Capsule Endoscope Localization based on Visual Features

Dimitris K. Iakovidis, *Member, IEEE*, Evaggelos Spyrou, Dimitris Diamantis, Ilias Tsiompanidis

***Abstract*—Computational analysis of wireless capsule endoscopy (WCE) videos has already proved its potentials in the discovery or characterization of lesions and in the reduction of the time required by the endoscopists to perform the examination. An open problem that has only partially been addressed is the localization of the capsule endoscope in the gastrointestinal (GI) tract. Previous works have been based mainly on external, wearable, sensors. In this paper we propose a novel approach based solely on visual information extracted from WCE videos. This approach is based on a feature tracking method for visual odometry, which enables the estimation of both the rotation and the displacement of a capsule endoscope from reference anatomical landmarks. Its implementation is based on a novel, open access Java Video Analysis (JVA) framework, which enables quick and standardized development of intelligent video analysis applications. The experimental evaluation presented in this paper, indicates the feasibility of the proposed methodological approach and the efficiency of its implementation.**

I. Introduction

RESEARCH on computer-aided analysis of video endoscopy data has been evolving during the last decade. Machine intelligence approaches were among the first considered for the analysis of conventional colonoscopy and gastroscopy videos [1, 2], and since then, a variety of methods have been proposed [3].

Wireless capsule endoscopy (WCE) has revolutionized the field of GI endoscopy [4]. It is performed by a swallowable capsule with the size of a large vitamin that includes a miniature color video camera, wirelessly transmitting thousands of video frames during its journey to the anus. It has become a standard imaging technique with a utility that has been extended to the whole GI tract, from the esophagus to the colon. It is recommended as a diagnostic or monitoring tool for various diseases including Crohn's disease, ulcers, polyps and cancer [5].

Challenges with regards to computer-aided analysis of WCE videos that have been addressed, include the discovery and characterization of specific types of lesions, including polyps and ulcers [6], and the reduction of the time required during the off-line examination of the videos by the endoscopists, which usually lasts for several hours [7]. Another challenge that has only partially been addressed by computational approaches is the localization of the capsule in the GI tract. Accurate capsule localization is necessary in cases such as surgical treatment of suspicious lesions, and the monitoring of the progress of a disease e.g. Crohn's disease, over time by follow up WCE examinations.

The localization of a capsule endoscope within the body is typically performed by external sensors [8]. Commercially available systems provide only a rough estimation of capsule's location based on wearable radio-frequency (RF) sensor arrays. In such systems, the location of the capsule is indicated roughly by a graphic unitless representation provided by the user interface of the video reading software, such as [9]. Recent methods that promise more accurate localization of the capsule endoscopes have been based on magnetic sensor arrays. However, they are still in experimental phase and tested only in laboratory setups [8]. Only a few approaches have been proposed for capsule localization based on visual features. These include methods addressing only the estimation of the rotation angle of the capsule [10, 11], and temporal video segmentation methods. The latter are capable of inferring the part of the GI tract in which the capsule is located i.e. oesophagus, stomach, small intestine and colon, by applying pattern recognition techniques for the identification either of the tissues of these parts [12] or of the transition points between these parts, such as the esogastric junction, the pylorus and the ileocecal valve [13].

In this paper we propose a novel application of a feature tracking method previously used in the context of visual odometry [14], capable of estimating both the rotation and the displacement of the capsule endoscope from reference anatomical landmarks such as the transition points between the parts of the GI tract that can be detected by pattern recognition [13]. As indicated in [15], visual odometry is not affected by wheel slip in uneven terrain or other adverse conditions, and it has been demonstrated that compared to wheel odometry, it provides more accurate trajectory estimates, with relative position error ranging from 0.1 to 2%.

Furthermore, we propose a novel approach to address the implementation aspects involved. We introduce the Java Video Analysis (JVA) framework[1] as an infrastructure for the development of future video analysis applications in a standard, reusable and extensible way. This framework could motivate further development of standardized open access or even open source video analysis components for

Manuscript received July 30, 2013. This work was supported in part by the Technological Educational Institute of Central Greece (formerly Technological Educational Institute of Lamia), Lamia, Greece.

D. K. Iakovidis, E. Spyrou and D. Diamantis are with the Department of Computer Engineering, Technological Educational Institute of Central Greece, 3rd km Old National Road Lamia-Athens, 35100 Lamia, Greece; e-mails: dimitris.iakovidis@ieee.org, {vspyrou, ddiamantis}@teilam.gr.

I. Tsiompanidis, is with the Department of Gastroenterology, University Hospital of Larisa, Larisa, Greece.

[1]JVA can be downloaded from http://innovation.teilam.gr/jva/, where documentation and examples are provided.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

endoscopy and other application domains.

The rest of this paper consists of four sections. Section II describes the proposed methodology. Section III introduces JVA and the details involved in the implementation of the proposed methodology. The experimental evaluation of the visual capsule endoscope localization is described in Section IV, and the last section summarizes the conclusions and suggests future research directions.

II. Capsule Endoscope Localization

The proposed approach is based on a relative pose estimation algorithm proposed by Nister [16]. It involves low-level feature extraction from consecutive video frames, detection of interest point correspondences between them, and estimation and decomposition of the essential matrix.

A. Feature extraction

The first step of the proposed approach consists of low-level visual feature extraction. We choose the well-known SURF (Speeded-Up Robust Features) [17], which have been proven to achieve high repeatability and distinctiveness while remaining invariant to many geometric and illumination changes. The first step of the algorithm extracts local interest key-points using a fast approximation of the Hessian Matrix and convolutions of the initial image with box filters at several scales (octaves). The second step extracts a visual descriptor that captures the intensity content distribution around each point. Additionally, their extraction time is significantly faster when compared to other algorithms such as SIFT [17]. In the following, with x and x' we will denote the extracted sets of points from two given frames, while with v and v' the respective visual descriptors.

B. Detection of point correspondences

The next step is to detect point correspondences within sets x and x' between two given frames, i.e. a mapping $x_i \leftrightarrow x'_i$. As the camera viewpoint changes, similar visual features may appear in a totally different spatial layout. One of the widely adopted techniques to estimate a set of pairs of point-to-point correspondences that follow the same geometric transformation, is RANSAC (RANdom Sample Consensus) algorithm [19].

Initially, we select a set of correspondences based solely on the sets of visual features, v and v', i.e. two points x_i and x'_i are considered similar when the distance between their visual descriptors v_i and v'_i is beneath a user defined threshold. This way, a set of "tentative" correspondences is formed. RANSAC then extracts a subset of correspondences that consists only of those that follow the same geometric transformation (inliers), after an iterative process. We should emphasize that RANSAC works well even in the presence of many "false" correspondences (outliers). An example of the inlier correspondences among interest points based on SURF features for the two consecutive WCE video frames of Fig. 1 is illustrated in Fig. 2. Should we assume that image transformations are affine, we are able to estimate scalings

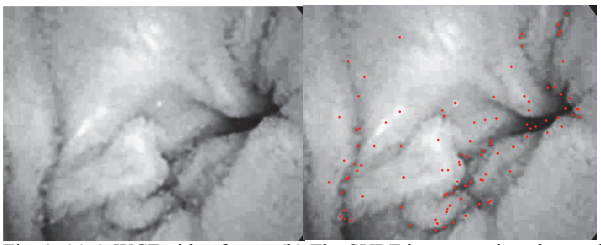

Fig. 1. (a) A WCE video frame. (b) The SURF interest points detected on the WCE video frame are denoted by red dots.

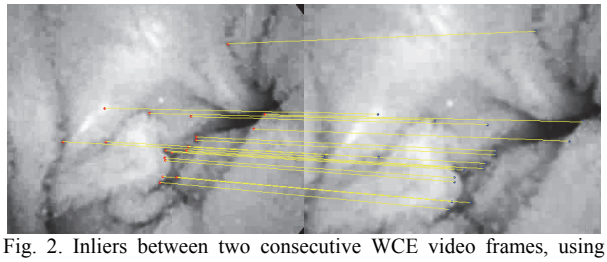

Fig. 2. Inliers between two consecutive WCE video frames, using RANSAC.

between two frames,as the ratio of the distances between the two pairs of inliers, e.g. for pairs $x_1 \leftrightarrow x'_1$ and $x_2 \leftrightarrow x'_2$, we can estimate scaling as $s_{12} = \|x_1 - x'_1\| / \|x_2 - x'_2\|$, where by $\| \bullet \|$ we denote the Euclidean norm.

C. Essential Matrix Estimation and Decomposition

In order to estimate the camera rotation between two video frames, we choose to adopt a well-known pose determination algorithm, proposed by Nister [16]. This algorithm reconstructs camera position and scene structure, based on point correspondences from different camera viewpoints. The input of the algorithm is the set of inliers, selected by RANSAC. Its baseline version requires exactly five point correspondences. We typically encounter such a number between consecutive WCE video frames.

The relative pose algorithm proceeds as follows: Each of the five point correspondences defines a constraint. To solve this set of equations in the best possible efficient manner, Nister applied QR-factorization. The essential matrix E is then estimated [20]. Let $y \subset x$ and $y' \subset x'$ denote the corresponding sets of inliers within the given pair of images. For the sake of simplicity we may assume that y' and x' consist of five random correspondences. We remind that E satisfies $y'TEy = 0$. We also adopt QR and decompose the essential matrix into its components. One of its components is the rotation matrix $R = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$, from which rotation θ is then calculated.

III. JVA Framework

State of the art frameworks for video analysis have either a methodological orientation, such as [21], or a software development orientation, such as the Java Media Framework (JMF) [22], which provides a unified architecture and messaging protocol for managing acquisition, processing, and delivery of time-based media data. Software libraries, such as OpenCV, JavaCV, BoofCV [23, 24], FFMpeg and

Xuggler [25], are often reported as frameworks; however, they mainly provide functions rather than a specific template for software development.

In this paper we introduce JVA, a software development framework developed in Java, which through its application programming interface (API) provides a higher level of abstraction, under which, various implementations of video acquisition, processing, analysis and visualization algorithms (e.g. based on other libraries or frameworks), can be integrated. In order to maximize code reusability and system reconfigurability, the code produced is not tightly coupled with the core framework, and the implementations can be seamlessly integrated within JVA and accessed in a standard and uniform way. JVA also offers platform independence, code reusability and easy application deployment to the World Wide Web. It consists of a core and four independent functional components, each of which can be deployed as a standalone or as a pluggable entity (plugin) to the core framework component. The core component provides a standard API for the main functionalities of the plugins.

The plugin-based architecture provides the framework with a remarkable reconfigurability for implementation of a variety of video analysis tasks without any modification to the core component, whereas each plugin can be manipulated independently. The parameters of the plugins can be configured through an external, extensible text configuration file. This enables easier experimentation for parameter tuning, and easier access to parameters from external wrapper graphical user interfaces (GUIs) addressing specific applications. Furthermore, debugging is supported by a functional GUI which visualizes the results of the video processing and analysis.

JVA framework consists of four plugin components with respective interfaces to be implemented in a video analysis application: Image Data-Source (IDS), Image Processor (IP), Image Analyzer (IA), and Output Handler (OH) component. IDS enables data acquisition, IP enables processing of the input images, IA is responsible for the analysis of single or multiple video frames and finally, the results of the video analysis are directed to OH, which is responsible for image/video display, storage and transfer to remote destinations.

The implementation of the described capsule tracking methodology exploits all four plugins. IDS utilizes FFMpeg [25] functionalities and was used for the acquisition of WCE video frames; IP utilizes ImageJ [26] functionalities and was used for cropping a grey-level rectangular sub-image, inscribed within the oval field of endoscopic content of each WCE video frame; IA describes the sampling policy required for SURF extraction, and utilizes functionalities of existing implementations [24] to estimate point correspondences and the essential matrix; OH implements displacement and rotation estimations, and displays output.

IV. Experiments and Results

A number of experiments were performed to assess the feasibility of the localization methodology and the performance of its JVA-based implementation. A total of 30 randomly selected WCE video clips from Given Imaging Atlas [27] formed the dataset used in our experiments. The experiments were performed on a laptop with Intel Pentium T4300 Core 2, 1MB Cache, 2.10 GHz, 800 MHz FSB processor, with 3GB RAM.

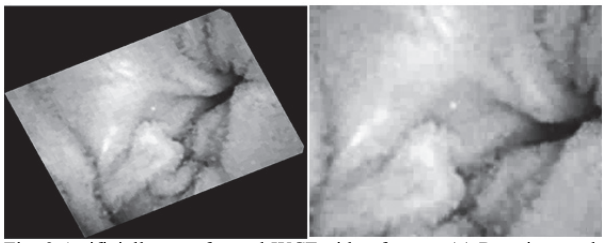

Fig. 3.Artificially transformed WCE video frames. (a) Rotation angle has been set to 25°. (b) Scaling factor has been set to 1.5.

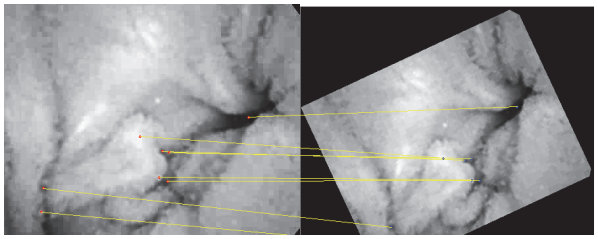

Fig. 4.Inliers estimated using RANSAC, for a video frame and its 25° rotated instance.

We extracted features from 4 octaves and from 4 scales per octave. RANSAC was then applied on the sets of points of two given video frames. As we noted in Section II, we first created a set of "tentative" correspondences, based solely on visual features. A large number of iterations (8000) was considered to ensure that at least one of the sets of random samples does not include an outlier, with a probability 99.9%. Since to the best of our knowledge there does not exist a WCE data set with known rotation and scaling factors amongst video frames, we evaluated the method of Section II using artificially transformed video frames from the available dataset; e.g. the frame of Fig. 1(a), rotated counter-clockwise by 25° is illustrated in Fig. 3(a), and scaled by a factor of 1.5 is illustrated in Fig. 3(b). Figure 4 illustrates the subset of inliers found by RANSAC for the case of rotation only.

Tables I and II summarize the experimental results obtained for the estimation of rotation and scaling, respectively. A total of nine different rotation angles were tested from 5° to 45° with a rotation step of 5°. As it can be noticed the error between the actual and the estimated rotation angle is rather low up even for a 45°rotation, where the state of the art methods [10, 11] produce very large errors. More specifically, the method of [10] is based on Kanade-Lucas-Tomasi (KLT) optical flow computation and produces very large errors for rotations larger than 30°, whereas in [11] the method based on homography matrix estimation produces very large errors for rotation angles of 40° and higher. As for scaling, the experiments indicate that the error between the actual and the estimated scale is of the order of 10^{-1}. For scaling factors larger than 3.0 the error and

its variance increases.

The processing times of all steps of the proposed algorithm and for a total of 100 video frames as implemented by the JVA framework, are summarized in Table III. It can be observed that the majority of the actual algorithm processing time is dedicated to RANSAC.

V. Conclusions

In this work we presented a novel methodology for localization of capsule endoscopes based solely on visual features, and implemented using our novel JVA framework. The experimental results indicate the feasibility of the approach and the efficiency of its implementation.

Within our future goals is to investigate approaches that would lead to lower scaling and rotation errors, optimize and evaluate the proposed framework in various applications, implement and integrate different video analysis tasks, such as image mining [7], and cope with practical issues related to capsule tracking such as evaluation in more realistic conditions e.g. phantom model where ground truth information about distance estimation will be available.

TABLE I
Experimental Results - Rotation

Actual Angle (degrees)	Mean Estimated Angle ± Std. Dev
5	5.63 ± 0.26
10	10.83 ± 0.47
15	15.90 ± 0.57
20	21.02 ± 0.41
25	25.86 ± 0.44
30	30.96 ± 0.51
35	35.92 ± 0.41
40	40.93 ± 0.59
45	45.78± 0.36

TABLE II
Experimental Results - Scaling

Actual Scale	Estimated Scale ± Std. Dev
0.2	0.33±0.17
0.4	0.52±0.97
0.6	0.62±0.15
0.8	0.97±0.37
1.0	1±0
2.0	2.19±0.25
3.0	3.37±1.23

TABLE III
Extraction times for an indicative WCE video sequence.

Step	Avg. Time ± Std. Dev
Frame Extraction (per frame)	5.41± 0.14 ms
Video Processing (per frame)	36.62± 0.46 ms
SURF Extraction (per frame)	29.71± 0.37 ms
RANSAC (per frame)	130.63 ± 7.8 ms
Essential Matrix (per frame)	3.33 ± 0.35 ms
Video Reconstruction	469.00 ms
Total Time	**20706 ms**

References

[1] S. Karkanis, D. Iakovidis, D. Maroulis, G. Magoulas, and N. Theofanous, "Tumor recognition in endoscopic video images using artificial neural network architectures," in *Euromicro Conference, 2000. Proceedings of the 26th*, vol. 2, pp. 423–429, IEEE, 2000.

[2] D. Iakovidis, D. Maroulis, and S. Karkanis, "An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy," *Comp.in Biology and Medicine*, vol. 36, no. 10, pp. 1084–1103, 2006.

[3] M. Liedlgruber and A. Uhl, "Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review," *Biomedical Engineering, IEEE Reviews in*, vol. 4, pp. 73–88, 2011.

[4] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, p. 417, 2000.

[5] J. Keller, C. Fibbe, U. Rosien, and P. Layer, "Recent advances in capsule endoscopy: development of maneuverable capsules.," *Expert Rev GastroenterolHepatol*, vol. 6, pp. 561–566, Sep 2012.

[6] A. Karargyris and N. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos," *Biomedical Engineering, IEEE Trans. on*, vol. 58, no. 10, pp. 2777–2786, 2011.

[7] D. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Comp.Med.Imaging and Graphics*, vol. 34, no. 6, pp. 471–478, 2010.

[8] T.D. Than, G. Alici, H. Zhou, and W. Li, "A review of localization systems for robotic endoscopic capsules," IEEE Transactionson Biomedical Engineering, vol. 59, no. 9, pp. 2387-2399, 2012.

[9] Given Imaging Software, http://www.givenimaging.com/en-int/Innovative-Solutions/Capsule-Endoscopy/Software/Pages/default.aspx, accessed Jun. 2013.

[10] L. Liu, C. Hu, W. Cai, and M. Meng, "Capsule endoscope localization based on computer vision technique," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 3711–3714, IEEE, 2009.

[11] E. Spyrou, and D. Iakovidis, "Homography-based orientation estimation for capsule endoscope tracking," in *Imaging Systems and Techniques (IST), 2012 IEEE Int. Conf. on*, pp. 101–105, IEEE, 2012.

[12] M. Mackiewicz, J. Berens, and M. Fisher, "Wireless capsule endoscopy color video segmentation," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 12, pp. 1769–1781, 2008.

[13] J.P. Cunha, M. Coimbra, P. Campos, J.M. Soares, "Automated Topographic Segmentation and Transit Time Estimation in Endoscopic Capsule Exams," IEEE Trans. Medical Imaging, vol. 27, no. 1, 2008.

[14] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in Proc. Int. Conf. Comp. Vision and Pattern Recognition, 2004, pp. 652–659.

[15] D. Scaramuzza and F. Fraundorfer, "Visual Odometry, Part I: The First 30 Years and Fundamentals," Robotics & Automation Magazine, IEEE, pp. 80-92,Dec. 2011

[16] D. Nister, "An efficient solution to the five-point relative pose problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 756–770, 2004.

[17] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[18] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.

[19] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun.ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[20] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, vol. 2. Cambridge Univ Press, 2000.

[21] S. Park, D. Sargent, I. Spofford, K. Vosburgh, *et al.*, "A colon video analysis framework for polyp detection," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1408–1418, 2012.

[22] S. Sullivan, L. Winzeler, D. Brown, and J. Deagen, *Programming with the Java media framework*. John Wiley & Sons, Inc., 1998.

[23] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Incorporated, 2008.

[24] P. Abeles, "Resolving implementation ambiguity and improving surf," *arXiv preprint arXiv:1202.0492*, 2012.

[25] F. Bellard, "FFMpeg multimedia system, 2006," 2006.

[26] M. Abramoff, P. Magalhaes, and S. Ram, "Image processing with ImageJ," *Biophotonics international*, vol. 11, no. 7, pp. 36–42, 2004.

[27] Given Imaging Atlas, http://capsuleendoscopy.org/, accessed April 2012.