depiction on the overlay image, while a negative result leads to a green-tinted depiction.

In figure 4(a), one may observe the large number of tentative correspondences. The majority of these correspondences may be easily characterized by a human observer as outliers, since it is obvious that they do not obviously follow any simple model. In contrast, determining inliers which leads to the selection of a valid model for the geometric transformation, e.g. as in figure 4(b), is not trivial to the human observer. Even if one may recognize a few pairs following approximately the same transformation, it still may not be obvious whether they are part of the largest possible set of inliers.

In the proposed method, we used RANSAC for the estimation of an affine transform that maps any given point of interest x_i of a given video frame to a corresponding point x'_i of the consecutive frame. Given the set of correspondences of points of interest between two consecutive frames, i.e. the pairs $x_i \leftrightarrow x'_i$, we can define a transformation T as

$$x'_i = Tx_i. \quad (1)$$

In this case, the mathematical model estimated by RANSAC is a transformation matrix. As previously mentioned, RANSAC will select the model supported by the largest possible set of inliers (in fact, the largest possible, after a user defined number of iterations). Since an affine transformation requires only three points in order to be estimated, it is intuitive that a significantly larger set increases the probability of its correctness.

The estimation of an image transformation using RANSAC has been applied very effectively in tasks such as stereoscopic camera calibration [10], a case where the

images captured by two cameras differ only by means of a perspective transform. This simple idea is extended herein, but instead of two images taken by a stereoscopic camera, we consider the case of the camera moving slowly and capturing approximately the same 'scene' by a different viewpoint. We should expect that since the variation of the viewpoint is not expected to be very high, we should encounter a small number of false correspondences. However, considering the nature of the problem at hand, many false correspondences may be introduced due to the equipment, e.g. noise, or the compression, e.g. MPEG artifacts. This justifies our choice, as we adopt RANSAC, due to its ability to correctly estimate a model even in the presence of a large number of outliers (false correspondences) present.

In the following we will present in brief all steps performed by RANSAC, towards the estimation of the correct model. RANSAC begins with a random sample which should contain a number of points equal to the number of model variables. Given a set of point correspondences between two consecutive video frames, we apply RANSAC in order to estimate the affine transformation between them as follows:

- (1) Initially, three points are randomly selected. Under the assumption that all are inliers, they define an affine transformation.
- (2) Using the model estimated under the assumption of the previous step, the number of inliers, i.e. correspondences that fit this model, is calculated. Their distance to the points estimated by the affine transformation should be smaller than the predefined threshold.
- (3) Random triplets of points are sampled. After every such iteration, both the maximum number of inliers that satisfy the model up to that point is kept, along with the corresponding model.
- (4) This process is repeated until a predefined number of iterations is reached, or the probability that new inliers

may be determined using another model falls beneath a certain threshold.

- (5) When the aforementioned relation is satisfied, the selected model is redefined, using all inliers selected from the previous steps.

Figure 4(b) illustrates an example of the inliers selected using RANSAC from the tentative correspondences of the two consecutive WCE video frames of figure 3, depicted in figure 4(a).

4. Capsule motion estimation

Once the transformation matrix T is calculated (equation (1)), we are able to extract rotation θ and scale s , from which the rotation and the displacement of the capsule in the GI tract can be derived. A forward (positive) displacement is measured for $s > 1 + s_0$, a backward (negative) displacement is measured for $s < 1 - s_0$, and for $1 - s_0 \leq s \leq 1 + s_0$ no displacement is measured. Constant $s_0 < 1$ controls the sensitivity of the motion estimation algorithm. First, we can write

$$T = \begin{pmatrix} s_c & s_s & t_x \\ -s_s & s_c & t_y \\ 0 & 0 & 1 \end{pmatrix} \quad (2)$$

where $s_c = s \cos \theta$, $s_s = s \sin \theta$. We should note that t_x and t_y denote the translations on x - and y -axis, respectively.

We are then able to make an estimation s_e of scale s as

$$s_e = \sqrt{s_c^2 + s_s^2} \quad (3)$$

and an estimation θ_e of the rotation θ as¹

$$\theta_e = \tan^{-1}(s_c/s_s). \quad (4)$$

In order to select a motion model, one should specifically take into account the characteristics of the problem at hand. Very often, certain assumptions should be made, since the goal is to select a simplified model, able to effectively fit the problem. While WCE travels in the GI tract, it is propelled by gastrointestinal tract GI [19]. This leads to small perspective changes. However, considering only the WCE camera motion and ignoring perspective parameters and the non-rigid deformation of the GI tract, as in [19], then two successive video frames may be related using a projective transformation [10]. We should also make the further assumption that the WCE camera we used approximates the pinhole model. Under these assumptions, our WCE camera performs a linear transformation from the 3D to the 2D projective space.

Let (X_i, Y_i, Z_i) denote the coordinates of a given point A_i in the 3D space. Let $A'_i(x_i, y_i)$ be the projection of A_i in the 2D space under a projective transformation. Using basic projective geometry, we are able to write

$$x_i = f \cdot \frac{X_i}{Z_i}, \quad (5)$$

where f denotes the focal length of the camera. Having estimated the scale, as we have previously described, then we are able to calculate Z_i as

$$Z_i = f \cdot \frac{X_i}{x_i}. \quad (6)$$

Given a pair of two corresponding points A_1 and A_2 , i.e. a pair of inliers, and since $X_2 = s X_1$, from (5) and (6) we estimate the (relative) displacement of the capsule endoscope as

$$\frac{Z_2}{Z_1} = s \cdot \frac{x_1}{x_2}, \quad (7)$$

where x_1/x_2 is measurable directly from images (i.e. the corresponding pixel coordinates), s is estimated from equation (3). Considering that several point correspondences are identified between two consecutive frames using RANSAC, the overall estimation of the displacement of the capsule is based on a set of estimated displacements. These displacements are aggregated by averaging to obtain a single, noise-tolerant displacement estimation. In order to estimate the displacement of the capsule endoscope in terms of a unit of measurement, e.g. in mm, several methods have been proposed [10, 26, 27]. These methods require specification of several capsule parameters, including optical and digital sensor parameters. In this paper, we address the problem of capsule displacement estimation regardless of the capsule used; therefore, any capsule-specific measurement is considered beyond the scope of this work.

5. Experiments and results

The proposed method was evaluated on a set of 117 WCE video clips with a resolution of 576×576 pixels and duration of approx. 20 s each. The video clips have been acquired from the small bowel with Given Imaging PillCam capsules, and they are available through the Given Imaging Atlas database [29]. This database includes video clips acquired also from the esophagus and the colon, but they have been excluded from this study because they were insufficient (5 and 13, respectively) to obtain statistically significant results for the respective parts of the GI tract.

Since to the best of our knowledge there does not exist any ground truth data set for the problems at hand, the only feasible way to evaluate the proposed approach was to use simulated scale and rotation transformations of sampled video frames². We adopted the same evaluation approach as in [3] and considering that the displacement of the capsule, estimated by the motion estimation model described in section 4, is tied to image scaling, we extended it for capsule's displacement estimation. More specifically, for a given video frame, we constructed a rotated and/or scaled version of it. Then, we extracted SURF features from both and, using RANSAC, we determined inlier correspondences between the original and rotated video frames. Based upon these correspondences we estimated a transformation matrix, from which it was trivial to extract rotation and/or scale accordingly. Finally, we estimated orientation (rotation) and/or displacement (scaling) errors in

¹ In our MATLAB implementation and in order to deal with all quadrants, the well-known atan2 function was used.

² The dataset used in this study can be constructed using the tools provided in: <http://innovation.teilam.gr/mst/dataset.htm>.

Table 1. Estimated orientation error per rotation angle, at the original scale.

Rotation angle actual (degrees)	Mean absolute orientation error proposed method (degrees)	Standard deviation of orientation error proposed method (degrees)	Mean absolute orientation error method in [3] (degrees)	Standard deviation of orientation error method in [3] (degrees)
5	0.022	0.017	0.072	0.024
10	0.022	0.017	0.139	0.033
15	0.026	0.020	0.502	0.051
20	0.028	0.026	1.702	2.265
25	0.689	0.058	0.787	2.274
30	1.558	2.105	23.742	41.356
35	4.719	1.913	35.789	54.624
40	20.084	43.190	40.621	58.885
45	30.771	51.593	48.253	61.083

comparison to the known transformation. The results obtained are organized in the following three respective subsections.

In order to obtain a representative set of frames from each small bowel video clip we have sampled them every 1 s, resulting in approx. 20 frames per video clip. Since this study makes use of independent video frames, which are artificially rotated and scaled, a sampling rate of 1 s is sufficient to obtain a set of representative video frames from the video clips. A higher frame rate would increase the redundancy of information and the computational cost of the experimentation. However, it should be noted that in case of application of the proposed methodology on non-simulated video frame sequences, a sampling rate higher than 1 frame per second (fps) would improve the correspondence between the SURF features, since the frame rate of contemporary capsules ranges between 2 and 4 fps. For the extraction of the SURF features we have used the OpenSurf [12] MATLAB implementation. We have chosen to extract features using five octaves. Experiments with more octaves showed no further improvement in the results of neither rotation nor displacement estimation, while less octaves led to a degradation of the results. The average extraction time was about 1.5 s per image, at the original scale.

In order to estimate the transformation between two images, we used the built-in RANSAC implementation of MATLAB [13]. For the determination of the set of the tentative correspondences, we have set $T_C = 5$. Since we desire the best possible accuracy, we have set $P_R = 99.8\%$. This increased estimation time, but certainly, lead to the most accurate results possible, given the extracted SURF features. The average time needed by RANSAC to estimate the transformation matrix between two given video frames is about 0.5 s per frame pair, at the original scale. Constant s_0 was empirically set at $s_0 = 0.2$.

5.1. Orientation

A total of eight different rotation angles were tested from 5° to 45° with a rotation step of 5° . The results obtained are given in table 1. This table shows that the error between the actual and the estimated rotation angle is rather low up to 35° and increases for angles over 40° . A further increase of the rotation angle causes larger errors, thus we do not present further experiments over 45° . Table 1 and the diagram in figure 5 compare the performance of the proposed method with

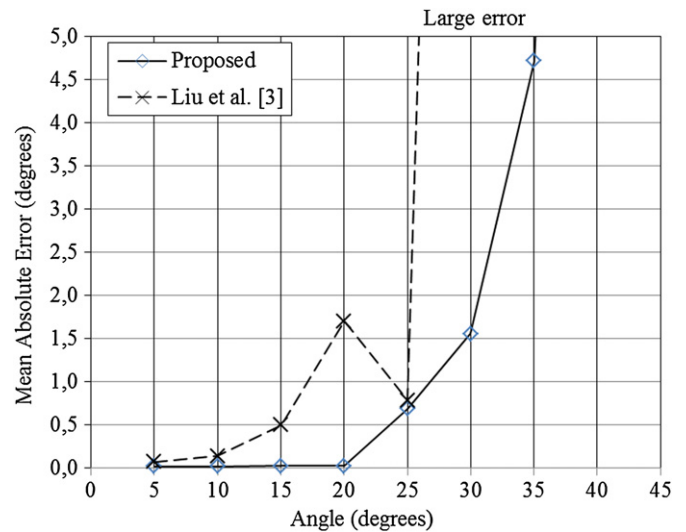


Figure 5. Comparison between the proposed method and the method proposed in [3], in terms of the mean absolute error observed in rotation angle estimation.

the state of the art, KLT-based method [3] in terms of mean absolute error. It is evident that the proposed method is better in the sense that it produces lower error rates for angles larger than 5° . More importantly, the KLT-based method, results quickly in a very large error as the angle increases from 25° to 30° , whereas the error obtained by the proposed method becomes very large for angles over 35° .

It should be noted that using these methods for rotation estimation in practice, i.e. with non-simulated scaling and rotation, the rotation estimation errors, as reported in table 1, will accumulate over a range of frames. However, this accumulation error cannot be properly evaluated with the available dataset, since as we have already mentioned, it is not annotated with actual measurements. Given the results of table 1, only a coarse estimation of the accumulation error is possible. For example, considering a video clip of 20 s from our dataset, where the capsule endoscope rotates up to 20° , the accumulated error will be of the order of only 1° , whereas it can increase significantly for rotation angles larger than 25° or 30° . However, the likelihood of such rotation angles decreases as the frame rate of the capsule endoscopes increases, considering that a higher frame rate will enable capturing capsule's rotation more progressively.

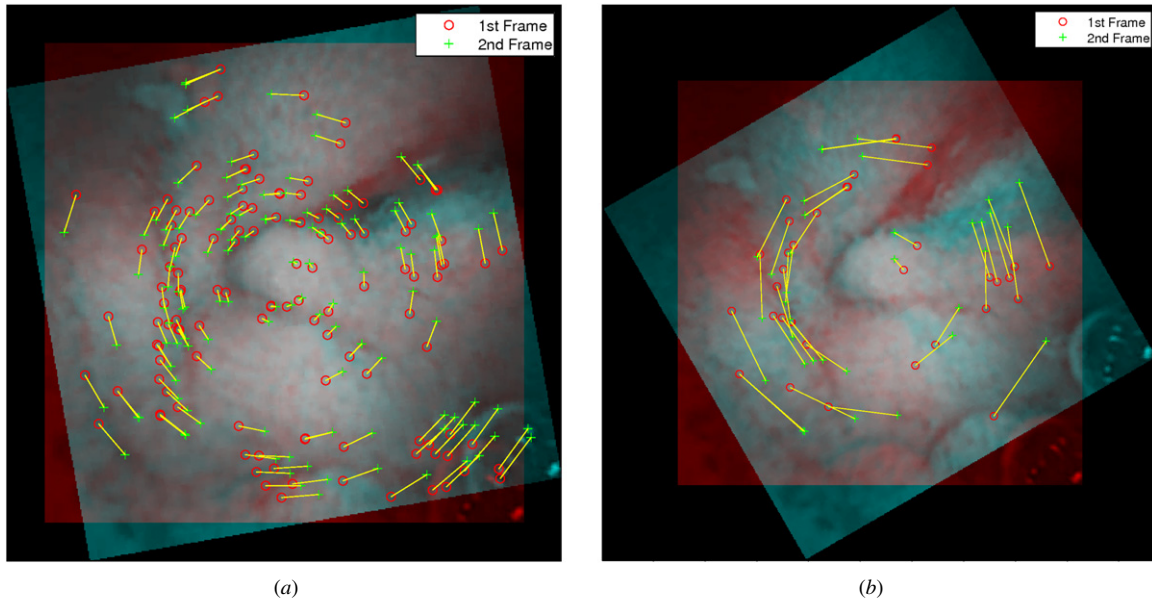


Figure 6. Inliers estimated using RANSAC, for a video frame and (a) its 10° rotated instance, (b) its 30° rotated instance. Scale was not modified in this example.

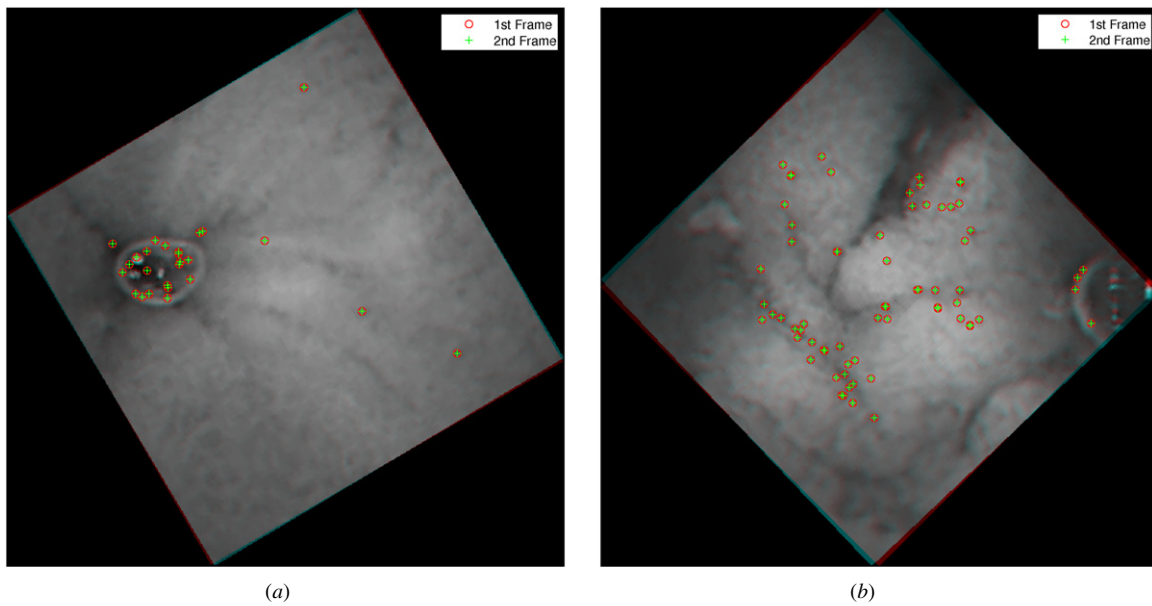


Figure 7. Two WCE frames artificially rotated by (a) 30° and (b) 45° . The rotated frames are denoted as first frame in both subfigures. Using the proposed method, the rotation estimation was 30.76° and 43.42° , respectively. The original frame, rotated accordingly, is denoted as second frame and was overlaid on the first frame. In both cases, one should easily observe that corresponding points match almost perfectly.

Representative examples of the correspondences detected between rotated video frames by the proposed method are illustrated in figure 6. More specifically, figure 6(a) illustrates the correspondences detected for a small rotation angle, of 10° , and figure 6(b) illustrates the correspondences detected for a larger rotation angle, of 30° . In both cases the estimated orientation using the proposed method is accurate, as indicated in table 1.

In order to demonstrate how accurately the feature points in the second frame are detected, figure 7 provides two examples of two different video frames rotated artificially

by 30° and 45° . We have overlaid the same frames rotated artificially by the angle estimated by the proposed method, which is 30.76° and 43.42° , respectively. One can easily notice that it is rather difficult to distinguish any deviation between the corresponding points of the two frames. This is justified since RANSAC selects a set of corresponding points, which is a subset of all interest points, and points that belong to this set follow exactly the same transformation. The example of figure 7(b) is one of many cases where the proposed method results in a small error for such a large rotation angle.

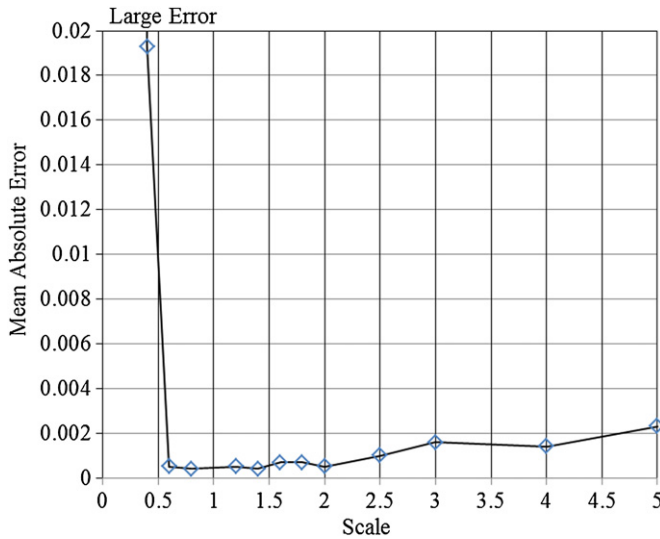


Figure 8. Estimated displacement error per scale, at the original rotation angle.

5.2. Displacement

A total of 14 different scales were tested, varying from a factor of 0.2 to a factor of 5.0. Lower or higher scales are less likely to appear in frame sequences of WCE videos. This assumption was confirmed by preliminary experimentation, using the SURF–RANSAC scheme as described in section 3, to estimate scaling in consecutive frames of the raw video clips of our dataset. The mean absolute scaling factor measured was 2.144, with the scaling factor to range between 0.3 and 4.8 in 87% of the consecutive frame pairs. Scaling factors ranging from 0 to 0.3 were rare, i.e. were encountered in 3% of the consecutive frame pairs. The results obtained by the application of the proposed method for displacement estimation of the capsule endoscope are given in table 2 and illustrated in figure 8. This table shows that the error between the actual and the estimated

Table 2. Estimated displacement error per scale, at the original rotation angle.

Actual scale	Mean absolute displacement error	Standard deviation of displacement error
0.2	N/A	N/A
0.3	13.3085	22.3591
0.4	0.0193	0.0895
0.6	0.0005	0.0005
0.8	0.0004	0.0003
1.2	0.0005	0.0004
1.4	0.0004	0.0004
1.6	0.0007	0.0007
1.8	0.0007	0.0007
2.0	0.0005	0.0004
2.5	0.0010	0.0008
3.0	0.0016	0.0016
4.0	0.0014	0.0010
5.0	0.0023	0.0017

scale is practically zero from various scales between 0.4 and 5.0, while it increases significantly when we further downscale below 0.4. We should emphasize, given the results of table 2, that the proposed method may be successfully applied on almost all video frames from scales of 0.4–5.0, since it results in small errors.

Figure 9 illustrates representative examples of the correspondences detected between scaled video frames, using the proposed method. More specifically, figure 9(a) illustrates the correspondences detected for scaling by a factor of 50% (downscaling) and figure 9(b) illustrates the correspondences detected for scaling by a factor of 150% (upscaling). In both cases the estimated scaling using the proposed method is accurate, as indicated in table 2. It should be noted that the error for downscaling by a factor of 0.3 or less is significantly large. Further downscaling results in even larger errors, making the proposed method not applicable (N/A), i.e. RANSAC cannot find enough correspondences between consecutive frames to

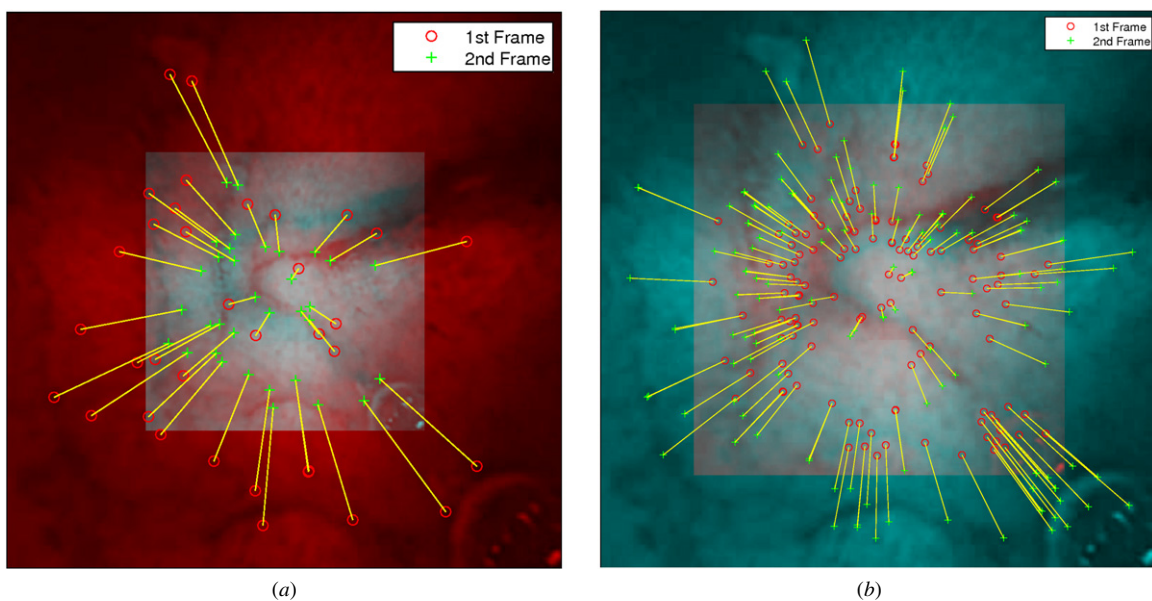


Figure 9. Inliers estimated using RANSAC, for a video frame and (a) scaled by a factor of 0.5, (b) scaled by a factor of 1.5. Frames have not been rotated in this case.

Table 3. Estimated orientation and displacement errors per angle, at a scale of 2.0.

Actual rotation angle (degrees)	Mean absolute orientation error (degrees)	Standard deviation of orientation error (degrees)	Mean absolute displacement error	Standard deviation of displacement error
5	0.0007	0.0006	0.0207	0.0172
10	0.0008	0.0007	0.0241	0.0181
15	0.0009	0.0007	0.0236	0.0194
20	0.0011	0.0010	0.0311	0.0274
25	0.0016	0.0020	0.0391	0.0348
30	0.0018	0.0022	0.0534	0.0916
35	0.0021	0.3019	0.1272	0.6991
40	1.0044	2.6429	34.8248	57.0366

Table 4. Estimated orientation and displacement errors per scale, at an angle of 30°.

Actual scale	Mean absolute orientation error (degrees)	Standard deviation of orientation error (degrees)	Mean absolute displacement error	Standard deviation of displacement error
0.4	18.4441	49.9357	0.1222	0.1009
0.6	0.3800	0.2338	0.0428	0.0212
0.8	0.0803	0.7231	0.0010	0.0182
1.2	0.0686	0.6634	0.0011	0.0276
1.4	0.0474	0.4321	0.0011	0.0342
1.6	0.2023	1.2120	0.0037	0.0523
1.8	0.0550	0.4211	0.0405	0.0620
2.0	0.0451	0.4298	0.0018	0.0222

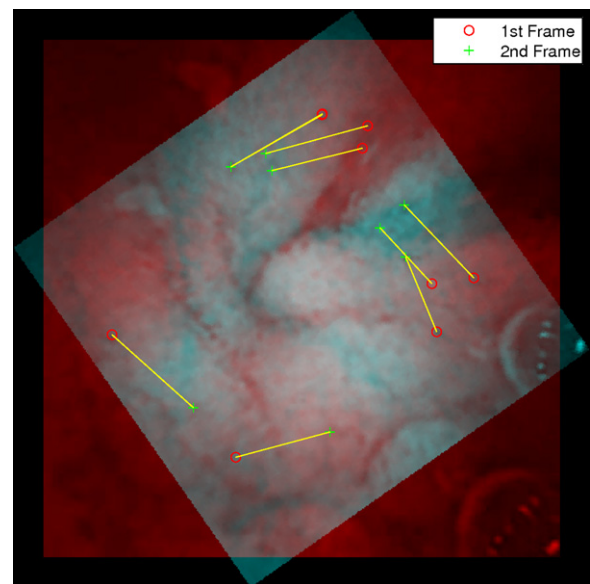
estimate the scaling factor, mainly due to the variance of the interest points between the consecutive frames. However, as mentioned in the previous paragraph, such scaling factors are rarely encountered in actual WCE video sequences, thus the large error values our method produces may not be considered as a serious disadvantage.

As in the case of rotation estimation, the application of the proposed method for the estimation of displacement in practice will be subject to accumulation of the displacement errors. Given the results of table 2, a coarse estimation of the accumulated displacement error for a video clip of 20 s from our dataset will be of the order of 0.1, for scaling factors of 0.6 or larger, whereas for smaller scaling factors, the accumulated error will be higher. However, as we have mentioned earlier such scaling factors are rarely observable and imply a quick motion of the capsule backwards, and their likelihood can further decrease for higher frame rates.

Unlike the case of rotation estimation (section 5.1), no previous displacement estimation method has been proposed to be able to compare the obtained results. Furthermore, the software of WCE provides only a visualization of the position of the capsule in the GI tract which can be used only for qualitative and not quantitative assessment of the displacement of the capsule based on sensor arrays. However, since all the experiments presented in this paper have been performed on publicly available data, comparisons with future methods are feasible.

5.3. Orientation and displacement

To further evaluate the performance of our method, we choose to experiment on the estimation of the rotation angle of an image and a rotated/scaled copy of itself. The results are

**Figure 10.** Inliers estimated using RANSAC, for a video frame and 35° rotated instance, scaled by a factor of 0.8.

presented in tables 3 and 4 and also in figures 11 and 12, respectively. We kept the scale fixed at 2.0 (i.e. a typical scaling factor already discussed in the beginning of this section) and increased the rotation as in subsection 5.1. The results showed that the rotation angle can be estimated using the proposed approach. The mean absolute error is calculated only on the set of frames where RANSAC can find enough correspondences between consecutive frames to be able to estimate rotation and scaling. Figure 10 illustrates a representative example of the correspondences detected between a frame and itself,

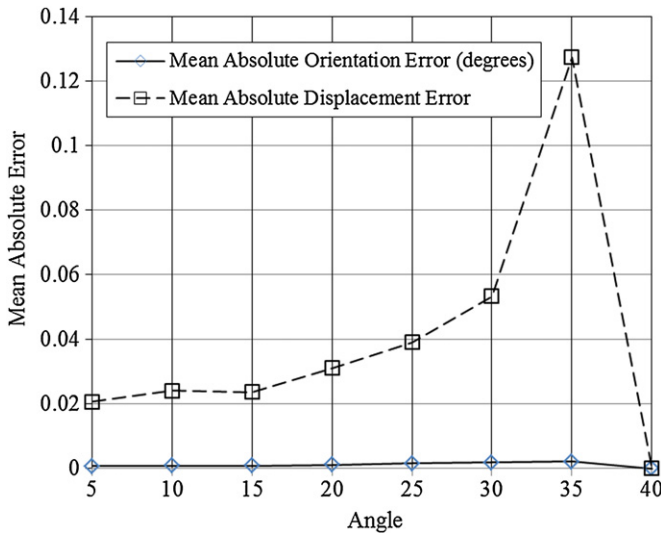


Figure 11. Estimated orientation and displacement errors per angle, at a scale of 2.0.

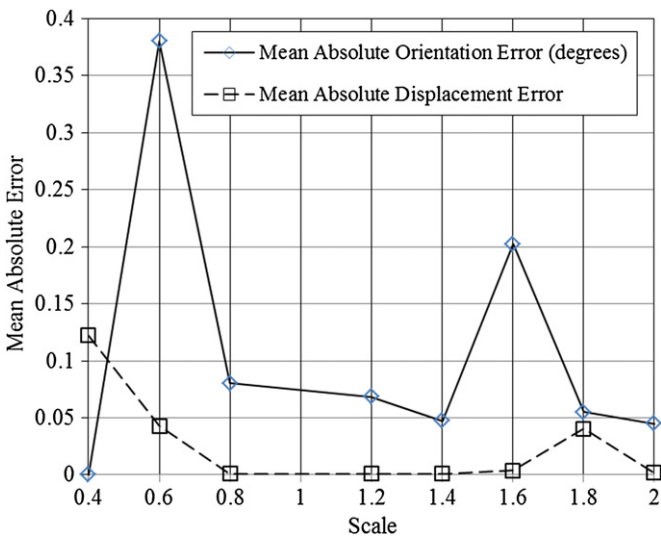


Figure 12. Orientation and displacement errors per scale, at an angle of 30°.

distorted by a rotation angle of 35° and scaled by a factor of 0.8. Although the number of inliers is rather small when compared with those of small rotation angles, they are proved enough for accurate rotation and scale estimation.

In the last experiment we have chosen the largest angle for which the proposed approach produces accurate results at the original scale, i.e. 30°, and distorted our video frames for scales from 0.2 to 5.0. The experimental results obtained for this case are summarized in table 4.

6. Discussion and conclusions

In this work, we presented a novel approach to the estimation of the displacement and orientation of wireless capsule endoscopes for WCE tracking. This approach is based solely on video features and a motion estimation model, whereas current approaches rely on external equipment, such as sensor arrays. Such equipment comes along with

overhead charges contributing to the increase of the overall WCE examination cost, whereas it can be discomforting for the patient by imposing motion limitations during the examination. Furthermore, currently the localization of an abnormality that may require surgical treatment is performed roughly, and thus it is prone to medical errors. Commercial sensor-based WCE can only provide rough 2D position information, presented to the user only as visualization. The proposed WCE tracking approach can be used in conjunction with such WCE devices, to enable not only a qualitative but also a quantitative estimation of the location of the abnormality in the GI tract. For example, if such an abnormality is observed in the WCE video, the endoscopist will be able to roughly estimate in which organ the capsule is located with the visualization tool, rewind the video to the point the capsule enters the organ by viewing the WCE images, and use the proposed video-based method to measure the travelling distance of the capsule from the entrance of the organ to the abnormality. Thus, if surgical treatment is necessary, a surgeon will be able to locate the abnormality more easily, with a lower probability of a mistaken intervention. This process can be further automated if it is combined with current topographic segmentation techniques of the WCE video [21, 22], which enables automatic detection of the entry points for each organ in the GI tract.

The travel distance of the capsule is derived by aggregation of the displacements estimated over the respective range of video frames. One could argue that the presence of intestinal folds quite often may lead to occlusions, hindering the detection of reliable point correspondences. However, considering that a reliable estimation of scaling can be produced by only three point correspondences by RANSAC [10], and that the number of correspondences per frame in our experiments was always greater than four, occlusions due to intestinal folds could hardly affect the capsule displacement or rotation measurements.

The proposed method has been evaluated on publicly available WCE small bowel video clips. The use of such a dataset facilitates the reproducibility of the results obtained and comparisons with future methodologies. The experimentation included simulated scale and rotation transformations of the original video frames serving as a gold standard for error estimation. The mean absolute displacement error is generally small even for a large displacement of the capsule endoscope. As the capsule travels in the GI tract the errors occurring at each step can accumulate and become larger. However, if it is applied for measuring the distance between the entry point of a capsule to an organ and an observed abnormality, the accumulated error will be significantly smaller. The mean absolute error in the rotation estimation is also small, but it may become large for angles larger than 35°.

At this point, we should remark that in a series of measurements, errors may be either positive or negative, thus in some cases the accumulated error can be significantly smaller than the mean absolute error. In figure 13 we illustrate such a case for the case of rotation only. We artificially rotated a frame using increments of 10°, from 0 to 70° and we estimated the rotation for each step using the proposed method. The errors

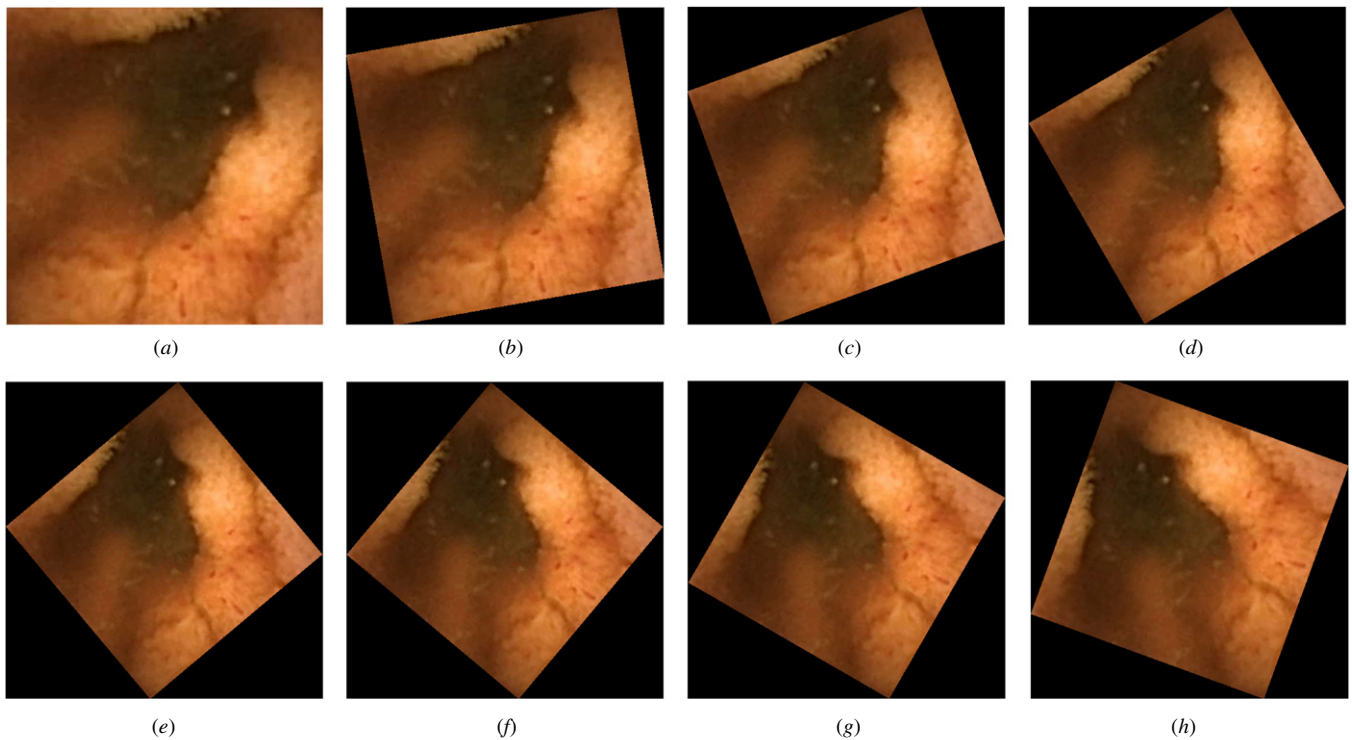


Figure 13. A WCE video frame rotated by 0–70° with a step of 10°.

Table 5. Estimated orientation errors for the sequence of figure 13.

Actual rotation angle (degrees)	Estimated rotation angle at current step (degrees)	Orientation error (degrees)	Estimated rotation angle (degrees)	Accumulative orientation error (degrees)
10	10.33	0.33	10.33	0.33
20	10.53	0.53	20.88	0.88
30	9.71	−0.29	30.59	0.59
40	10.49	0.49	41.08	1.08
50	9.76	−0.24	50.84	0.84
60	9.56	−0.44	60.40	0.40
70	9.79	−0.21	70.19	0.19

obtained are summarized in table 5. It can be observed that three steps had a positive error sign, while four had a negative error sign. The estimated rotation with the proposed method is 70.17°, i.e. the accumulative error is 0.17°, whereas the mean absolute error in this case is 0.36°, which is significantly larger. However, a large error in the case of rotation estimation would not have a significant impact in practice, since the position of an abnormality is primarily derived based on the displacement of the capsule.

The results of this study validate the effectiveness of the proposed method and show that tracking of a capsule endoscope using video-based measurements is feasible. A more thorough evaluation of this method would require ground truth information about the actual distance travelled by the capsule so as to be able to compare it with the results of the video-based measurement methods. A non-invasive approach to cope with this issue would be to develop and use phantom models of the human GI tract in which the capsule motion and visible content can be simulated, e.g. using animal intestines with *a priori* known lengths.

Another issue to be investigated is the effect of the motility of the different sections of the GI tract in the performance of the proposed method, considering also that it can be different for different patients. The observable contractions from a physiological viewpoint can be grouped into phasic and tonic [23]. The former are characterized by a sudden closing of the intestinal lumen, followed by a posterior opening, while the latter corresponds to a sequence in which the intestinal lumen remains completely closed for an undetermined span in time (figure 14). Such contractions may introduce errors in capsule motion estimation. Indicatively we applied the proposed method in this case and we obtained the results illustrated in figure 15. For the first pair of images, i.e. figures 14(a) and (b), a rotation of 11.223° and a scaling factor of 1.126 are obtained, and for the second pair, i.e. figures 14(b) and (c), a rotation of −8.1320° and a scaling factor of 0.7116 are obtained.

The MDR method, which is another, more complicated capsule motion estimation approach, has proved effective in capturing various intestinal motility patterns, and it can be

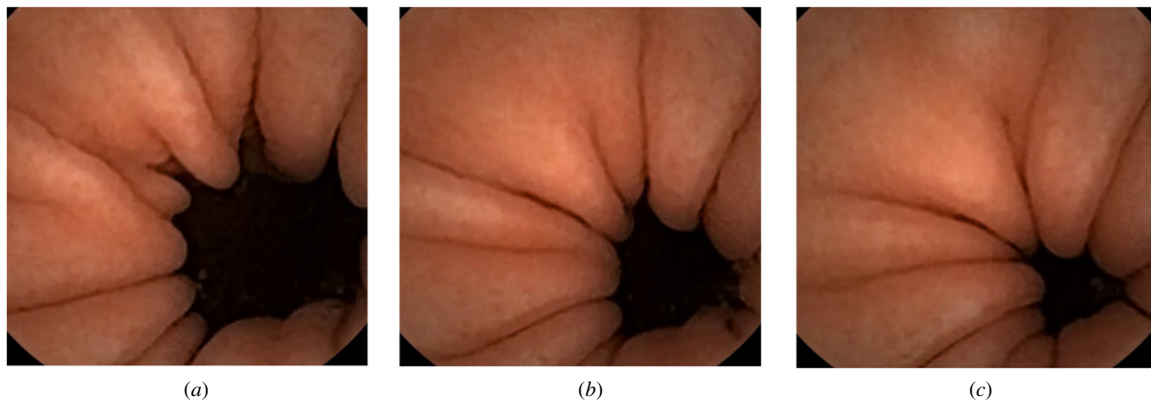


Figure 14. Three consecutive WCE frames, illustrating a contraction.

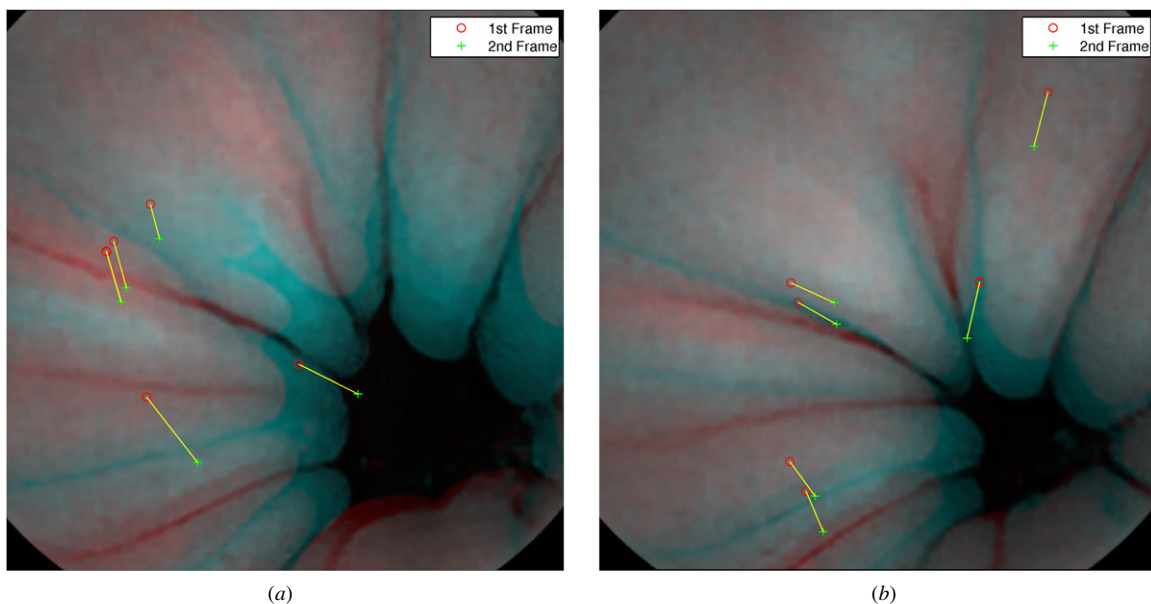


Figure 15. Inliers estimated using RANSAC for the consecutive pairs of the WCE frames of figure 14.

used to indicate video fragments which exhibit segmentary contractions, peristalsis, refraction phases and areas of capsule retention [28]. Similarly to the proposed approach it provides the ability to track only longitudinal motion and roll rotation of the capsule. It has been applied for the assembly of a 2D map representation of the internal surface of the GI tract, but not for capsule localization measurements. An MDR is composed of active interconnected nodes arranged so that they form concentric rings with their centers located near the center of the image frame. Its size changes are used for longitudinal velocity estimation of the WCE within the GI tract. However, since the capsule's motion is not always smooth and parallel to its symmetry axis, the velocity estimation can be inaccurate. The proposed approach is based on interest points extracted from the whole image, rather than interconnected nodes of a specific arrangement. This provides it with the advantage of being less sensitive to abrupt changes of the WCE video content from frame to frame.

Machine-learning-based methods for automatic detection of contractions [24], could be considered in conjunction with the proposed method to prohibit misinterpretation

of contractions as capsule motion, e.g. by excluding the respective parts of the WCE video. Machine-learning-based methods [25] would also be suitable for the detection of irrelevant objects that could affect displacement and rotation estimation, such as stools, usually floating within the lumen with independent motion profiles. Furthermore, we plan to investigate the efficiency of other interest point and low-level feature extraction methods, so as to make our method more robust to larger rotation angles and scaling factors.

Other challenges for future research include experiments on full-length WCE videos obtained from patients with a systematic clinical protocol, performance analysis on different sections of the GI tract (e.g. esophagus, colon, etc) using different capsules, and enhancement of the proposed video-based measurement method to be able to cope with the aforementioned issues.

References

- [1] Iddan G and Swain P 2004 History and development of capsule endoscopy *Gastrointest. Endosc.* **14** 1–9

- [2] Ciuti G, Menciassi A and Dario P 2011 Capsule endoscopy: from current achievements to open challenges *IEEE Rev. Biomed. Eng.* **4** 59–72
- [3] Liu L, Hu C, Cai W and Meng M Q H 2009 Capsule endoscope localization based on computer vision technique *Proc. IEEE Eng. Med. Biol. Soc. (EMBS)* pp 3711–4
- [4] Spyrou E, Kalantidis Y, Toliass G, Mylonas P and Kollias S 2010 Intelligent content retrieval using a visual vocabulary and geometric constraints *FUZZ-IEEE 2010, in conjunction with The 2010 IEEE World Congress on Computational Intelligence (IEEE WCCI 2010)*
- [5] Bay H, Ess A, Tuytelaars T and Van Gool L 2008 Speeded-up robust features (SURF) *Comput. Vis. Image Underst.* **110** 346–59
- [6] Lowe D G 2004 Distinctive image features from scale invariant keypoints *Int. J. Comput. Vis.* **60** 91–110
- [7] Lindeberg T 1994 *Scale-Space Theory in Computer Vision (Kluwer International Series in Engineering and Computer Science)* (Dordrecht: Kluwer) p 256
- [8] Neubeck A and Gool L V 2006 Efficient non-maximum suppression *Int. Conf. on Pattern Recognition (ICPR)* vol 3, pp 850–5
- [9] Fischler M A and Bolles R C 1981 Random sample consensus—a paradigm for model fitting with applications to image analysis and automated cartography *Commun. ACM* **24** 381–95
- [10] Hartley R and Zisserman A 2000 *A Multiple View Geometry in Computer Vision* (Cambridge: Cambridge University Press)
- [11] Given Imaging 2012 www.givenimaging.com
- [12] OpenSurf Library 2012 www.chrisevansdev.com/computer-vision-opensurf.html
- [13] Matlab www.mathworks.com
- [14] Pan G, Xin W, Yan G and Chen J 2011 A video wireless capsule endoscopy system powered wirelessly: design, analysis and experiment *Meas. Sci. Technol.* **22** 065802
- [15] Hara A K, Leighton J A, Sharma V K, Heigh R I and Fleischer D E Imaging of small bowel disease: comparison of capsule endoscopy, standard endoscopy, barium examination, and CT *Radiographics* **25** 697–711
- [16] Salerno M, Ciuti G, Lucarini G, Rizzo R, Valdastrì P, Menciassi A, Landi A and Dario P 2012 A discrete-time localization method for capsule endoscopy based on on-board magnetic sensing *Meas. Sci. Technol.* **23** 015701
- [17] Spyrou E and Iakovidis D K 2012 Homography-based orientation estimation for capsule endoscope tracking *Proc. IEEE Int. Conf. on Imaging Systems and Techniques (Manchester, UK)* pp 101–5
- [18] Liu L, Liu W, Hu C and Meng M Q H 2011 Hybrid magnetic and vision localization technique of capsule endoscope for 3D recovery of pathological tissues *Proc. World Congress on Intelligent Control and Automation (Taiwan)* pp 1019–23
- [19] Liu H, Pan N, Lu H, Song E, Wang Q and Hung C-C 2012 Wireless capsule endoscopy video reduction based on camera motion estimation *J. Digit. Imaging* **26** 287–301
- [20] Liu Y, Tillo T, Xiao J, Lim E and Wang Z 2011 2D to cylindrical inverse projection of the wireless capsule endoscopy images *Int. Congress on Image and Signal Processing*
- [21] Than T D, Alici G, Zhou H and Li W 2012 A review of localization systems for robotic endoscopic capsules *IEEE Trans. Biomed. Eng.* **59** 2387–99
- [22] Cunha J P, Coimbra M, Campos P and Soares J M 2008 Automated topographic segmentation and transit time estimation in endoscopic capsule exams *IEEE Trans. Med. Imaging* **27** 19–27
- [23] Kellow J *et al* 1999 Principles of applied neurogastroenterology: physiology/motility-sensation *Gut* **45** II17–24
- [24] Vilariño F, Spyridonos P, Deiorio F, Vitria J, Azpiroz F and Radeva P 2010 Intestinal motility assessment with video capsule endoscopy: automatic annotation of phasic intestinal contractions *IEEE Trans. Med. Imaging* **29** 246–59
- [25] Muthukudage J, Oh J, Tavanapong W, Wong J and De Groen P C 2011 Color based stool region detection in colonoscopy videos for quality measurements *Advances in Image and Video Technology (Lecture Notes in Computer Science* vol 7087) (Including Subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*) (Dordrecht: Springer) pp 61–72
- [26] Tsai R Y 1987 A versatile camera calibration technique for 3D machine vision *IEEE J. Robot. Autom.* **RA-3** 323–44
- [27] Faugeras O 1996 *Three-Dimensional Computer Vision: A Geometric Approach* (Cambridge, MA: MIT Press)
- [28] Szczypiński P M, Sriram R D, Sriram P V and Reddy D N 2009 A model of deformable rings for interpretation of wireless capsule endoscopic videos *Med. Image Anal.* **13** 312–24
- [29] Given Imaging Atlas 2012 capsuleendoscopy.org