Efficient Homography-Based Video Visualization for Wireless Capsule Endoscopy

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

***Abstract*—Wireless capsule endoscopy (WCE) is performed by a swallowable pill capsule equipped with a camera wirelessly transmitting color video frames to an external receiver. The resulting video consists usually of several thousands of frames and its visual examination requires hours of endoscopists' undivided attention. In this paper we propose a novel visualization approach for WCE which enables faster examination of the endoscopic video, while providing a broader field of view. This is achieved by an algorithm that iteratively samples clusters of consecutive frames from the original video. The frames of each cluster are geometrically transformed, so as to generate a seamless collage subsequently projected into a new frame without any information loss. The new frames compose a new WCE video with a smaller number of frames. The video frame collage is based on homography matrix estimation from frame correspondences. The experiments show that the length of the WCE video, and therefore the reading times required can be significantly reduced.**

I. Introduction

WIRELESS capsule endoscopy (WCE) is recommended as a diagnostic or monitoring tool for various diseases including Crohn's disease, ulcers, polyps and cancer [1,2]. It is performed by a swallowable capsule with a size of a large vitamin that includes a color video camera wirelessly transmitting thousands of video frames during its journey to the anus. A full-length WCE video consists of approx. 50000 frames and the time required by endoscopists to read it can range from 45 minutes to several hours, while their attention should remain undivided. This challenge for human capabilities is often the cause of the very low detection rate of clinically significant findings, which has been estimated to be of the order of only 40% [3].

A number of methods have been proposed to improve the WCE video reading efficiency. These include video summarization methods based on image mining for the selection of most representative frames from the whole WCE video [4, 5], and visualization methods. Commercial software such as RAPID Reader [6] provides a straightforward but effective visualization of multiple consecutive frames simultaneously in a 2D arrangement. This enables the endoscopists to evaluate two, four or more frames at the same time, with a potential for a proportional reduction of the reading times. However, this potential is limited by human capabilities, as it is very difficult for humans to fully perceive the content of more than four different images projected simultaneously.

Another commercially available approach to efficient visualization of WCE videos is QuickView. According to this approach only highlights from a whole video are presented. However, the validity of this approach has been strongly questioned and criticized [7, 8], since frames with important diagnostic information may be skipped. In order to avoid frame skipping, Hai et al [9] proposed a method for adaptive control of video display. The principle behind this approach is that the WCE video can be played back at high frame rate in stable smooth frame sequences to save time and when sudden rough changes happen, the frame rate can decrease; thus enabling the assessment of possibly suspicious findings in detail.

Chu et al [10] proposed a methodology that constructs epitomes from a set of consecutive frames based on prior knowledge regarding the normal and the possibly abnormal tissues, and the expected non-informative contents. The visual summarization obtained by the epitomes aims to be semantically organized, and its experimental evaluation showed that it can reduce the number of images down to less than 10% of the original videos. It should be noted that not all visualization methods aim to WCE reading times reduction; for example, the method proposed in [11] provides a visualization method for intestinal motility inspection.

The concept of automatic generation of image collages has recently opened new perspectives in medical image/video visualization. Representative examples include methods for improving visualization of retinal images [12], and construction of surface mosaics of the bladder from endoscopic video [13]. Recently, in the context of WCE the utility of this concept has been outlined for visualization of WCE images acquired with an unconventional capsule endoscope equipped with special optics capable of capturing 360^{o} panoramic images [14]. That study advises the application of image stitching as a general approach to form a dissected view of the whole GI tract.

Motivated by that study, we propose a novel application of automatic image stitching for efficient visualization of conventional WCE videos. This approach aims to improve the commercially adopted multi-frame visualization [6] by enabling visualization of even more than four frames simultaneously as a collage of frames in a single image of a broader field of view, without any information loss. The video frame collage is based on homography matrix estimation from frame correspondences.

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.

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

The rest of this paper consists of three sections. Section II describes the proposed visualization approach. The results from its experimental evaluation are apposed in Section III, and the conclusions of this study are summarized in the last section.

II. Collage Visualization

The viewpoint of a camera used for WCE video acquisition changes with its motion; therefore, the visual features of two consecutive frames displaying the same planar surface of the GI tract can appear in a different spatial layout. By identifying correspondences between key-points of the two frames, a projective transformation model, known as homography [15], can be estimated to describe how these frames are spatially related.

Given a 3×3 homography matrix H and two points $p = (x_1, x_2, x_3)$ and $p' = (x_1', x_2', x_3')$, where x_i and x_i' represent their homogeneous coordinates, the projective transformation which maps p to p' can be expressed as

$$p' = H \cdot p \tag{1}$$

which can also be expressed in a matrix form as

$$\begin{pmatrix} p_1' \\ p_2' \\ p_3' \end{pmatrix} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} \cdot \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \tag{2}$$

The proposed visualization approach involves: a) detection of key-points based on visual features of the video frames; b) detection of correspondences between the key-points; c) homography estimation; and d) clustering of the video frames and formation of frame collages that seamlessly integrate their whole visual content into a compact single-frame representation.

A. Detection of key-points

The detection key-points within the WCE video frames is based on SURF (Speeded-Up Robust Features), which are local low-level features, that have shown robustness in most geometric transformations, as well as illumination and viewpoint invariance [16]. SURF extraction involves an algorithm for the detection and the description of invariant key-points. The first step of this algorithm relies on 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 image texture around each point.

In the rest of this paper, the sets of SURF key-points for two consecutive WCE frames will be denoted by k and k', respectively, while by v and v' we will denote their respective visual descriptions. A typical WCE video frame and the detected key-points are illustrated in Fig. 1. The next step of the proposed approach aims to determine of a set of geometrically consistent key-point correspondences between two consecutive video frames.

B. Detection of key-point correspondences

A set of tentative correspondences between key-points is created, by assessing the visual similarity of the key-points detected between consecutive WCE video frames. This process involves the estimation of the Euclidean dissimilarity matrix between v and v'. Based on this matrix, the most similar vector pairs are selected, and subsequently, the largest subset of correspondences that follow the same geometric transformation is investigated. These useful correspondences, known as *inliers*, will be used for the estimation of H in the next step of the proposed visualization approach. The remaining correspondences, known as *outliers*, will be considered as noise to be discarded.

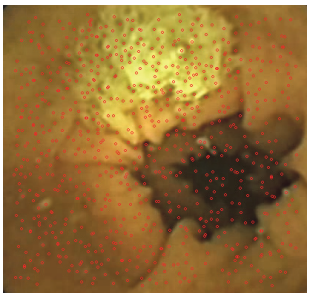


Fig. 1. Representative WCE video frame. A total of 798 SURF key-points were detected. They are depicted as red dots overlaid to the original frame.

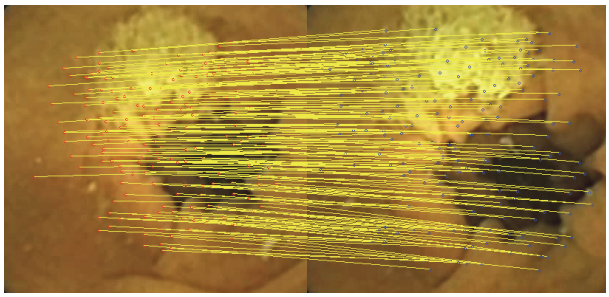


Fig. 2. Inliers between the WCE video frame illustrated in Fig. 1 (on the right) and its previous frame (on the right), using RANSAC. A total of 804 key-points were detected in the previous frame, a total of 301 correspondences have been detected between the key-points. The total number of inliers, identified by the yellow lines connecting the respective pixels, is 180.

In order to find a mapping $k \leftrightarrow k'$ such that a near optimal homography can be estimated, we apply RANSAC (RANdom SAmple Consensus) algorithm. RANSAC is particularly useful when the number of outliers is large with regard to inliers [17].

C. Homography estimation

In the video frame registration task, p and p' represent the pixels of the key-point correspondences found in two consecutive video frames and they are therefore more naturally described by their respective Euclidean coordinates (x, y) and (x', y') respectively. The homography matrix as expressed by Eq. (2) can then be formulated as

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} \cdot \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} \tag{3}$$

The application of RANSAC for the detection of the

inliers and the estimation of the parameters of Eq. (3) proceeds as follows:

1. Random quadruplets are selected from the set of tentative correspondences.
2. A homography is estimated and all other correspondences are tested against it. When a correspondence fits to the estimated homography, it is considered to be an inlier.
3. The homography is re-estimated based on all inliers.
4. The best homography is the one supported by the largest set of inliers (consensus set).

The output of RANSAC is the homography matrix H that maximizes N. The consensus set estimated by RANSAC consists of N inliers. Let $R = \{r_i\}$ and $R' = \{r_i'\}$, $i = 1,2,\dots,N$ denote the sets of the inliers found in the first and in the second frame of two consecutive frames respectively. It is obvious that we can write $r_i \leftrightarrow r_i'$, since each inlier of the first frame corresponds to exactly one of the second. The homography matrix H is then defined as the matrix that satisfies $r_i' = Hr_i$, for $i = 1,2,\dots,N$.

An illustrative example of the application of RANSAC in WCE frames is presented in Fig. 2. One may expect that we should not encounter a large number of outliers between two consecutive WCE video frames. However, this could happen in an ideal case. In WCE videos many false correspondences are introduced due to image noise, such as the MPEG-1 block artifacts which were very common in the dataset used in the experiments described in the next section. However, RANSAC overcomes this problem and is able to estimate this homography.

D. *Frame clustering*

After estimating H from video frames f_{i-1} and f_i we apply its inverse H^{-1} to f_i, so that it matches the spatial layout of the GI tract presented in f_{i-1}. Considering that p is a pixel of f_i, , its transformation to p_i' using H is estimated by solving Eq. (3) as

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix}^{-1} \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (4)$$

Once this transformation is applied, f_i is superimposed to f_{i-1} based on the inliers selected by RANSAC and the pairs of SURF key-points that constitute the inlier correspondences should approximately match. In order to achieve a seamless blending between the superimposed frames the method proposed in [18] is applicable.

This process is repeated for a number of c_j WCE video frames, where j = 1,2,…C, until the number of inliers between f_{i-1} and f_i falls below an empirically defined threshold T. These c_j video frames are all superimposed in a single image, which has the form of an image collage. This image is then added as a new frame f_i' in a new video of collages.

Considering that the frame rate of contemporary WCE cameras ranges between 2 and 4 frames per second (fps), most of the consecutive WCE video frames should contain redundant information (visualized as overlapping image regions). Therefore, the number of clusters C, in which the whole WCE video is temporally segmented, is expected to be significantly smaller than the number of video frames of the whole video. Since each frame cluster is visualized in a single frame, the length of the new video of collages and the time required to read it, are expected to be significantly smaller than the length and the reading time required for the original video.

III. Results

The proposed visualization approach was implemented as a plugin in Java Video Analysis (JVA) framework, a novel platform-independent software development framework for video analysis applications recently released by our research group[1]. Its experimental evaluation was based on a total of 30 randomly selected WCE video clips from Given Imaging Atlas [19].

The SURF extraction parameters used include 4 octaves and 4 scales per octave. The creation of tentative correspondences was performed using a similarity threshold set to 300. The number of iterations of RANSAC was set to 8000, in order to ensure that at least one of the sets of random samples would consist solely by inliers, with a 99.9% probability. In the collage generation process, we considered only consecutive frames with $T = 5$ or more inlier correspondences.

According to Section II, the application of the proposed method resulted in a set of smaller video clips than the originals, containing collages of frame clusters. The overall length reduction achieved for the available dataset was 85.6%. Figure 3(a) illustrates a representative cluster of 10 consecutive WCE video frames, and the respective generated collage is illustrated in Fig. 3(b).

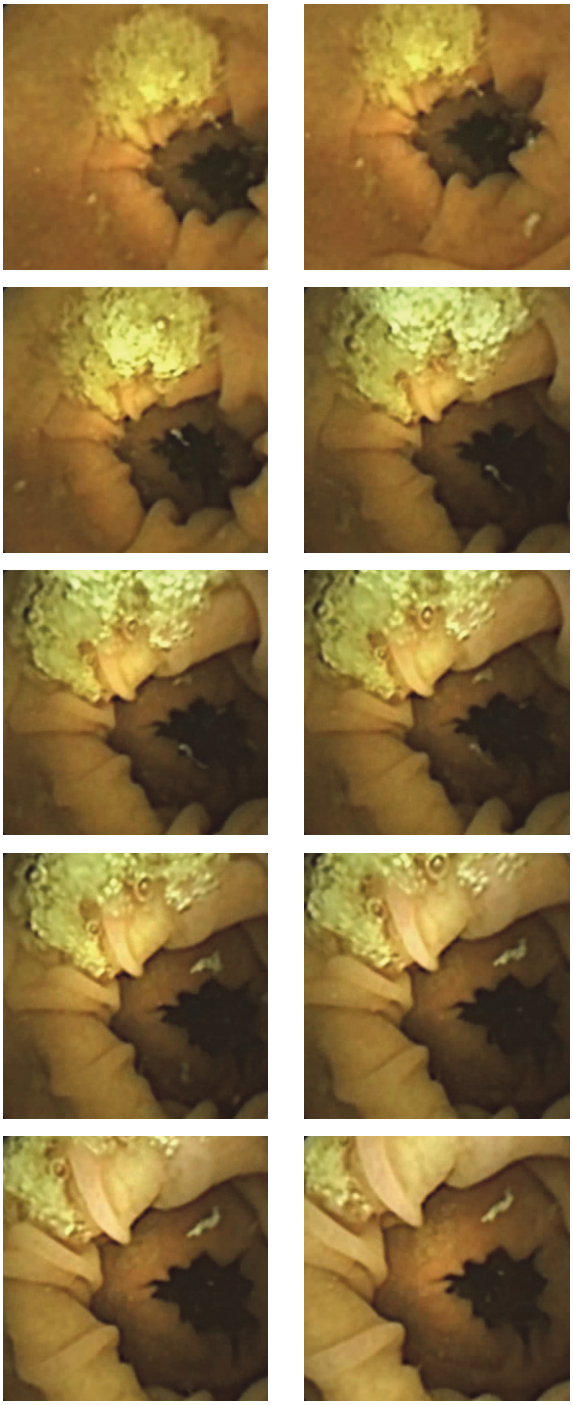
The probability distribution of generating collages from clusters of various cardinalities is illustrated in Fig. 4. It can be noticed that collages of 8-frame clusters are most probable, with a probability of 16%. Overall, only 2.9% of the collages were composed of more than 12 frames, whereas 2% of frames did not have any matching points neither with their previous, nor with their next frame.

IV. Conclusions

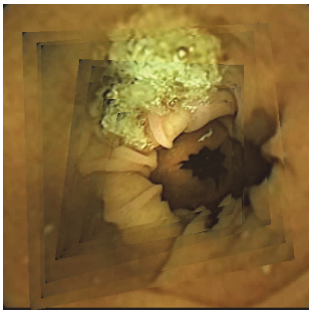
We proposed a visualization approach for WCE videos, based on homography matrix estimation. It generates frame collages concentrating the information of multiple frames without any information loss. The experiments validate that it can be used for WCE video reading time reduction.

The video length reduction obtained as compared with the one reported in [10] can be considered comparable as it possible to reduce the number of images down to less than 14.4% of the original videos, and better than the 25% that can be achieved by the straightforward approach of simultaneously projecting four frames in quad-view [6].

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



(a)



(b)

Fig. 3. Automatically generated collage of a 10-frame cluster. (a) Input video frames. (b) Output visualization added in the new video.

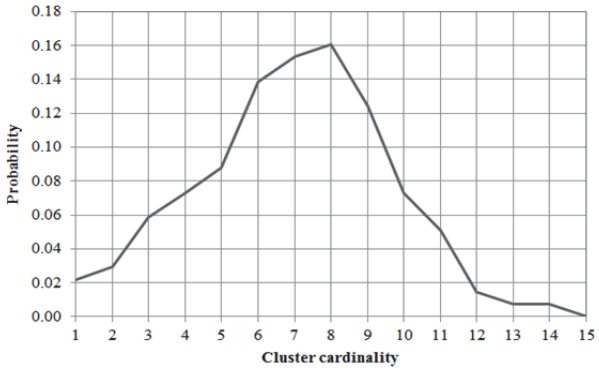


Fig. 4. Probability distribution of cluster cardinalities.

REFERENCES

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

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

[3] Y. Zheng, L. Hawkins, J. Wolff, O. Goloubeva, and E. Goldberg, "Detection of lesions during capsule endoscopy: physician performance is disappointing.," *Am J Gastroenterol*, vol. 107, pp. 554–560, Apr 2012

[4] D.K. Iakovidis, S. Tsevas, D. Maroulis, A. Polydorou, "Unsupervised Summarisation of Capsule Endoscopy Video," in *Proc. IEEE Int. Conf. Intelligent Systems*, Varna, Bulgaria, 2008, pp. 3-15-3-20

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

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

[7] Günther U, Daum S, Zeitz M, et al. Capsule endoscopy: comparison of two different reading modes. Int J Colorectal Dis 2011

[8] A. Smirnidis, A. Koulaouzidis, S. Douglas and J. Plevris, "PTU-143 Quickview in capsule endoscopy: is it enough?," *Gut*, vol. 61, no. Suppl 2, pp. A244-A244, 2012.

[9] V. Hai, T. Echigo, R. Sagawa, Y. Keiko, M. Shiba et al, "Controlling the display of capsule endoscopy video for diagnostic assistance," *IEICE Trans. Inf. Systems*, vol. 92, no. 3, pp. 512-528, 2009.

[10] X. Chu, C. K. Poh, L. Li, K. L. Chan, S. Yan and W. Shen et al, "Epitomized summarization of wireless capsule endoscopic videos for efficient visualization," in *Proc. MICCAI*, pp. 522-529, 2010.

[11] M. Drozdzal, S. Seguí, J. Vitrià, C. Malagelada, F. Azpiroz and P. Radeva, "Adaptable image cuts for motility inspection using WCE," Computerized Medical Imaging and Graphics, 2012.

[12] L. Wei, L. Huang, L. Pan and L. Yu, "The retinal image mosaic based on invariant feature and hierarchial transformation models," Image and Signal Processing, 2009. CISP'09, pp. 1-5, 2009.

[13] T. D. Soper, M. P. Porter and E. J. Seibel, "Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance," IEEE Transactions on Biomedical Engineering, vol. 59, no. 6, pp. 1670-1680, 2012.

[14] S. Yi, J. Xie, P. Mui and J. A. Leighton, "Achieving real-time capsule endoscopy (CE) video visualization through panoramic imaging," IS&T/SPIE Electronic Imaging, pp. 86560I-86560I, 2013.

[15] O. Faugeras, "Three-Dimensional Computer Vision: A Geometric Approach", MIT Press, 1996.

[16] 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.

[17] D. Monnin, E. Bieber, G. Schmitt and A. Schneider, "An effective rigidity constraint for improving RANsAC in homography estimation," Advanced Concepts for Intelligent Vision Systems, pp. 203-214, 2010.

[18] W. Wang and M. K. Ng, "A Variational Method for Multiple-Image Blending," IEEE Trans. Im. Proc., vol. 21, no. 4, pp. 1809-1822, 2012.

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