Xiongbiao Luo · Tobias Reichl
Austin Reiter · Gian-Luca Mariottini (Eds.)

LNCS 9515

Computer-Assisted and Robotic Endoscopy

Second International Workshop, CARE 2015
Held in Conjunction with MICCAI 2015
Munich, Germany, October 5, 2015, Revised Selected Papers

Springer

EXTRAS ONLINE

Lecture Notes in Computer Science **9515**

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

More information about this series at http://www.springer.com/series/7412

Xiongbiao Luo · Tobias Reichl
Austin Reiter · Gian-Luca Mariottini (Eds.)

Computer-Assisted and Robotic Endoscopy

Second International Workshop, CARE 2015
Held in Conjunction with MICCAI 2015
Munich, Germany, October 5, 2015
Revised Selected Papers

Springer

Editors
Xiongbiao Luo
Xiamen University
Fujian
China

Austin Reiter
Johns Hopkins University
Baltimore, MD
USA

Tobias Reichl
KUKA Robotics
Augsburg
Germany

Gian-Luca Mariottini
University of Texas at Arlington
Arlington, TX
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-29964-8 ISBN 978-3-319-29965-5 (eBook)
DOI 10.1007/978-3-319-29965-5

Library of Congress Control Number: 2016931292

LNCS Sublibrary: SL6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

The original version of the cover was revised. The logo was inserted by mistake. The Erratum to the book frontmatter is available at *10.1007/978-3-319-29965-5_16*

Preface

Welcome to the proceedings of the second edition of the International Workshop on Computer-Assisted and Robotic Endoscopy (CARE) that was held in conjunction with MICCAI on October 5, 2015, in Munich, Germany.

CARE aims at bringing together researchers, clinicians, and medical companies to advance the scientific research in the field of computer-assisted and robotic endoscopy and thereby improve current endoscopic medical interventions. The next generation of CARE systems promises to integrate multimodal information relative to patient anatomy, the control status of medical endoscopes and surgical tools, and the actions of surgical staffs to guide endoscopic interventions. To this end, technical advances should be introduced in many areas, such as computer vision, graphics, robotics, medical imaging, external tracking systems, medical device controls systems, information processing techniques, as well as endoscopy planning and simulation.

The technical program of this workshop comprised original and high-quality papers that, together with this year's keynote speakers, explored the most recent scientific, technological, and translational advancements and challenges in the next generation of CARE systems. We selected 15 high-quality papers from 12 countries this year. All the selected papers were revised and resubmitted by the authors in accordance with the reviewers' comments and the volume editors' suggestions.

It was also our great honor and pleasure to welcome the keynote speakers, Prof. Guang-Zhong Yang (Imperial College London, UK), Prof. Emanuele Trucco (University of Dundee, UK), Prof. Robert J. Webster III (Vanderbilt University, USA), and Dr. Mahdi Azizian (Intuitive Surgical Inc., USA), who gave fantastic talks on recent advances on robotic endoscopic interventions representing both the academic and industrial fields.

The CARE 2015 Organizing Committee would like to sincerely thank the Advisory Committee members for their suggestions and assistance in selecting the best paper and all the Program Committee members for their great efforts in reviewing all the submissions. We also extend our thanks and appreciation to KUKA Robotics, Germany, for sponsoring the best paper award and Springer for accepting to publish the CARE proceedings in the *Lecture Notes in Computer Science* series. We warmly thank all the authors, researchers, and attendees at CARE 2015 for their scientific contribution, enthusiasm, and support. We look forward to all the continuing support and participation in our next CARE event that will be held in conjunction with MICCAI 2016 in Istanbul, Turkey.

January 2016

Xiongbiao Luo
Tobias Reichl
Reiter Austin
Gian-Luca Mariottini

Organization

Organizing Committee

Xiongbiao Luo	Xiamen University, China
Tobias Reichl	KUKA Robotics, Germany
Austin Reiter	Johns Hopkins University, USA
Gian-Luca Mariottini	The University of Texas at Arlington, USA

Advisory Committee

Stephen Aylward	Kitware, Inc., USA
Kevin Cleary	Children's Research Institute, USA
Randy Ellis	Queen's University, Canada
Hubertus Feussner	Klinikum rechts der Isar, Germany
Alejandro Frangi	University of Sheffield, UK
Robert Howe	Harvard University, USA
Pierre Jannin	Université de Rennes 1, France
Leo Joskowicz	The Hebrew University of Jerusalem, Israel
Thomas Lango	SINTEF, Norway
Andreas Maier	Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Kensaku Mori	Nagoya University, Japan
Nassir Navab	Technische Universität München, Germany; JHU, USA
Terry Peters	Western University, Canada
Josien Pluim	Eindhoven University of Technology, The Netherlands
Daniel Rueckert	Imperial College London, UK
Tim Salcudean	The University of British Columbia, Canada
Dinggang Shen	University of North Carolina at Chapel Hill, USA
Milan Sonka	The University of Iowa, USA
Russel Taylor	Johns Hopkins University, USA
Emanuele Trucco	Dundee University, UK
Pietro Valdastri	Vanderbilt University, USA
Stephen Wong	Methodist Hospital-Weill Cornell Medical College, USA
Guang-Zhong Yang	Imperial College London, UK

Program Committee

Jorge Bernal	The Autonomous University of Barcelona, Spain
Duane Cornish	Johns Hopkins University, USA
Bernhard Furst	Johns Hopkins University, USA
Stamatia Giannarou	Imperial College London, UK

Uditha Jayarathne	Western University, Canada
Rahul Khare	Blue Belt Technologies, USA
Takayuki Kitasaka	Aichi Institute of Technology, Japan
David Kwartowitz	Clemson University, USA
Feng Li	Sunnybrook Health Sciences Centre, Canada
Paul Loschak	Harvard University, USA
Lena Mair-Hein	German Cancer Research Center, Germany
Siyamalan Manivannan	University of Dundee, UK
Jonathan McLeod	Western University, Canada
Daniel Mirota	Intel Corporation, USA
John Moore	Western University, Canada
Peter Mountney	Siemens Corporation, USA
Masahiro Oda	Nagoya University, Japan
Yoshito Otake	Nara Institute of Science and Technology, Japan
Wu Qiu	Western University, Canada
Holger Roth	National Institute of Health, USA
Amit Shah	Technische Universität München, Germany
Chaoyang Shi	University of Toronto, Canada
Timothy Soper	Intuitive Surgical, USA
Raphael Sznitman	University of Bern, Switzerland
Nanda van der Stap	University of Twente, The Netherlands
Huafeng Wang	BeiHang University, China
Shijun Wang	Alibaba Group, USA
Sebastian Wirkert	German Cancer Research Center, Germany
Guorong Wu	University of North Carolina at Chapel Hill, USA
Wei Xiong	Institute for Infocomm Research, Singapore
Menglong Ye	Imperial College London, UK
Guoyan Zheng	University of Bern, Switzerland
Siyang Zuo	Imperial College London, UK

Contents

Impact of Lossy Image Compression on CAD Support Systems for Colonoscopy

Peter Elmer[1], Michael Häfner[2], Toru Tamaki[3], Shinji Tanaka[4], Rene Thaler[1], Andreas Uhl[1(✉)], and Shigeto Yoshida[5]

[1] Department of Computer Sciences, University of Salzburg, Salzburg, Austria
uhl@cosy.sbg.ac.at
[2] St. Elisabeth Hospital, Vienna, Austria
[3] Department of Information Engineering, Hiroshima University, Higashihiroshima, Japan
[4] Hiroshima University Hospital, Hiroshima, Japan
[5] Hiroshima General Hospital of West Japan Railway Company, Hiroshima, Japan

Abstract. In a large experimental study, the impact of lossy image compression standards on CAD support systems based on texure classification is assessed using colonoscopic imagery as an example. Results clearly indicate that (1) it is important to compress both training and evaluation data involved in the classification process, (2) there is a big difference if initial data is precompressed or uncompressed, and (3) in the latter case significant improvements in terms of classification accuracy may be achieved, even and especially in case of high compression ratios. Moreover it is found that compression efficiency in terms of image quality metrics and/or human perception is not correlated with the impact compression has on texture classification accuracy.

1 Introduction

The amount of medical image data produced on a daily basis is tremendous and the necessity to store (or transfer) these data in a compact manner is obvious. However, medical image compression is constrained by the fact that most radiologists are not willing to base a diagnosis on an image that has been compressed in a lossy way. This is partially due to legal reasons (depending on the corresponding country's laws) and partially due to the fear of misdiagnosis because of lost data in the compression procedure [19]. Therefore, in many scenarios, only lossless techniques are accepted, which limits the amount of compression to a factor of about 3 (in contrast to factors of 100 or more achievable in lossy schemes). On the other hand, many medical professionals are convinced that the future of health care will be shaped by technologies such as telemedicine (see [14] for a discussion of compression in tele-endoscopy). Applications of this type demand lower data rates as are achievable with lossless schemes [3]. This immideately shows the need for efficient and widely accepted techniques for medical image compression. At present state, JPEG2000 is included in the DICOM standard, thus, represents the most accepted solution in this area.

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 1–11, 2016.
DOI: 10.1007/978-3-319-29965-5_1

Image compression algorithms are classically either assessed with respect to human perception (using mean option score - MOS - or similar) or with respect to rate-distortion criteria (e.g. employing distortion measures like PSNR or SSIM). For medical image data, this strategy has been followed as well: E.g., [1,13] compare different compression schemes for MRI based on image quality measures while [17] determine perceptual quality of laprascopic video after compression based on medical experts scores. However, an assessment w.r.t. the impact on the actual diagnostic aim of the acquired imagery is more beneficial and usually drastically increases acceptance of such techniques among medical personnel [3]. For example, [11] investigates the effect of image compression and scaling on automated scoring of immunohistochemical stainings and segmentation of tumor epithelium, while [16] studies effects of MR image compression in tissue classification quality.

In this paper, we study the effect of lossy image compression techniques on texture classification schemes as used in CAD support systems. In particular, as an example, we focus on computer assisted tumor staging in colonoscopy. The impact of compression on image classification has been well studied in areas like remote sensing [5,12,20] or face recognition [4]. Medical imagery and especially endoscopic data is widely inexplored in this context. We aim to close this gap with the present paper.

For some applications in pattern recognition, it has been found that optimisation of image compression with respect to either human perception or rate-distortion criteria is not necessarily the optimal solution. For example, in [2] the JPEG Q-table is tuned for application in the pattern recognition context by emphasising middle and high frequencies and discarding low frequencies (the standard JPEG Q-table is rotated by 180°) leading to better results as the classical, perceptually optimised Q-table. Also, JPEG Q-table optimisation has been considered in biometrics, e.g. in face recognition [8] and iris recognition [10] (both approaches led to improved recognition results as compared to the original Q-table). A further example is optimisation of JPEG 2000 Part 2 wavelet packet decomposition structures with respect to optimising iris recognition accuracy which provides better results compared to rate-distortion optimised wavelet packet structures [7].

These observations raise the general question if compression algorithms exhibiting better rate-distortion performance are also better in a recognition or general pattern recognition context. The answer is “no” obviously, at least for specific applications as already seen before – as another example, it has been found recently that although significantly inferior as compared to more recent standards in terms of image quality measures, JPEG turns out to support iris segmentation much better compared to its more recent competitors JPEG 2000 and JPEG XR [15].

In Sect. 2, we briefly review the most important lossy still image compression schemes as standardised by ISO. Section 3 presents a large scale experimental study on compressing imagery of two different colonoscopic databases and the impact when using these data in automated mucosa texture classification aiming towards tumor staging. Conclusions are drawn in Sect. 4.

2 Lossy Image Compression Standards

We consider three different lossy image compression algorithms for increasing compression rates up to 100 using the respective default configurations unless stated otherwise:
JPEG (JPG): The well-known (ISO/IEC IS 10918-1) DCT-based image compression method. By adjusting the divisors in the quantization phase, different compression ratios can be achieved. We adjust the quality parameter iteratively to achieve a file size closest to the desired compression rate.
JPEG 2000 (J2K): The wavelet-based image compression standard (ISO/IEC IS 15444-1) can operate at higher compression ratios. J2K is also a part of the DICOM standard where it replaced lossless JPEG compression. Results typically do not generate block-based artefacts as the original DCT-based JPG standard. J2K facilitates explicit rate control, i.e. target bitrates are met with high accuracy.
JPEG-XR (JXR): This compression standard based on Microsoft's HD Photo is known to produce higher quality than JPEG, but provides faster conversion than JPEG 2000. In the default configuration the Photo Overlay/Overlap Transformation is only applied to high pass coefficients prior to the Photo Core Transformation (ISO/IEC IS 29199-2). We adjust quantization levels iteratively to achieve a target bitrate closest to the desired one.

In terms of image quality measures and human perception, it is commonly agreed that JPG is clearly the weakest algorithm, especially for high compression ratios, while J2K and JXR perform quite close on many datasets with slight but consistent advantages for J2K.

3 Experiments

3.1 Experimental Settings

In colonic tumor staging we often distinguish between a 2-classes case and a 3-classes case. In the former we simply distinguish between normal mucosa (non-neoplastic) and mucosal changes which need a medical intervention (neoplastic). A more fine-grained classification was proposed in [9]. In this classification scheme the images are divided into three classes: normal lesions, non-invasive lesions, and invasive lesions. This classification scheme is of particular clinical importance since normal mucosa needs not to be removed, non-invasive lesions must be removed endoscopically, and invasive lesions must not be removed endoscopically.

The High-magnification Colonic Polyp Database (HM-DB) [6] is based on 327 endoscopic color images and is provided by the Department of Gastroenterology and Hepatology (Medical University of Vienna) using a zoom-colonoscope (Olympus Evis Exera CF-Q160ZI/L) with a magnification factor of 150. In order to acquire the images, 40 patients underwent colonoscopy. To obtain a larger set of images, subimages (regions of interest) have been extracted manually from the original images by a medical expert with a size of 256×256 pixels. This resulted in an extended image set containing 716 images in total.

Lesions found during colonoscopy have been examined after application of dye-spraying with indigocarmine, as routinely performed in colonoscopy. Biopsies or mucosal resection have been performed in order to get a histopathological diagnosis which serves as a ground truth in our experiments.

The NBI database (NBI-DB) is an endoscopic image database consisting of 908 patches extracted from frames of zoom- endoscopic (CF-H260AZ/I, Olympus Optical Co) videos using the NBI technology and is provided by the Hiroshima University and the Hiroshima University Hospital [18]. The image patches are rectangular and have varying sizes – for providing comparable experimental data, 256×256 pixels squares have been extracted from the center of the original images. The database consists of 359 images of type A, 462 images of type B and 87 images of type C3, all taken from different patients. From this dataset, due to size restrictions, only a certain share of images allows the extraction of appropriately sized patches from the center – for the others (about 40 %), we mirror the data across the images' edges to get comparable results to the HM-DB. Image labels were provided by at least two medical doctors and endoscopists who are experienced in colorectal cancer diagnosis and familiar with pit pattern analysis and NBI classifications (the two class case corresponds to distinguishing type A from types B and C3).

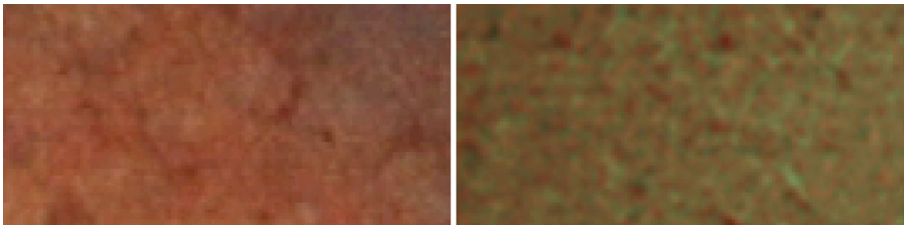


Fig. 1. Zoom into example images of HM-DB and NBI-DB, respectively.

There is one important difference between the two datasets. As can be seen from Fig. 1, HM-DB original images exhibit blocking artifacts from prior DCT-based compression, while NBI-DB images come without any visible degradations from prior compression. As we shall see, this makes a big difference.

Band-pass type Fourier descriptors have been used successfully in classification of the HM-DB imagery [6]. As we are not interested in maximising classification rates, we do not perform fusion of several different band-pass descriptors across frequency bands but look into the discriminative power of individual frequency bands with small frequency support. In particular, we consider 128 concentric rings in Fourier space with increasing the frequency by 1 in each step following the main coordinate axes – for each ring, values in the power spectrum closest to the ring under consideration are used to compute mean and variance for each ring. These two features, computed for each of the three RGB colour bands, comprise the final feature vector used in a k-nn classification employing

Euclidian distance, where $1 \leq k \leq 19$. While classically magnitude is considered only [6], here we also use phase information in the same manner and we also fuse magnitude and phase information after feature value normalisation (leading to a twelve-component feature vector). Please note that using these feature vectors, we gain insight of the sensitivity of the information contained in different frequency bands w.r.t. compression, thus we are able to derive statements about the compression sensitivity of other features as well (as long as these features can be localised in frequency space in a sensible way).

For determining the overall classification rate (which is finally considered to rate the impact of compression for each frequency band), a leave-one-patient-out (LOPO) cross validation protocol is used (identical to leave-one-out cross validation (LOOCV) for NBI-DB due to the structure of this dataset). When conducting classification on compressed data, either both images compared can be compressed (scenario CC) or only one of the two images involved (scenario UC). The one-image compressed case corresponds to either having the image to be classified in compressed form (e.g. due to a prior transmission in a tele-endoscopic set-up) or to having the training data in compressed form (e.g. due to excessive storage requirements).

3.2 Experimental Results

In the following figures, the x-axis represents the 128 distinct frequency bands considered, while the y-axis shows the achieved overall classification accuracy for each of the 128 bands individually. We do not further report on other settings for k as $k = 19$ in the classifier, since the overall trend is that the best results are obtained for this configuration.

Results obtained from the HM-DB in the 3 classes case somehow correspond to the expectations. Figure 2 displays the results for the CC scenario for using magnitude features. We see that overall, magnitude features deliver consistently better results with the uncrompressed data (denoted as "png" in the plots) as data compressed with compression ratio 100. Phase features exhibit virtually identical behaviour.

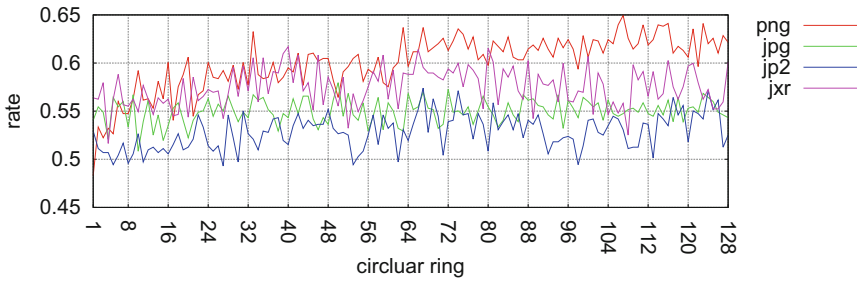


Fig. 2. HM-DB, magnitude, 3classes, compression ratio 100, scenario CC

However, the ranking of the compression algorithms is surprising: JPG, clearly inferior in terms of image quality measures especially for high compression ratios according to common knowledge, is competitive to J2K, while JXR is clearly superior to both other standards, for low and medium frequencies even comparable to the uncompressed case.

Figures 3 and 4 compare the results for the UC scenario. We immediately notice a significant difference to the CC scenario. While for the magnitude features (Fig. 3) JXR and JPG behave similarly to the CC scenario, J2K compression results are significantly detoriated. For the phase features (Fig. 4) only JXR can compete with the uncompressed case for very low frequencies, in all other cases a significant decrease of classification accuracy is observed.

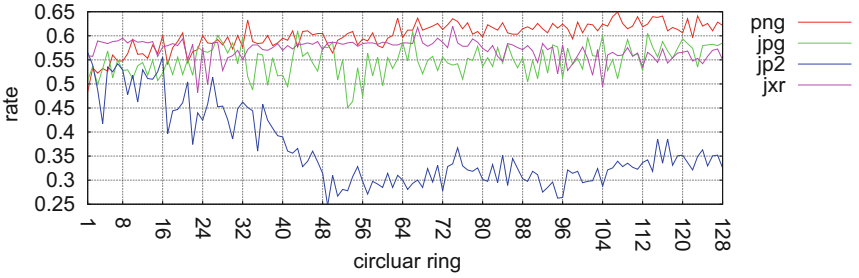


Fig. 3. HM-DB, magnitude, 3classes, compression ratio 100, scenario UC

The results for the 3 classes case as discussed so far are similar down to a compression ratio of 10, however, the differences among the compression schemes tend to get smaller for lower compression ratios and JXR takes the lead. For the 2 classes case we do not show graphical results here. Besides delivering higher classification accuracy overall of course (up to 75 %), the impact of compression is not very high for both scenarios (CC and UC) and leads even to slightly better classification rates as compared to uncompressed data, in most cases for JPG and JXR. This is a surprising effect which leads us to the results for the second dataset.

In the following, results corresponding to the NBI-DB are discussed. Figures 5 and 6 compare the CC and UC scenarios for compression ratio 100 in the 2 classes case (magnitude feature), respectively. Interestingly we notice for CC, while J2K is comparable to the uncompressed case, JXR as well as JPG improve classification accuracy, the latter significantly so (top accuracy is improved by more than 10 % !).

This surprising effect cannot be observed for the UC scenario, where overall, the uncompressed case is top performing, closely followed by JXR, J2K, and JPG delivers the worst accuracy (contrasting to the CC scenario this corresponds to the image quality measure results). Please note that for JPG compression, the difference in terms of classification accuracy between the CC and UC scenarios can be up to 30 % depending on the frequency band considered.

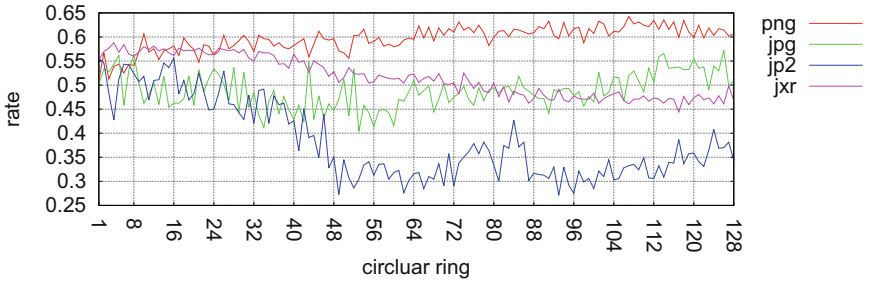


Fig. 4. HM-DB, phase, 3classes, compression ratio 100, scenario UC

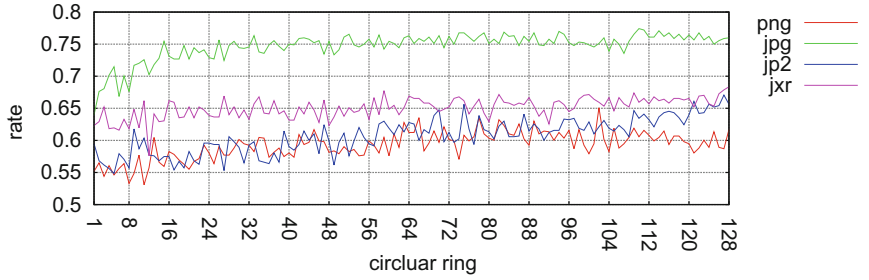


Fig. 5. NBI-DB, magnitude, 2classes, compression ratio 100, scenario CC

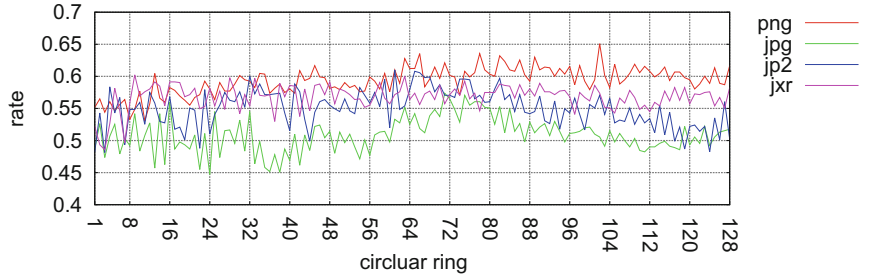


Fig. 6. NBI-DB, magnitude, 2classes, compression ratio 100, scenario UC

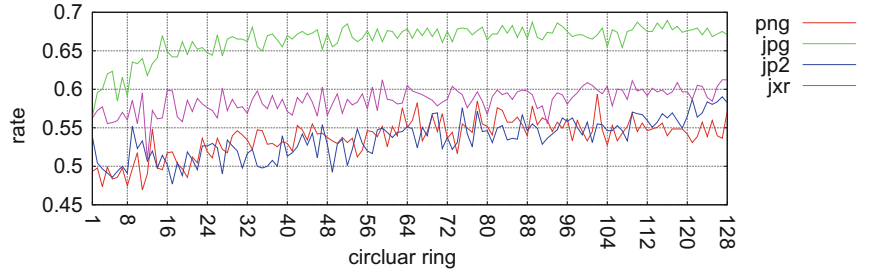


Fig. 7. NBI-DB, magnitude, 3classes, compression ratio 100, scenario CC

Figure 7 confirms the same behaviour for the 3 classes case and the magnitude feature. For the UC scenario, we observe the same behaviour as for the two class case (not shown), except that all three compression schemes deliver almost the same results (inferior to the uncompressed case). It should be noted that we observe exactly the same phenomenon for the phase feature and the magnitude-phase feature fusion (not shown as well).

Figures 8 and 9 investigate the observed results in more details for the most significant case of JPG compression (using fused magnitude and phase features), considering compression ratios of 1 (only file conversion), 10, 33, 50, and 100. Figure 8 shows the result for the CC scenario, where we see that classification accurracy consistently increases for increasing compression ratio, which is a very surprising result.

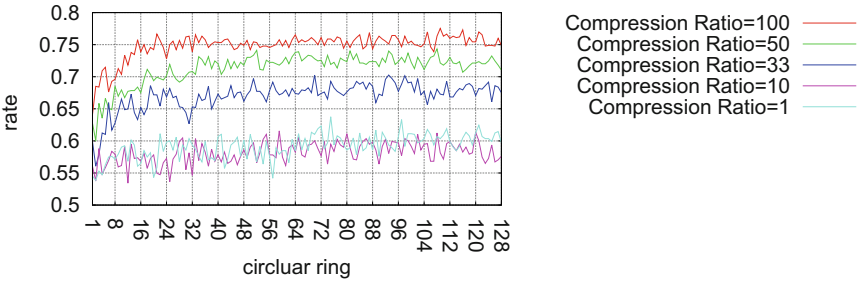


Fig. 8. NBI-DB, fusion, 2classes, JPG, scenario CC

For the UC scenario (see Fig. 9) we observe the expected behaviour with decreasing classification results for increasing compression ratio, seen almost across the entire range of considered frequency bands.

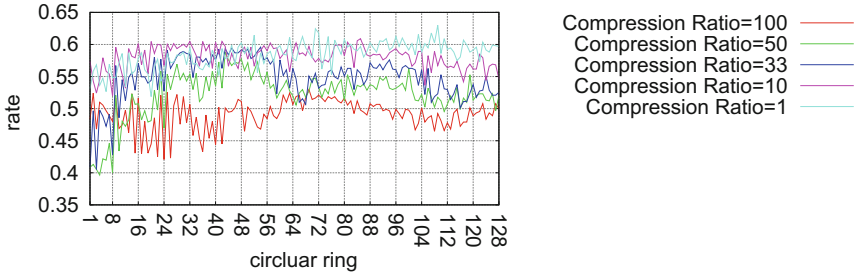


Fig. 9. NBI-DB, fusion, 2classes, JPG, scneario UC

4 Discussion and Conclusion

From the results obtained it is quite obvious that no matter which dataset is being considered, it is better to have all involved data being compressed (scenario CC). However, apart from that, we observe quite different results for the two datasets, and it is plausible that the precompression of HM-DB is the reason for this difference.

For the HM-DB, the major difference is seen between the 2 classes and 3 classes setting, but not between the CC and UC senarios. The reason is that we have compression artifacts already present in this dataset (see Fig. 1, corresponding to DCT-block structures). Recompressing these data generates double compression effects which propagate into the features vectors, thus we have JPG artifacts present in both images involved in compression, no matter if we operate in the CC or UC scenario. While for the 2 classes case the discriminative power of the features is sufficient also under compression to even slightly improve the uncompressed case, this is not true for the 3 classes case where the difference of the image features needs to be exploited more thoroughly, thus we see result degradation under compression.

For the NBI-DB, the situation is fairly different. Image data is virtually uncompressed, thus applying compression introduces artifacts, which of course contribute to the features extracted. Thus, in the UC scenario, we compare feature vectors computed from undistorted data to features vectors containing compression artifacts. Obviously, having artifacts in both images involved is the better solution.

The remaining question is why JPG delivers the best results in the CC scenario for the NBI-DB and why results do get better for increasing compression ratio. It is clear that JPG introduces the most significant artifacts (especially at high compression ratios), both in terms of visual quality and in terms of image quality measures. However, based on the results observed, it is likely that these strong artifacts are specific to the image class and this difference can be expoited in feature extraction and subsequent classification (i.e. JPG compression acts as a specific type of additional pre-feature extraction). The stronger those artifacts are, the better the classifier is able to exploit them, which explains the improving classification for increasing compression ratio. Please note that recently, a somewhat related result has been demonstrated in iris biometrics: JPG turns out to support iris segmentation much better compared to its more recent competitors J2K and JXR [15]. The explanation given relates to the stronger artifacts as produced by JPG along iris boundaries, which actually aid in the segmentation process.

Although experiments are being restricted to colonoscopic imagery, there is strong evidence that the results do carry over to any texture-based CAD support system based on classification. Future work will consider the actual impact on state-of-the-art feature descriptors, while the present study is restricted to the general discriminative power of frequency-band based descriptors.

References

1. Belloulata, K., Baskurt, A., Benoit-Cattin, H., Prost, R.: Fractal coding of medical images. In: Kim, Y. (ed.) Medical Imaging 1996: Image Display. SPIE Proceedings, vol. 2707, pp. 598–609. SPIE, Newport Beach (1996)
2. Chen, M., Zhang, S., Karim, M.: Modification of standard image compression methods for correlation-based pattern recognition. Opt. Eng. **43**(8), 1723–1730 (2004)
3. Cosman, P.C., et al.: Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy. Proc. IEEE **82**(6), 919–932 (1994)
4. Delac, K., et al.: Image compression in face recognition - a literature survey. In: Delac, K., et al. (ed.) Recent Advances in Face Recognition, pp. 236–250. I-Tech (2008)
5. Garcia-Vichez, F., Munoz-Mari, J., Zortea, M., Blanes, I., Gonzales-Ruiz, V., Camps-Valls, G.: On the impact of lossy compression on hyperspectral image classification and unmixing. IEEE Geosci. Remote Sens. Lett. **8**(2), 253–257 (2011). doi:10.1109/LGRS.2010.2062484
6. Häfner, M., et al.: Computer-aided classification of zoom-endoscopical images using fourier filters. IEEE Trans. Inf. Technol. Biomed. **14**(4), 958–970 (2010)
7. Hämmerle-Uhl, J., Karnutsch, M., Uhl, A.: Evolutionary optimisation of JPEG2000 part 2 wavelet packet structures for polar iris image compression. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) CIARP 2013, Part I. LNCS, vol. 8258, pp. 391–398. Springer, Heidelberg (2013)
8. Jeong, G.M., Kim, C., Ahn, H.S., Ahn, B.J.: JPEG quantization table design for face images and its application to face recognition. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E69–A**(11), 2990–2993 (2006)
9. Kato, S., et al.: Magnifying colonoscopy as a non-biopsy technique for differential diagnosis of non-neoplastic and neoplastic lesions. World J. Gastroenterol. **12**(9), 1416–1420 (2006)
10. Konrad, M., Stögner, H., Uhl, A.: Custom design of JPEG quantisation tables for compressing iris polar images to improve recognition accuracy. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1091–1101. Springer, Heidelberg (2009)
11. Konsti, J., et al.: Effect of image compression and scaling on automated scoring of immunohistochemical stainings and segmentation of tumor epithelium. Diagn. Pathol. **7**(29) (2012). doi:10.1186/1746-1596-7-29
12. Lau, W.L., Li, Z.L., Lam, W.K.: Effects of JPEG compression on image classification. Int. J. Remote Sens. **24**(7), 1535–1544 (2003)
13. Panych, L.: Theoretical comparison of Fourier and wavelet encoding in magnetic resonance imaging. IEEE Trans. Med. imaging **15**(2), 141–153 (1997)
14. Rabenstein, T., et al.: Tele-endoscopy: influence of data compression, bandwidth and simulated impairments on the usability of real-time digital video endoscopy transmissions for medical diagnoses. Endoscopy **34**(9), 703–710 (2002)
15. Rathgeb, C., et al.: Effects of severe image compression on iris segmentation performance. In: Proceedings of the IAPR/IEEE International Joint Conference on Biometrics (IJCB 2014) (2014)
16. Santalla, H., et al.: Effects on MR images compression in tissue classification quality. J. Phys. Conf. Ser. **90**(1) (2007)
17. Schoeffmann, K., et al.: Investigation of the impact of compression on the perceptional quality of laparoscopic videos. In: Proceedings of the 27th International Symposium on Computer-Based Medical Systems (CBMS 2014), pp. 153–158 (2014)

18. Tamaki, T., et al.: Computer-aided colorectal tumor classification in nbi endoscopy using local features. Med. Image Anal. **17**(1), 78–100 (2013)
19. Wong, S., et al.: Radiologic image compression - a review. Proc. IEEE **83**(2), 194–219 (1995)
20. Zabala, A., Pons, X.: Effects of lossy compression on remote sensing image classification of forest areas. Int. J. Appl. Earth Obs. Geoinf. **13**(1), 43–51 (2011)

Pointing with a One-Eyed Cursor for Supervised Training in Minimally Invasive Robotic Surgery

Martin Kibsgaard[1] and Martin Kraus[1(✉)]

Department of AD:MT, Aalborg University,
Rendsburggade 14, 9000 Aalborg, Denmark
{kibsgaard,martin}@create.aau.dk
http://www.create.aau.dk/graphics/

Abstract. Pointing in the endoscopic view of a surgical robot is a natural and efficient way for instructors to communicate with trainees in robot-assisted minimally invasive surgery. However, pointing in a stereo-endoscopic view can be limited by problems such as video delay, double vision, arm fatigue, and reachability of the pointer controls. We address these problems by hardware-based overlaying the stereo-endoscopic view with a one-eyed cursor, which can be comfortably controlled by a wireless, gyroscopic air mouse. The proposed system was positively evaluated by five experienced instructors in four full-day training units in robot-assisted minimally invasive surgery on anaesthetised pigs.

Keywords: Minimally invasive surgery · Robot-assisted surgery · Teleoperation · Telesurgery · Telementoring · Training · Robotic endoscopy · Stereoscopic endoscopy · Mixed reality · Augmented reality · Head-up display · One-eyed cursor · Telestration

1 Introduction

Training in robot-assisted minimally invasive surgery is costly [3] but also important in order to achieve the best possible outcomes [5]. In the training on actual robots, pointing and line drawing (so-called "telestration") in the endoscopic view is often useful to support referential verbal communication by instructors (e.g., "look at this," "cut here," etc.) [8]. However, many surgical robots are operated using an immersive interface that blocks the visual communication between instructor and trainee (see Fig. 1). In these cases, one common approach is to overlay the endoscopic video image with the video image of a pointer and/or a line drawing and present the resulting video image to the trainee, who operates the robot, as well as the instructor, who controls the pointer [8]. One advantage of this solution is that the instructor or expert advisor (in general called "mentor") does not have to be physically present but can be located at a large distance (so-called "telementoring") [8,11].

In this work, however, we only consider the case of supervised training where the instructor is in the same room as the trainee who operates a da Vinci S

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 12–21, 2016.
DOI: 10.1007/978-3-319-29965-5_2

Fig. 1. Setup of a da Vinci S HD Surgical System. Left: surgeon (in our case the trainee) operating the console. Center: patient cart and assistant (in our case the instructor) using the telestration feature of a touch screen. Right: vision cart. Copyright 2015 Intuitive Surgical, Inc.

HD Surgical System since this is the situation in the training courses at Aalborg University Hospital that are offered by our collaboration partner MIUC (Minimal Invasive Udviklings Center). The instructors are experienced surgeons and an experienced surgical assistant, who is often performing additional tasks (e.g., operating a suction device) while instructing the trainees.

The da Vinci S HD Surgical System offers the possibility to draw lines on a monoscopic touch screen and to overlay the stereo-endoscopic video image with a stereoscopic version of the line drawing [8]. This stereoscopic image includes an automatically added stereoscopic effect such that the drawing appears — to the trainee operating the robot — on a plane in front of the operating field. While this telestration feature works without affecting the resolution, frame rate or delay of the stereo-endoscopic image, the exact position of the drawn lines can appear ambiguous since they do not appear at the same depth as the tissue that the instructor points at. Specifically, the overlaid line drawing appears at different positions in the left and right image of the stereoscopic image and, in general, both positions are different from the position that the instructor touched on the monoscopic screen.

As reported by our collaborators and well-known in the literature [10,12,14], this ambiguity can make exact pointing very difficult. Furthermore, it is difficult for the surgical assistant to reach the touch screen while operating, for example, a suction device in the current setup of the training room. For these reasons, our collaborators usually do not use the telestration feature of the da Vinci S HD Surgical System in their courses. Another potential problem with a touch screen

at eye's height is arm fatigue [6]; however, this issue has not been mentioned by our collaborators.

In order to support exact pointing, we implemented a one-eyed cursor [14], which is controlled by a wireless gyroscopic air mouse, which can be held at a comfortable height to avoid arm fatigue [6]. One-eyed line drawings are supported by pressing a button of this air mouse. To overlay the stereo-endoscopic image with the image of the pointer and/or line drawing at the original resolution and frame rate without noticeable delay, we employ a recently proposed framework for hardware-based overlaying [7].

We are still adjusting details of the system based on the observed usage and feedback by instructors and trainees. So far, slightly different prototypes of our system were used and positively evaluated by five experienced instructors (including one experienced surgical assistant) in a total of four full-day training units in robot-assisted minimally invasive surgery on anaesthetised pigs.

The first main contribution of our work is to present the design and implementation of a one-eyed cursor for the da Vinci S HD Surgical System, which is comfortably controlled by a wireless, gyroscopic air mouse and does not affect the resolution, frame rate or latency of the stereo-endoscopic view; see Sect. 3. The second main contribution is the successful evaluation of a developing prototype of the proposed system in an operational environment, i.e., in actual training courses in robot-assisted minimally invasive surgery at Aalborg University Hospital; see Sect. 4. Before discussing these contributions, Sect. 2 reviews previous work.

2 Previous Work

Pointing at objects in stereoscopic images is basically a two-dimensional task, but it is usually considered a special case of pointing in three dimensions [10,12,14]. There appears to be a wide consensus that displaying a stereoscopic cursor at a different depth than the depth of the object that the cursor is pointing at should be avoided in order to avoid cursor diplopia (double vision). Instead, the cursor either should be displayed at the same stereo depth as the object or the cursor should only be displayed to one eye as first suggested by Ware and Lowther [14]. Schemali and Eisemann [10] observed better user performance with the first option and attributed this to the discomfort that a one-eyed cursor can cause (due to binocular rivalry). On the other hand, Teather and Stuerzlinger [12] observed — for certain pointing techniques — better user performance with a one-eyed cursor; in particular for objects far away from the screen depth. They attributed this to the problems of diplopia and accommodation-vergence conflicts, which do not occur with a one-eyed cursor.

In the case of stereo-telestration for robotic surgery, Hasser et al. [4] proposed to mark positions at the same depth as objects in the stereo-endoscopic image by computing a disparity map of the stereo-image. (See also Lamprecht et al. [8] and Zhao et al. [15].) Ali et al. [1] reported results of a user study with a prototype of such a system using a da Vinci surgical robot where three participants ("trainees") had to identify pins that another participant (the "mentor") pointed

at. In comparison to 2D telestration, the trainees required significantly more time and committed non-significantly more errors with the three-dimensional marks. Similarly to the study reported by Teather and Stuerzlinger [12], these results might have been caused by diplopia and/or accommodation-vergence conflicts.

These works show that stereo-telestration at object depth for robotic surgery requires considerably more hardware and more complex software while impairing user performance — even if the software worked perfectly. Therefore, a one-eyed cursor appears to be an interesting and viable option to avoid the problems of stereo-telestration and at the same time retain the advantages of a stereo-endoscopic view.

Overlaying a stereo-endoscopic image with a computer-generated image usually results in a noticeable delay of more than 100 ms (e.g., [13]). Azuma et al. [2] state that delays as small as 10 ms can lead to a significantly worse user performance for certain tasks. This is consistent with results for low-latency direct touch which showed "noticeable improvement continued well below 10 ms" [9]. We assume that any noticeable delay (or reduction in frame rate) would reduce the user acceptance of our system.

An alternative to delaying the stereo-endoscopic image is to show it without delay side-by-side with a delayed image that is overlaid with another image. This approach is supported by the "TilePro" feature of da Vinci S HD Surgical Systems but it reduces the size and resolution of both, the original image and the delayed image with the overlay. Therefore, at least some surgeons appear to turn off this feature whenever possible [13]. Thus, we assume that any noticeable reduction in size or resolution of the stereo-endoscopic image would reduce the user acceptance of our system.

In order to overlay the stereo-endoscopic view with the image of a pointer without noticeable delay nor reduction of frame rate, size, or resolution, we employ a framework that we have recently presented [7]. Our specific usage of this framework is described in Sect. 3.

Hincapié-Ramos et al. [6] proposed a series of guidelines for the design of fatigue-efficient mid-air interfaces. In particular, they concluded that mid-air gestures at the height of the shoulder joint are more tiring than gestures between the height of the shoulder and the waist. Furthermore, they found that a clicking device for selection minimizes fatigue. Therefore, we assume that an air mouse that can be held at any height is more fatigue-efficient than a touch screen at eye's height.

3 Proposed One-Eyed Cursor for Stereo-Endoscopy

To overlay the stereo-endoscopic video image of the da Vinci system with the computer-generated image of a pointer, we have employed our recently proposed system [7]. The core of the system is a desktop computer with two PCIe video cards (Blackmagic Design's DeckLink HD Extreme 2), which is capable of overlaying the two channels of the stereo video image with any computer-generated imagery at full resolution and frame rate with less than 1 ms delay. We have

also included a fail-safe system and a 3D TV and used a wireless, gyroscopic air mouse for user input to avoid arm fatigue [6] and to allow instructors to control our system from most positions in the training room.

To render the one-eyed cursor, the image of a pointer is only displayed to one eye by overlaying only one channel of the stereo video image. By default, the cursor is shown to the right eye, as this is the channel that the monoscopic displays of the da Vinci system default to. In some cases it is useful for the instructor to switch which eye the one-eyed cursor is displayed to, e.g., if the trainee is unable to perceive stereoscopic images or if the trainee is uncomfortable with a one-eyed cursor that is displayed to his or her non-dominant eye [14]. With our system, instructors can switch from one eye to the other by clicking the scroll-button of the air mouse. The console and the 3D TV that we introduced in the setup will then show the cursor to the other eye. In order for the cursor to show up also on the monoscopic displays of the da Vinci system, the instructor (or an assistant) has to change the channel shown on those displays by using the touch screen controls of the da Vinci system.

As described in Sect. 1, the telestration feature of the da Vinci system allows instructors to draw lines that are directly visible in the console. Due to the ambiguous position of the drawings, this feature has been rarely used in the training at Aalborg University Hospital in the past; however, our collaborating instructors are familiar with it and expected our solution to provide the same functionality. In our implementation (see Fig. 2), a green line is drawn from the tip of the pointer when the instructor presses and holds the left mouse button of the air mouse. We chose to use the color green based on observations and feedback from our collaborators, who stated that green is the least frequent color when operating on pigs and humans.

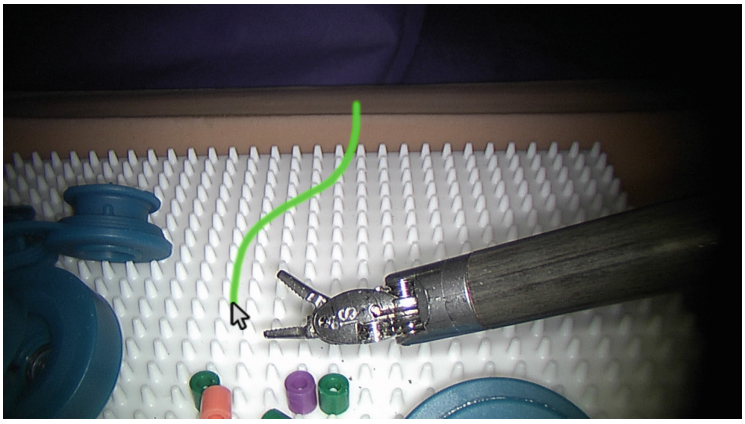


Fig. 2. The cursor and a line drawing overlaid on one of the video channels of the da Vinci S HD system. Note that a monoscopic image cannot convey the appearance of a one-eyed cursor.

The telestration feature of the da Vinci system removes any line drawings when the endoscopic camera is moved. Alternatively, they can be removed by pushing a button on the touch screen. Initially, it was a user requirement that our system behaves similarly to the da Vinci telestrator, i.e., line drawings should stay on the screen until removed. However, when evaluating and regularly using the prototype, it proved more useful to have the drawings automatically disappear a few seconds after the instructor stops drawing. In this way, the instructors can keep the drawings on the screen as long as desired by holding down the drawing button, and there is no need for an additional button to remove the drawings.

To control the cursor, we have tested several wireless, gyroscopic air mice and found two candidates for the scenario at Aalborg University Hospital. We considered ease of use, precision, price, number of buttons, and ability to clutch, which is similar to lifting up a regular mouse to reposition it without moving the mouse cursor. The Gyration Air Mouse Elite was the most precise of the tested air mice, but it is also the most expensive one and still introduces some interaction problems (see Sect. 4). It has a "reverse clutch," i.e., the user needs to hold a button on the bottom of the mouse to move the cursor. This turned out to be an intuitive clutching mechanism and also avoids unintended cursor movements, which would be distracting to the trainees.

The Measy RC9 Air Smart Mouse is a little less precise and offers a "toggle clutch," i.e., the user has to press the same button to activate and to deactivate the control over the cursor. This turned out to be a less intuitive clutching mechanism and requires users to remember to toggle the clutch to avoid unintended cursor movements. The Measy RC9 Air Smart Mouse is significantly cheaper (less than one third of the price of the Gyration Air Mouse Elite), which might be important since the environment in which the air mice are used can be rough on electronic devices as fluids (blood, water, etc.) often get on the instructors' hands when they are assisting. Waterproofing the air mouse by putting it into a plastic bag could protect it, but this would make it more difficult to use.

We have also investigated several other input methods, but based on initial testing they have proved either impractical in our setting (Kinect, LEAP motion) or simply too imprecise (Wii Remote).

4 Evaluation of Prototype in Operational Environment

Before evaluating a prototype of our system in training courses at Aalborg University Hospital, we observed several eight-hours training sessions without our system in order to assess the communication problems between instructors and trainees. The main conclusion from these observations was that the instructors did not use the telestration feature of the da Vinci system. Instead, they usually tried to rely on verbal communication and tended to take over the console of the robot when verbal communication alone proved to be insufficient. This approach was inefficient as considerable time was spent on unsuccessful verbal communication and taking over the console resulted in interruptions of the trainees' operation of the robot and reduced their training time on the robot.

Since our collaborators were not actively using the telestration feature of the da Vinci system in their training courses, we decided against comparing it with our system in these courses since we are trying to interfere as little as possible with the courses. Furthermore, a comparative study between a one-eyed cursor and a stereo cursor at a different depth than the object that it is pointing at is very likely to confirm the previously published result that a one-eyed cursor is preferable in this comparison [12].

Therefore, we chose to evaluate a developing prototype of our system by installing it in the training room at Aalborg University Hospital and observing its impact on the training and the communication during the training. Moreover, we also observed and got feedback on the interaction with the air mouse and the perception of the one-eyed cursor and line drawings.

Our system has been in continuous use during the four most recent full-day training sessions at Aalborg University Hospital. Some of the interaction problems that were revealed in these sessions were fixed between sessions. For example, the instructors sometimes left the cursor in the middle of the screen without using it to point. To solve this problem, we hide the pointer when it has been in the same position for more than two seconds, as was also suggested by the instructors. Another improvement was to decrease the time before line drawings are removed from five seconds to two seconds after the instructors release the drawing button.

Of the two air mice that the instructors evaluated in the training courses, the Gyration air mouse was the preferred one due to its clutching mechanism, which appears to be the single most significant aspect of the usability of the air mice. As mentioned earlier, the reverse clutch helps to avoid unintended cursor movements and the instructors were able to use it immediately — presumably because it is similar to the clutching mechanism of the robot. The toggle clutch of the Measy air mouse proved to be unintuitive and was quickly abandoned by the instructors.

Our system clearly improved the visual communication from instructors to trainees, and with it, the communication overall. This was apparent by much more interactive communication between the instructors and trainees. The instructors used the cursor and line drawings to guide anatomical explorations by the trainees and to give task instructions, e.g., by pointing with the cursor and saying “cut here,” or drawing along a nerve and saying “the nerve is running here,” “grab here,” etc. — activities that previously often resulted in the instructors taking over control of the console.

We neither directly nor indirectly observed any need for switching the cursor to the other eye. None of the trainees reported any issues with perceiving the cursor and we did not observe any apparent miscommunication in relation to pointing. However, the way we implemented the switch caused some confusion as the instructors accidentally switched the channel in which the cursor was shown, causing them to lose sight of the cursor on the monoscopic displays. To avoid this, the instructors suggested that we make it more difficult to accidentally switch (e.g., by requiring to hold the button down for five seconds) and to add

a message after the switch that tells them when the monoscopic displays are showing the channel without the cursor.

Generally, our system was evaluated positively by instructors and trainees. The trainees only had two complaints. First, the white part of the pointer was too bright which caused a slight flickering on the displays of the console. We have consequently changed the color to a bright gray. Second, the cursor was sometimes not hidden when it was left in the middle of the screen. This problem occurred because the Gyration air mouse can also be used as a regular mouse, which caused the mouse to unintentionally move when its proximity sensor was triggered such that our system assumed that the mouse was still in use for pointing. To solve this problem, we have blocked the infrared light that the mouse uses to measure distance.

In summary, the instructors found the prototype of our system very useful. In particular, they found it better and more precise than the telestration feature of the da Vinci S HD Surgical System. While the trainees never experienced the telestration feature of the da Vinci system, they benefitted from the improved visual communication with the instructors as compared to the training without any telestration system.

5 Discussion

The feedback that we received and the observed impact of the prototype of our system on the training in robot-assisted surgery at Aalborg University Hospital is very encouraging as it suggests that a one-eyed cursor that is controlled by a wireless air mouse can in fact improve the communication between instructors and trainees. However, we are fully aware that we are biased observers of our own system and that some of the instructors are similarly biased since they contributed to the development of the system. As most of the trainees have no prior experience with the da Vinci robot, their feedback cannot be used to compare our system with the telestration feature of the da Vinci system. Thus, further user studies are necessary to establish the benefits of our system once its development is completed.

6 Conclusion and Future Work

Based on the concept of a one-eyed cursor, we have developed a new telestration system for pointing and drawing in stereo-endoscopic views of the da Vinci S HD Surgical System. A prototype of the system has been integrated in training courses on robot-assisted minimally invasive surgery at Aalborg University Hospital and was positively evaluated by five experienced instructors.

Future work includes further observations and improvements of the system in regular use. This also includes improvements of the way the system is used by the instructors. For example, we assume that it would be beneficial to some trainees if instructors showed them the one-eyed cursor for each eye such that

each trainee can choose the more comfortable alternative. Once the system and its usage are finalized, formal user studies are necessary to prove its benefits.

Furthermore, it would be very interesting to determine the level of discomfort that a one-eyed cursor can cause, the effect of eye dominance on this discomfort, and the percentage of affected users.

Our observations of the training with the proposed system showed that instructors still take over the robot in some situations, e.g., to demonstrate skills such as knot tying. Some instructors also ask trainees to look up from the robot, e.g., in order to communicate the best orientation of a needle with hand gestures. These situations could be addressed by overlaying the stereo-endoscopic view with a simulation of virtual robotic instruments that are controlled by instructors and displayed to trainees while they operate the console. Whether there is any advantage in displaying these virtual instruments to one eye only, is another open question.

Acknowledgments. The authors would like to thank the participating instructors and trainees of the relevant MIUC courses at Aalborg University Hospital; in particular, chief surgeon Johan Poulsen and registered nurse first assistant in robotic surgery Jane Petersson.

References

1. Ali, M.R., Loggins, J.P., Fuller, W.D., Miller, B.E., Hasser, C.J., Yellowlees, P., Vidovszky, T.J., Rasmussen, J.J., Pierce, J.: 3-D telestration: a teaching tool for robotic surgery. J. Laparoendosc. Adv. Surg. Tech. **18**(1), 107–112 (2008)
2. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. IEEE Comput. Graph. Appl. **21**(6), 34–47 (2001). http://dx.org/10.1109/38.963459
3. Buchs, N., Pugin, F., Volont, F., Morel, P.: Learning tools andsimulation in robotic surgery: state of the art. World J. Surg. **37**(12), 2812–2819 (2013). http://dx.org/10.1007/s00268-013-2065-y
4. Hasser, C.J., Larkin, D.Q., Miller, B., Zhang, G.G., Nowlin, W.C.: Medical robotic system providing three-dimensional telestration. US Patent 2007/0167702 A1 (2007)
5. Ontario, H.Q., Secretariat, M.A.: Robotic-assisted minimally invasive surgery for gynecologic and urologic oncology: an evidence-based analysis. Ont. Health Technol. Assess. Ser. 2010 **10**, 1–118 (2010). Health Quality Ontario
6. Hincapié-Ramos, J.D., Guo, X., Moghadasian, P., Irani, P.: Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 1063–1072. ACM (2014)
7. Kibsgaard, M., Kraus, M.: Real-time augmented reality for robotic-assisted surgery. In: Proceedings of the 3rd Aalborg University Workshop on Robotics 2014 (2015, accepted for publication)
8. Lamprecht, B., Nowlin, W., Stern, J.: Stereo telestration for robotic surgery. US Patent 7,907,166 B2 (2011)

9. Ng, A., Lepinski, J., Wigdor, D., Sanders, S., Dietz, P.: Designing for low-latency direct-touch input. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, UIST 2012, pp. 453–464 (2012)
10. Schemali, L., Eisemann, E.: Design and evaluation of mouse cursors in a stereoscopic desktop environment. In: 2014 IEEE Symposium on 3D User Interfaces (3DUI), pp. 67–70. IEEE (2014)
11. Shenai, M.B., Dillavou, M., Shum, C., Ross, D., Tubbs, R.S., Shih, A., Guthrie, B.L.: Virtual interactive presence and augmented reality (VIPAR) for remote surgical assistance. Neurosurgery **68**, ons200–ons207 (2011)
12. Teather, R.J., Stuerzlinger, W.: Pointing at 3D target projections with one-eyed and stereo cursors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, NY, USA, pp. 159–168. ACM, New York (2013). http://doi.acm.org/10.1145/2470654.2470677
13. Volonté, F., Buchs, N.C., Pugin, F., Spaltenstein, J., Schiltz, B., Jung, M., Hagen, M., Ratib, O., Morel, P.: Augmented reality to the rescue of the minimally invasive surgeon. The usefulness of the interposition of stereoscopic images in the da vinci robotic console. Int. J. Med. Robot. Comput. Assist. Surg. **9**(3), e34–e38 (2013)
14. Ware, C., Lowther, K.: Selection using a one-eyed cursor in a fish tank VR environment. ACM Trans. Comput.-Hum. Interact. **4**(4), 309–322 (1997). http://doi.acm.org/10.1145/267135.267136
15. Zhao, W., Wu, C., Hirvonen, D., Hasser, C.J., Miller, B.E., Mohr, C.J., Zhao, T., Di Maio, S., Hoffman, B.D.: Efficient 3-D telestration for local and remote robotic proctoring. US Patent 2015/0025392 A1 (2015)

Instrument Tracking with Rigid Part Mixtures Model

Daniel Wesierski[1(✉)], Grzegorz Wojdyga[1], and Anna Jezierska[2]

[1] Multimedia Systems Department, Faculty of Electronics, Telecommunications, and Informatics, Gdansk University of Technology, Gdańsk, Poland
daniel.wesierski@pg.gda.pl
[2] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Abstract. Tracking instruments in video-assisted minimally invasive surgeries is an attractive and open computer vision problem. A tracker successfully locating instruments would immediately find applications in manual and robotic interventions in the operating theater. We describe a tracking method that uses a rigidly structured model of instrument parts. The rigidly composed parts encode diverse, pose-specific appearance mixtures of the tool. This rigid part mixtures model then jointly explains the evolving structure of the tool parts by switching between mixture components during tracking. We evaluate our approach on publicly available datasets of *in-vivo* sequences and demonstrate state-of-the-art results.

Keywords: Instrument tracking · Video-assisted minimally invasive surgery · Part-based models

1 Introduction

Locating instruments in videos for augmented assistance [1] during minimally invasive surgeries (MIS) has recently received much attention. Minimally invasive surgeries offer a number of advantages over open surgeries. Less postoperative pain, reduced blood loss, minor scarring, and shorter recovery time and hospitalization are attractive factors for inpatients and clinicians. Carrying out such a surgery, though, is a challenging task. The surgeons first make keyhole incisions in the body to insert elongated surgical instruments. Confronted with lost vision and hampered dexterity, the surgeons require additional sensing devices to monitor the instruments maneuvering within the body. While robotic manipulators can control the instruments with high flexibility and stability, their encoders accumulate errors in forward kinematics and lead to inaccurate estimations of the absolute instrument location [11]. On the other hand, specialized hardware sensors and encoders require extensive hardware integration and suffer from lower accuracy [2] thereby cumbersomely integrating to multiple operating rooms. Arguably, widespread color cameras in MIS offer a natural, visual feedback to surgeons. Other imaging modalities such as depth-only sensing devices would be

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 22–34, 2016.
DOI: 10.1007/978-3-319-29965-5_3

hardly interpretable. Amenable to easy transfer between operating rooms and motivated by steady progress of computer vision, vision-based instrument tracking thus constitutes an encouraging approach to improving the guidance and navigation of manual and robotic surgeries.

Description of image features plays a significant role in MIS tool tracking setting. Registered videos may suffer from degraded quality, e.g., due to motion blur. Moreover, adverse lighting conditions in the form of globally varying illumination of the scene, specular reflections on the tool and tissue regions, as well as shadows left by the tool are factors that make tool detection a challenging task in practice. Past work has explored color and gradient features [2,11] to discern greyish tool foreground from reddish and whitish tissue background, markers [13], and used elaborate classification schemes during detection [4,5]. Bootstrapping object appearance from initial frame, that reported remarkable results in the general object tracking setting [16], has recently also been applied to tracking MIS instruments [10] with state-of-the-art performance.

We describe a rigid part mixtures model of a surgical instrument and a detection procedure for tracking its 2D pose (i.e., center and orientation) in MIS videos. As the 3D pose can be recovered from stereo-cameras [1], here we focus on the problem of 2D pose estimation in a single image. While motion models can be used for filtering of, e.g., instrument location and size [9], we achieve good tracks by detecting the instrument pose in each frame independently from neighbor frames. Our model is a spatial assembly of instrument parts that encode mixtures of dedicated pose appearances. By capturing such appearances of an object part at various poses, our approach relates to poselets [15] that reason about fragmented object pose from rigid parts. It differs from poselets by jointly modeling the compositions of small and large part mixtures that can explain full pose of the instrument. Consequently, our approach leverages successful flexible part mixtures model [6] that can be trained with datasets of modest size [17]. Structured part-based models use deformation constraints that act like springs to flex the model to regions with putative objects. Arranged under a tree-graph, they can efficiently explain previously unseen configurations of the flexible object structure but, at the same time, such models can overlap two tip parts on one tip of the tool. In the spirit of poselets, we avoid double-counting image evidence [18] by rigidly modeling end-effector articulations with larger, rigid parts. Hence, our approach differs from past work by enforcing strictly rigid, global compositions of part mixtures and by consistently capturing variable instrument structure.

Our contribution is two-fold. Firstly, we develop a springs-free, structured part-based model of an instrument. It imposes a rigid structure on spatially distributed local features to discard putative tool regions, e.g. in tool neighborhood, that might prompt models with springs to incorrect or flexed structure detections. Secondly, we demonstrate that a structured part-based model can be successfully applied to instrument tracking in MIS. Estimating instrument pose is typically approached in a disjoint manner by first detecting individual parts and then fusing detections with, e.g., a Kalman filter [1] or RANSAC sampling [5]. By exploiting rigidly structured relations between instrument parts,

our tracker detects the end-effector and shaft parts jointly thereby recovering the instrument pose. Applying a structured model, though, is challenging as this requires frequent updates of its underlying structure. Object appearance can vary significantly between frames, especially due to frequent truncations. Specifically, the rigid, straightly elongated shaft has often been used as a discriminative visual cue in detecting the tool and in estimating its 3D pose [3,14]. However, observing that surgeons often prefer to work in close proximity to tissue, [12] ignore the shaft and focus on tracking the articulating end-effector with thousands of efficiently matched templates.

This leads to a dilemma. On the one hand, one would like to take advantage of the shaft part when it is visible. On the other hand, one has to take into consideration the varying, truncated tool structure. Our model-based tracker exploits the rigid shaft while adapting to its changing length. We then discriminatively train dedicated models on a series of training images for each video sequence and show that our method is on par with or exceeds state-of-the-art results in instrument tracking on publicly available datasets.

2 Method

The structure of MIS instruments, e.g. for laparoscopy, retinal microsurgery, in image I can operationally be represented as a pair composition of two parts: (i) a rigid, straight, elongated shaft $\mathcal{S}$ and (ii) a rigid or an articulated end-effector $\mathcal{E}$, as depicted in Fig. 1. Let G_I denote a two-dimensional regular tessellation of the pixel grid of image I, $L \subset \mathbb{N}^{1\times 2}$ a discrete set of locations on the whole grid G_I, and $L_b \subset L$ a discrete set of locations on an arbitrarily shaped (e.g., rectangular, circular) one-dimensional border stripe of this grid.

The $\mathcal{E}$-part is enclosed in a single window in the grid with the center location $l_{\mathcal{E}} \in L \setminus L_b$. An $\mathcal{S}$-part is a collection of N_e subparts that are outlined by adjacent windows. We restrict possible locations of these windows $l_{\mathcal{S}(k)} \in L$, where $1 \leqslant k \leqslant N_e$, to an oriented raster line segment[1] that anchors at $l_{\mathcal{E}}$ and ranges from the border stripe $l_{\mathcal{S}(1)} \in L_b$ to the end of the shaft $l_{\mathcal{S}(N_e)} \in L$. Then, let $l = (l_{\mathcal{E}}, l_{\mathcal{S}(1)})_{1\times 4}$ denote the line segment. As the shaft is often truncated and partially occluded, we represent the $\mathcal{S}$-part as a subcollection of $N \leqslant N_e$ subparts for each new image frame I. As a result, in our model the location of the $\mathcal{S}$-part $l_{\mathcal{S}}(l) = \big(l_{\mathcal{S}(k_1)}(l), \ldots, l_{\mathcal{S}(k_N)}(l)\big)_{1\times 2N}$ determines some ordering of these subparts on the line segment l of the instrument.

In practice, both parts slightly rotate during a surgery while instrument pose admits *non-circumvolving* motion. In general, though, the shaft is oriented at an arbitrary angle as the locations of body incisions vary between surgical scenarios. Moreover, the grippers of the end-effector articulate and take various forms, i.e. the length and shape of the grippers varies. In view of this, we approach the

[1] The location of each subpart is $l_{\mathcal{S}(k)} = l_{\mathcal{S}(1)} + \big[s_k \left(l_{\mathcal{E}} - l_{\mathcal{S}(1)}\right)\big]$, where $[\cdot]$ is the nearest integer function, s_k is a scaling factor $0 \leqslant s_k \leqslant s_{N_e} < 1$, and s_{N_e} ensures that the location $l_{\mathcal{S}(N_e)}$ of the window of the last subpart of the $\mathcal{S}$-part does not overlap with the window of the $\mathcal{E}$-part.

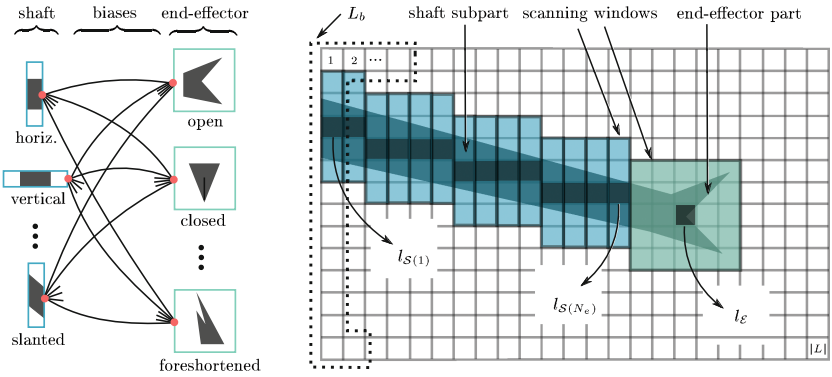


Fig. 1. Our rigid part mixtures model (left) and its instantiation on the grid G_I (right). The model uses (i) a set of appearance templates (i.e., part mixture) that represent a single subpart of the shaft at multiple orientations, (ii) a set of appearance templates that represent various articulations of the end-effector part (e.g., rotated, open or closed gripper), and (iii) a set of biases that promote or discourage rigid appearance compositions of mixture components of the shaft and end-effector parts.

problem of tracking 2D instrument pose by capturing the appearance variation of the tool with a structured model of rigid mixtures of parts that jointly encodes pose-dependent tool appearance.

Model. We represent the appearance and structure of the instruments under graph $M = \{V, E\}$. The appearance mixtures of the end-effector part are chained with the appearance mixtures of the shaft parts. The nodes $V = \{w^i_{\mathcal{E}}, l_{\mathcal{E}}\}^{n_{\mathcal{E}}}_{i=1} \cup \{w^j_{\mathcal{S}}, l_{\mathcal{S}}\}^{n_{\mathcal{S}}}_{j=1}$ denote particular appearances of the $n_{\mathcal{E}}$ end-effector and $n_{\mathcal{S}}$ shaft mixtures, respectively. The i-th component of the appearance mixture of the end-effector part at location $l_{\mathcal{E}}$ is specified by template $w^i_{\mathcal{E}}$ that rigidly encodes specific articulation of this part, as encountered in poselets-based approaches to object recognition [15] and in MIS tool tracking scenarios [12]. The j-th component of the appearance mixture of the shaft part at location $l_{\mathcal{S}}$ is specified by template $w^j_{\mathcal{S}}$ that can capture specific perspective and orientation of the part, e.g. an outwards slanted shaft. The edges $E = \{b^{ij}_{\mathcal{ES}}\}^{n_{\mathcal{E}} \times n_{\mathcal{S}}}_{ij=1}$ model rigid compositions of the end-effector mixture with the shaft mixture. Specifically, the scalar-valued co-occurrences $b^{ij}_{\mathcal{ES}}$ bias configurations of mixtures such that certain, rigidly encoded articulations $w^i_{\mathcal{E}}$ may form more consistent compositions with certain orientations $w^j_{\mathcal{S}}$. In effect, our model admits a strictly rigid structure.

We define the mixture of the shaft part as orientation templates. On the other hand, the $\mathcal{S}$-part lies on the oriented line segment l. Hence, the mapping $j : l_{1\times 4} \rightarrow \mathbb{N}^1$ of a given instance of this oriented line uniquely determines the j-th mixture component of the shaft. Then, instantiating a composition of particular mixture components of the $\mathcal{ES}$-parts in image I at location $l = (l_{\mathcal{E}}, l_{\mathcal{S}(1)})$ is scored with our model as:

$$S(I,l,i) = w^i_{\mathcal{E}}\phi^i_{\mathcal{E}}(I,l_{\mathcal{E}}) + \sum_{p=1}^{N} w^{j(l)}_{\mathcal{S}}\phi^{j(l)}_{\mathcal{S}}(I,l_{\mathcal{S}(k_p)}(l)) + b^{ij(l)}_{\mathcal{ES}} \quad (1)$$

where $\phi^i_{\mathcal{E}}(I,l_{\mathcal{E}})$ and $\phi^{j(l)}_{\mathcal{S}}(I,l_{\mathcal{S}(k_p)})$ are image descriptors (e.g., a `HOG` [8], a color histogram) in the window of the i-th mixture component of the $\mathcal{E}$-part at $l_{\mathcal{E}}$ and in the window of the subpart of the j-th mixture component of the $\mathcal{S}$-part at $l_{\mathcal{S}(k_p)}$, respectively.

The varying length of the shaft notwithstanding, our model allows for taking advantage of the discriminative evidence for this part in each image during tracking. We achieve this with N subparts of the shaft that are anchored at $l_{\mathcal{S}}$. As the elongated shaft roughly admits consistent appearance along the image plane, we deem all subparts of its j-th mixture component to be alike and dedicate a single, canonical template $w^{j(l)}_{\mathcal{S}}$ for representing their appearance. In effect, the subparts, which share the single template, render our model less complex in learning from and matching to images.

Detection. We cast the problem of instrument tracking within the tracking-by-detection framework. We infer the rigid composition of mixture components of the $\mathcal{ES}$-parts at location l that best explains current video frame I by solving the inference problem $\texttt{argmax}_{l,i}\ \ S(I,l,i)$, as depicted in Fig. 2.

Matching the appearance templates $\{w^i_{\mathcal{E}}\}_{i=1}^{n_{\mathcal{E}}}$ and $\{w^j_{\mathcal{S}}\}_{j=1}^{n_{\mathcal{S}}}$ to corresponding image descriptors at each location in L amounts to the convolution in the feature space[2] that yields tables of appearance scores for each mixture component. As our graph M is a mixture of chains, in which $\mathcal{E}$-part mixtures are parents and $\mathcal{S}$-part mixtures are children, we employ dynamic programming as an exhaustive search algorithm over the state space (l,i) to combine the appearance scores across plausible locations and mixture components.

To this end, the search procedure commences by partitioning the border stripe L_b of the grid G_I into $n_{\mathcal{S}}$ disjoint segments $L_b = \bigsqcup_{j=1}^{n_{\mathcal{S}}} L^j_b(l_{\mathcal{E}})$ at given $l_{\mathcal{E}}$. All pairs $(l_{\mathcal{E}}, l_{\mathcal{S}(1)} \in L^j_b(l_{\mathcal{E}}))$ together determine a pencil of line segments. The segments, in turn, indicate all possible orientations of the $\mathcal{S}$-part at $l_{\mathcal{E}}$ within the angular range of the j-th mixture component. As the $\mathcal{S}$-part is represented by N subparts, the score of each hypothesized orientation of the shaft depends on finding such a configuration $l_{\mathcal{S}}(l)$ of image descriptors that best match to the $w^{j(l)}_{\mathcal{S}}$ template. This results in selecting N-best scoring subparts of the shaft within the given line segment.

The search proceeds by enumerating all possible compositions of mixture components of the $\mathcal{ES}$-parts. After aggregating the score $b^{ij(l)}_{\mathcal{ES}}$ of each composition with the N-best scores of the shaft part, the best location $l_{\mathcal{S}(1)}$ of the shaft is selected at given $l_{\mathcal{E}}$ for each i-th mixture component of the end-effector. We then retrieve the best i-th mixture component at $l_{\mathcal{E}}$.

[2] Since the target object can change its scale during tracking, we search over the feature pyramid of $\phi(I,\cdot)$ at run-time.

After repeating this search procedure for each $l_{\mathcal{E}}$, we select $l_{\mathcal{E}}$ with the best aggregated score (1), then backtrack to the best i-th mixture component stored at that location, and terminate at the best $l_{\mathcal{S}(1)}$ pointed by this component.

Learning. We learn the parameters of our rigid part mixtures model in the supervised manner. Our instrument model uses a mixture of appearance templates per part, where only a single template of this part is present in a given positive training image. As we assume a given collection of positive training images contains only keypoint annotations, we retrieve the missing ij-labels of mixture components for each image based on these annotations, as shown in Fig. 3.

We automatically obtain a mixture label of the $\mathcal{E}$ part by first (i) binning the manually labeled end-effector keypoints in a coarse grid (e.g., G_I) and then (ii) grouping the bins features into $n_{\mathcal{E}}$ disjoint sets across all training images. We discard sets with the number of features $< K$. In effect, a given unique spatial arrangement of bins captures a particular articulation of the end-effector. The labels for $\mathcal{S}$-part mixture components are obtained by slicing the image plane into $n_{\mathcal{S}}$ angular intervals. We note when the end-effector part is rigid, we assign the corresponding label of the $\mathcal{S}$-part mixture component to the $\mathcal{E}$-part.

Our rigid part mixtures model is inspired by the flexible part mixtures model [6]. Hence, its array of model parameters is learned jointly and takes the form: $\beta = \left[b^{11}_{\mathcal{ES}}, \ldots, b^{ij}_{\mathcal{ES}}, \ldots, b^{n_{\mathcal{E}} n_{\mathcal{S}}}_{\mathcal{ES}}, w^1_{\mathcal{E}}, \ldots, w^i_{\mathcal{E}}, \ldots, w^{n_{\mathcal{E}}}_{\mathcal{E}}, w^1_{\mathcal{S}}, \ldots, w^j_{\mathcal{S}}, \ldots, w^{n_{\mathcal{S}}}_{\mathcal{S}}\right]$. Since β uses a canonical appearance template $w^j_{\mathcal{S}}$ of a single subpart to *generalize* the appearance of all shaft subparts for j-th mixture component, the function (1) scoring a training feature vector x_n yields the following dot-product form:

$$S(I_n, l, i) = \beta \cdot (0 \ldots 1 \ldots 0 \ldots \phi^i_{\mathcal{E}}(I_n, l_{\mathcal{E}}) \ldots 0 \ldots \phi^{j(l)}_{\mathcal{S}}(I_n, l_{\mathcal{S}(k)}(l)) \ldots 0) = \beta \cdot x_n \tag{2}$$

It induces a sparse structure on x_n that depends on the pre-assignment of mixture labels to respective parts in a given training image I_n.

We then learn the model parameters β with an objective function under linear SVM regime:

$$\begin{aligned} \text{ar gmin}_{\beta,\xi} \quad & \frac{1}{2}\|\beta\|^2 + C^+ \sum_{n=1}^{m^+} \xi_n + C^- \sum_{n=1}^{m^-} \xi_n \\ \text{s.t.} \quad & \beta \cdot x^+_n \geq 1 - \xi_n \ , \quad \forall x^+_n \\ & \beta \cdot x^-_n \leq -1 + \xi_n \ , \quad \forall x^-_n \end{aligned}$$

that can be optimized with, e.g., a dual coordinate-decent solver [6]. The above formulation states that our model β should learn to assign scores higher than 1 to positive examples x^+_n of rigid compositions of respective mixture components and assign scores lower than -1 to negative examples x^-_n. The objective function penalizes violations of these constraints with slack variables $\xi_n \geq 0$, weighted by constants C^+ and C^-. The negative examples x^-_n constitute incorrect detections of the instrument that are mined as hard-negatives on images with masked instruments, as e.g. in [4]. We slightly rotate the positive training images to augment the training set of positive examples.

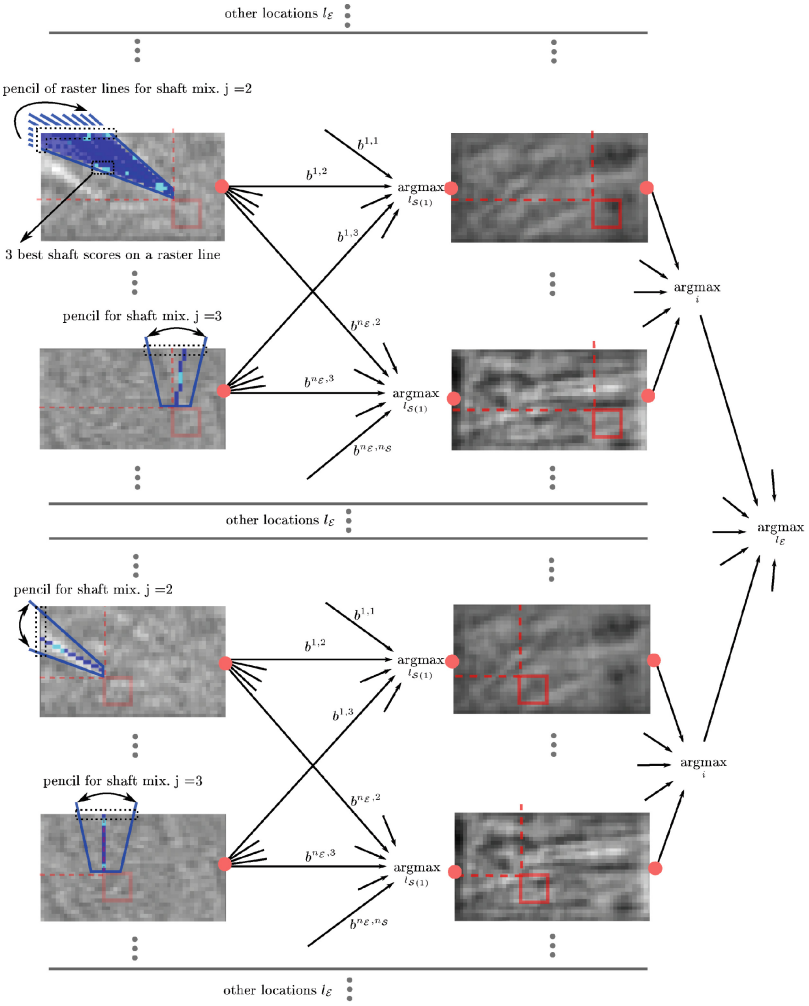


Fig. 2. Instrument detection by matching our model to an image (in feature space). We visualize (best seen in color) two separate iterations. The iterations correspond to two hypothesized locations $l_\mathcal{E}$ of the $\mathcal{E}$-part thereby leading to two different partitions of L_b. In general, for each $l_\mathcal{E} \in L \setminus L_b$, we instantiate $l_{\mathcal{S}(1)}$ yielding all possible oriented line segments that anchor the subparts of the shaft. Each table of appearance scores (in gray) of the $\mathcal{S}$-part (left column) corresponds to j-th mixture component and is selected according to particular instantiation of the line segment. By recursively summing the scores and storing the pointers to selected locations and mixture components, we select (i) N-best scoring subparts of the shaft $l_\mathcal{S}(l)$, followed by (ii) the best line segment $(l_\mathcal{E}, l_{\mathcal{S}(1)})$ per i-th mixture component after adding respective biases, then (iii) best scoring i-th mixture component after adding appearance scores of the $\mathcal{E}$-part (right column), and (iv) terminate by selecting the location $l_\mathcal{E}$ with maximal overall score. As an implementation detail, in the tables the score locations are shifted from the center to upper left corner of every window (Color figure online).

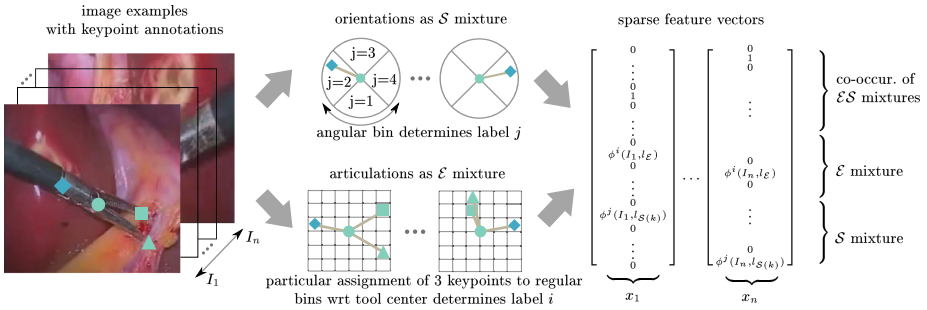


Fig. 3. Learning the mixture labels of the shaft and end-effector parts from image examples determines the sparse structure of feature vectors x_n for SVM classification. The annotations of positive training examples indicate the keypoint locations of two tool tips, tool center $l_\mathcal{E}$, and the end of the shaft. The number of mixture components of both parts, $n_\mathcal{E}$ and $n_\mathcal{S}$, is obtained automatically and depends on the resolution of two respective coarse grids. We use a polar grid to retrieve an orientation type of the shaft part. Then, we rigidly capture articulations of the end-effector part. We first quantize the locations over a regular grid that result in binary occupancy features. We then find their unique groups to retrieve the types of end-effector articulation. Finally, compositions of $\mathcal{ES}$ mixtures serve to store the mixtures co-occurrence indicator and feature descriptors at respective locations in the sparse feature vectors.

3 Results

In this section, we extensively evaluate our method on the task of *in-vivo* single instrument tracking in (i) retinal microsurgery - RM (dataset with 3 sequences [4]), and (ii) spine and pelvic surgery - SPS (dataset with 3 sequences [5]). Both datasets are publicly available.

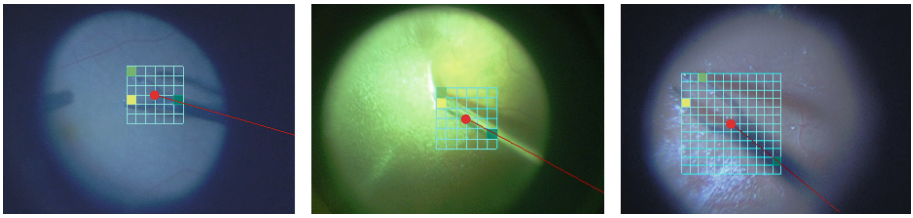


Fig. 4. Instrument pose detection during retinal microsurgery (best viewed in color). Our model detects the tool center (red dot) and the orientation of the shaft (red line). Here, we visualize the windows of end-effector mixture components which are detected on the `HOG` grid (blue). In the spirit of poselets, the model can reason about the articulation of the end-effectors (filled bins on the grid) (Color figure online).

Table 1. Results of our method wrt [5] for estimating the tool center and tool orientation in sequences from RM and SPS datasets. We use the protocol of [5] and evaluate the performance based on the angular mean (Ang. M.) and angular standard deviation (Ang. St.D.) (left) and mean distance (M.) and standard deviation (St.D.) (right). The ratios (train/test) indicate the number of images used for training and testing the model, after the usage guidelines of the SPS dataset. Following the evaluation protocol that omits false negatives and false positives when the tool is present and absent in test images, respectively, we only evaluate the test images that contain the instrument. We note, though, that our tracker could run for images without the tool as the detection threshold is learned within SVM margins. Best results are indicated in bold.

		Ret. 1	Ret. 2	Ret. 3
# Images	[5]	200/152	-	200/297
	Ours	198/152	100/121	196/287
Ang. M.	[5]	4.18	-	**5.31**
	Ours	**3.42**	3.62	5.66
Ang. St.D.	[5]	3.9	-	4.9
	Ours	**2.37**	1.91	**4.53**

		Pelvic 1	Pelvic 2	Spine
# Images	[5]	400/400	100/490	150/322
	Ours	76/213	337/339	91/247
M.	[5]	24.32	16.15	**3.98**
	Ours	**13.87**	**9.07**	13.09
St.D.	[5]	**15.21**	**10.8**	**1.75**
	Ours	45.75	30.25	33.72

Implementation Details. For all sequences, we *equally* configure our model and use *fixed* parameter settings. To make the comparison fair, we follow [4,5] and use the training set of each sequence, as specified in the datasets, to train dedicated models of the instruments. We compute window sizes of the end-effector and shaft parts from the keypoint annotations in the training images.

The appearance templates are defined in `HOG` feature space [7]. We set `sbin`= 8 for `HOG` cells, $K = 10$ for pruning groups of bins features when learning the $\mathcal{E}$ mixtures, and $N = 3$ for the number of detected shaft subparts. To specify the orientation labels for the $\mathcal{S}$-part, we follow the `HOG` specification of 18 equal orientation intervals over $(-\pi, +\pi\rangle$. The number of labels $n_{\mathcal{S}}$ is then determined based on the annotated instruments in the training set. We set $C^+ = 0.004$ and $C^- = 0.002$ to account for $m^+ < m^-$ imbalance in the training set.

Qualitative Evaluation. We qualitatively show that our model can detect the 2D pose of the instrument (i.e., center location of the end-effector and orientation of the shaft) as well as the articulation of the end-effector, as shown in Fig. 4. In RM sequences, the tracker yields robust tracks despite illumination variations and disrupting tool-like shadows. In SPS sequences, the instruments significantly change their scale, are partially occluded, and often heavily truncated. Our method is able to successfully locate the end-effectors in these sequences. It can adapt to the varying length of the shaft by searching for the best-scoring subparts along the hypothesized, oriented shafts (Fig. 5).

Quantitative Evaluation. We report quantitative results in Table 1 as mean distance precision and standard deviation from the ground truth (i) tool center for SPS and (ii) tool orientation for RM. In addition, we report the percentage

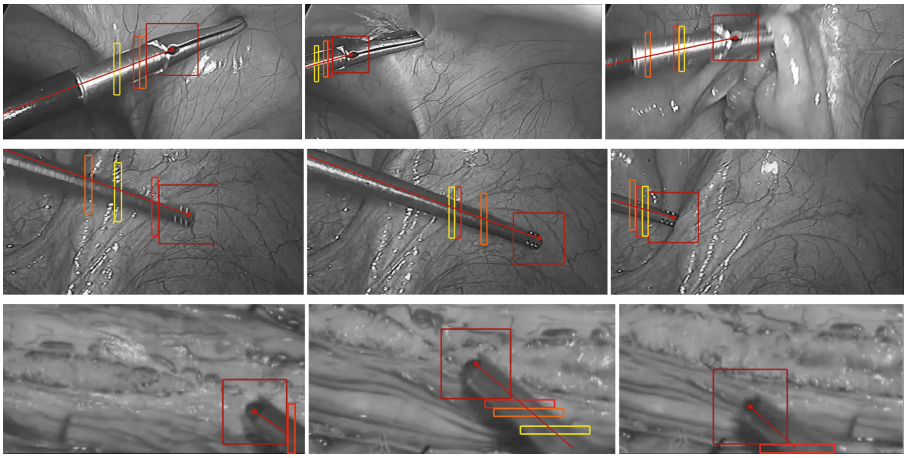


Fig. 5. Instrument pose detection (best viewed in color) in pelvic 1 (top row), pelvic 2 (middle), and spine (bottom) sequences. We show (i) the windows of 3 best shaft subparts that are detected with a single, canonical template and (ii) the window and its center of the end-effector part. We learn the appearance of the $\mathcal{E}$-part based on its window center as indicated by the SPS dataset annotations. The end-effector and shaft mixture labels are equal as the $\mathcal{E}$-part is considered rigid having no articulations in the sequences. Note how mixture components of the shaft part switch to explain the varying orientation of the instrument. Also, the colors red–orange–yellow of the shaft subparts indicate the 1^{st}, 2^{nd}, and 3^{rd} best detection, respectively. The tracker detects the tool at multiple scales (top row). By selecting the best scoring subparts along the shaft, the tracker takes advantage of the discriminative appearance of the shaft (middle row) while at the same time it copes with heavy truncations (bottom row) (Color figure online).

of accurate detections of the tool center within a given pixel range for RM in Fig. 6. We demonstrate that the proposed rigid part mixtures model achieves state-of-the-art results on both benchmarks.

In Table 1, we do well in terms of smaller mean distance precision measure. Our high deviation error wrt [5] for SPS sequences comes from far but rare misdetections of the tool center. In general, though, our method yields stable tracks in the RM and SPS sequences.

In Fig. 6, we are on par with other trackers. We outperform other methods in Ret. 1, but do worse in Ret. 2 wrt [10]. However, while [10] successfully tracks the tool center, our tracker also outputs tool orientation (Figs. 4 and 5). Finally, we examine the reliance of our detector on the length of the shaft. We show that our model, augmented with more subparts of the shaft, better stabilizes the detections thereby leading to improved performance (Fig. 6d).

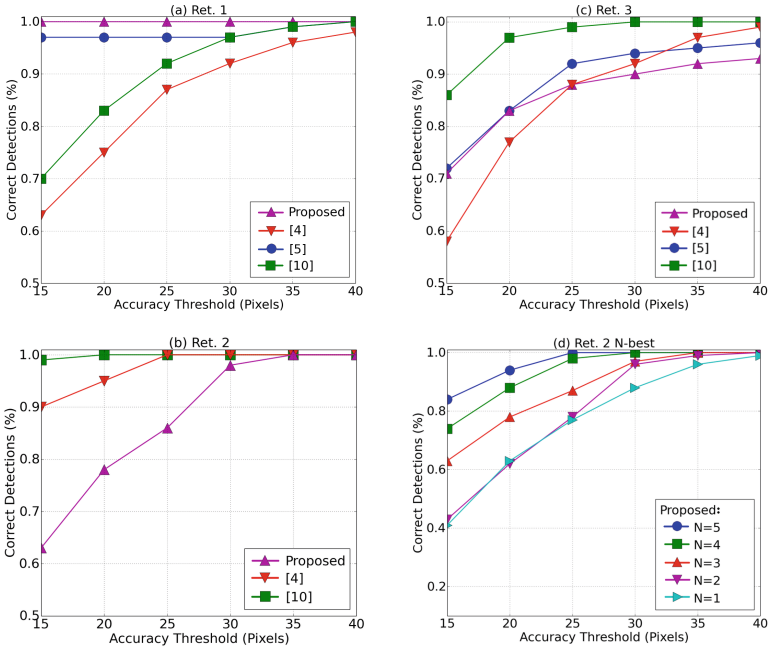


Fig. 6. The results for retina microsurgery dataset on the task of end-effector localization. Our method performs best in Ret. 1 and is on par with [4,5] in Ret. 3. However, we do worse in Ret. 2 and Ret. 3 wrt [10] but additionally output tool orientation. In the last graph (d), we show that the performance of our method scales proportionally with N-best subparts of the shaft. When our model uses 5 subparts, it effectively levels up with [10].

4 Conclusions and Future Work

We proposed a rigid part mixtures model for structurally representing the appearance of surgical instruments in MIS videos. The model robustly explains the evolving object structure by switching between part mixture components that rigidly encode pose-specific appearances of the tool. In effect, our versatile approach to tracking 2D instrument pose reaches state-of-the-art results on two public benchmarks and often improves the estimation of tool location and orientation upon other trackers. We also showed that increasing visual shape cues by a larger pool of shaft subparts leads to more stabilized tool tracking.

Tracking instruments in MIS scenarios is a challenging task. The shaft undergoes frequent truncations, the end-effector can have many degrees of freedom in articulation, such as the da Vinci instruments, and both parts can be occluded when multiple tools are present. At the same time, a tool tracking algorithm should run at frame rates ideally exceeding real-time to minimize the latency of visual feedback and thereby to improve augmented assistance in MIS. Our future work will concentrate on these challenges.

References

1. Reiter, A., Allen, P.K., Zhao, T.: Feature classification for tracking articulated surgical tools. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 592–600. Springer, Heidelberg (2012)
2. Allan, M., Thompson, S., Clarkson, M.J., Ourselin, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: 2D-3D pose tracking of rigid instruments in minimally invasive surgery. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 1–10. Springer, Heidelberg (2014)
3. Wolf, R., Duchateau, J., Cinquin, P., Voros, S.: 3D tracking of laparoscopic instruments using statistical and geometric modeling. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 203–210. Springer, Heidelberg (2011)
4. Sznitman, R., Ali, K., Richa, R., Taylor, R.H., Hager, G.D., Fua, P.: Data-driven visual tracking in retinal microsurgery. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 568–575. Springer, Heidelberg (2012)
5. Sznitman, R., Becker, C., Fua, P.: Fast part-based classification for instrument detection in minimally invasive surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 692–699. Springer, Heidelberg (2014)
6. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2013)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893. IEEE Press, New York (2005)
9. Sznitman, R., Basu, A., Richa, R., Handa, J., Gehlbach, P., Taylor, R.H., Jedynak, B., Hager, G.D.: Unified detection and tracking in retinal microsurgery. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 1–8. Springer, Heidelberg (2011)
10. Li, Y., Chen, C., Huang, X., Huang, J.: Instrument tracking via online learning in retinal microsurgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 464–471. Springer, Heidelberg (2014)
11. Reiter, A., Allen, P.K., Zhao, T.: Appearance learning for 3D tracking of robotic surgical tools. Int. J. Robot. Res. (2013)
12. Reiter, A., Allen, P.K., Zhao, T.: Marker-less articulated surgical tool detection. In: Computer Assisted Radiology and Surgery (2012)
13. Zhao, T., Zhao, W., Halabe, D.J., Hoffman, B.D., Nowlin, W.C.: Fiducial marker design and detection for locating surgical instrument in images. Patent US 068395, 07 08 (2010)
14. Doignon, C., Nageotte, F., de Mathelin, M.: Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision. In: Vidal, R., Heyden, A., Ma, Y. (eds.) WDV 2005/2006. LNCS, vol. 4358, pp. 314–327. Springer, Heidelberg (2007)
15. Lubomir, B., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: ICCV, pp. 1365–1372. IEEE (2009)

16. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409–1422 (2012)
17. Zhu, X., Vondrick, C., Ramanan, D., Fowlkes., C.: Do we need more training data or better models for object detection? In: BMVC (2012)
18. Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., Sheikh, Y.: Pose machines: articulated pose estimation via inference machines. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 33–47. Springer, Heidelberg (2014)

Stereoscopic Motion Magnification in Minimally-Invasive Robotic Prostatectomy

A. Jonathan McLeod[1,2(✉)], John S.H. Baxter[1,2], Uditha Jayarathne[1,2], Stephen Pautler[3], Terry M. Peters[1,2], and Xiongbiao Luo[1]

[1] Robarts Research Institute, Western University, London, Canada
jmcleod@robarts.ca
[2] Biomedical Engineering Graduate Program, Western University, London, Canada
[3] Division of Urology, Department of Surgery, Western University, London, Canada

Abstract. The removal of the prostate is a common treatment option for localized prostate cancer. Robotic prostatectomy uses endoscopic cameras to provide a stereoscopic view of the surgical scene to the surgeon. Often, this surgical scene is difficult to interpret because of variants in anatomy and some critical structures such as the neurovascular bundles alongside the prostate, are affected by variations in size and shape of the prostate. The objective of this article is to develop a real-time stereoscopic video processing framework to improve the perceptibly of the surgical scene, using Eulerian Motion Magnification to exaggerate the subtle pulsatile behavior of the neurovascular bundles. This framework has been validated on both digital phantoms and retrospective analysis of robotic prostatectomy video.

Keywords: Robotic prostatectomy · Motion magnification · Stereoscopic video processing

1 Introduction

Prostate cancer accounts for 27 % of all new male cancer diagnoses in the United States in 2014 and almost 30,000 deaths [10]. Prosatectomy, in which the prostate is surgically resected, is a common treatment for localized prostate cancer. Although open surgery is possible, an increasing number of centers have used a minimally invasive technique using laproscopic guidance. Robotic prostatectomy, in which the intervention is performed with the aid of a series of robotic arms, has become the dominant form of minimally invasive prostatectomy procedure owing to improved usability, faster recovery times, fewer complications, and lower blood loss [2,4,5].

Critical structures to avoid during prostatectomy include the neurovascular bundles, two sets of coupled arteries and nerves running along both sides of the prostate, damage to which can lead to complications in terms of urinary and erectile function [12]. Movement towards nerve (and vessel)-sparing procedures has thus been identified as crucial in improving patient quality-of-life post-surgically [9].

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 35–45, 2016.
DOI: 10.1007/978-3-319-29965-5_4

Fluorescence contrast agents such as those used with the daVinci FireFly system have previously been used for real-time prostate lymphangiography to enhance the visual salience of the neurovascular bundles [6]. However, such techniques may be limited by the availability of fluorescence imaging equipment and the additional instrumentation costs imposed on what is already considered an expensive procedure [3].

The arterial pulsation of the neurovascular bundles provides a valuable cue in their localization for vessel sparing procedures [12] but this pulsation is subtle and difficult to detect. The goal of this article is to use recent advances in real-time motion analysis and video processing to enhance the pulsation of sensitive vasculature to make it more visible to the surgeon.

1.1 Related Work

Recently, motion analysis for the detection of vasculature in endoscopic video has been investigated for a number of minimally-invasive procedures. We previously proposed an intensity based monoscopic motion magnification pipeline for vessel-sparing in endoscopic procedures, specifically endoscopic third ventriculostomy and laproscopic prostatectomy. This pipeline made use of adaptive filtering to track the heart-rate and proposed several methods for artifact reduction [8].

Amir-Khalili et al. [1] used a pipeline similar to phase-based motion magnification to detect major arteries and veins during the Hilar dissections in partial nephrectomy. Rather than enhance the pulsatile motion, their pipeline used the filter response to segment the occluded vessels.

Another technique for motion based segmentation of pulsating structures was developed to detect dural pulsation during ultrasound-guided epidural injections and on relies extended Kalman filtering to fit a parametric model to pixel intensities [7].

In this paper we apply motion magnification to stereoscopic video. Causal and computationally efficient spatial-temporal filtering methods are developed to handle the challenges presented by online processing of stereoscopic high definition video. A digital phantom is developed for simulating pulsatile motion of vessels with a known ground truth and both intensity and phase based pipelines are evaluated on this phantom and restrospective human video.

2 Methods

Eulerian motion magnification as proposed by Wu et al. [15] and Wadhwa et al. [13,14] takes advantage of small changes in the intensity or local image phase of a video stream corresponding to the movement. Magnifying these changes directly gives the perception of amplified motion without requiring the estimation of an explicit motion field making magnification computationally efficient and suitable for real-time applications. This framework relies on a two-stage process where the video is first spatially filtered using either a Laplacian or a Riesz pyramid then temporally filtered to extract coherent intensity changes.

Although originally intended for monoscopic video analysis, Eulerian motion magnification can also be used to exaggerate motion in stereoscopic videos without inducing perceptual artifacts regarding depth cues. The primary concern of this approach is the translation of three-dimensional motion to a two-dimensional representation. To illustrate this, consider the Taylor series expansion of the image intensity.

$$I(y(x+\delta x)) \approx I(y(x)) + \nabla I J_y \delta x \tag{1}$$

where $I(y)$ is the intensity of the image at pixel index y, ∇I is the gradient of the image intensity with respect to the pixel index, $y(x)$ is the camera projection of the 3D position x, and J_y is the Jacobian of said projection operator. For a pin-hole camera model, this term is equal to:

$$J_y = \begin{bmatrix} \frac{fx_1}{x_3}, & 0, & \frac{-fx_1}{x_3^2} \\ 0, & \frac{fx_2}{x_3}, & \frac{-fx_2}{x_3^2} \end{bmatrix} \tag{2}$$

where the coordinate origin is defined to be at the camera focus and the depth direction, x_3, perpendicular to the image plane. The pin-hole camera projection is linear with respect to x_1 and x_2. If we were to magnify any motion in the 3D co-ordinates, that is, magnify δx by α, this equation yields:

$$\begin{aligned} I(y(x+(1+\alpha)\delta x)) &\approx I(y(x)) + (1+\alpha)\nabla I J_y \delta x \\ &\approx I(y(x+\delta x)) + \alpha\left(I(y(x+\delta x)) - I(y(x))\right) \end{aligned} \tag{3}$$

So long as $\delta x_3 << x_3$ the linear approximation will be valid to the traditional bounds proposed by Wu et al. [15].

The motion magnification pipelines examined follow along the lines of those proposed by Wu et al. [15] and Wadhwa et al. [13,14] as shown in Fig. 1. This linear approximation is applied to both cameras to synthesize stereoscopic video with enhanced motion. One key consideration in this framework is the computational efficiency of the pipeline, as the intended use requires real-time processing of two high-definition endoscopic video streams.

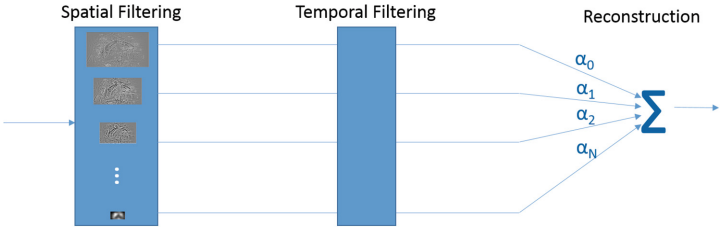


Fig. 1. Overview of a typical motion magnification pipeline

Rapid artifact reduction [8] can be performed after reconstruction which ensures that intensity estimates are clamped within a reasonable range based on the intensity distribution of the area in which motion is magnified.

2.1 Spatial Filtering

A critical aspect of the original Eulerian motion magnification process is the decomposition of the video into spatial frequency bands through a Laplacian pyramid. This decomposition allows for the same pixel-wise temporal processing to be performed over each spatial frequency band but with differing amplification factors. The differentiation of amplification factors allows for the framework to be adjusted according to the application in which it is used. Because of the linearity of this pipeline each frequency band in the pyramid is the result of linear filtering (aside from minor interpolation artifacts) and the subsequent temporal filtering is also linear as is the weighted summation used for pyramid reconstruction. As a result, the order of these operations can be swapped. The pipeline is equivalent to performing the temporal filtering on a bandpassed image. If the first and last levels of the pyramid are given a weight of zero as is traditionally done to remove high frequency noise and changes in overall background (illumination), and the remaining levels are fully amplified, this pyramid is equivalent to a difference of Gaussian filter with standard deviations corresponding to the first and last pyramid level. Further attenuation of the high frequency components can be achieved by adjusting the difference of Gaussian filter. This modification is highly attractive from a computational aspect as it reduces the amount of spatial and temporal filtering required, especially if the Gaussian filtering is approximated using a combination of IIR box-car filters.

2.2 Temporal Filtering

A key consideration for temporal filtering is the selection of the motion frequencies to be amplified. To remove the signal from background motion and only enhance pulsation, the filter must be selective to a relatively narrow bandwidth around the heart-rate. For offline processing, this can be accomplished by taking the Fourier transform of the entire video. However, for real-time applications the temporal filtering must be causal, i.e. dependent only on previous frames, and should be computationally efficient. In addition to their frequency domain analysis, Wu et al. [15] constructed 2nd order IIR filters from the difference of two lowpass filters. Here, we consider the general biquad filter with transfer function:

$$H(z) = K\frac{1 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{4}$$

When implemented in Direct Form II, [11] this filter requires only two delay registers, four additions and five multiplications per pixel. The gain can be factored out as shown here and combined with the magnification factor to reduce additional computations. To design the filter we would like the following properties: complete DC rejection, relatively narrow bandwidth about the target frequency, and a unit gain with zero phase shift at the target frequency. The biquad filter has two complex poles, the radius of which control the bandwidth of the filter. One of the zeros is fixed at 1+ 0i to remove DC. The natural frequency of the poles and the location of the remaining zero can then be chosen to achieve a

peak gain at the target frequency with zero phase shift. Values for these last two parameters can be found to satisfy the target frequency and zero phase shift constraints through numerical optimization methods, such as the downhill simplex algorithm. The coefficients corresponding to a heart rate of 90 bpm sampled at 30 fps with a pole radius of 0.95 are shown in Table 1. This filter was used for all experiments presented in this paper. In practice, the filter coefficients could be updated based on the heart-rate obtained through an ECG or pulse-oximeter, or be updated adaptively using a least mean squares algorithm as presented in our previous work [8].

Table 1. Filter coefficients for heart-rate of 90 bpm sampled at 30 fps

b_1	b_2	a_1	a_2	K
−0.0336	−0.966	−1.809	0.9025	0.050

2.3 Phase-Based Filtering

One issue with a purely intensity-based formulation of Eulerian motion magnification is the presence of artifacts at high spatial frequencies. These artifacts are especially noticeable around high contrast edges and textured objects. To address this, Wadhwa et al. [13,14] developed a phase-based approach which used Riesz pyramids to extract local amplitude and phase information. Using this information, the magnified motion can be added as a phase shift in each sub-band. This pipeline cannot be collapsed, as obtaining the local phase is non-linear and only valid over a narrow frequency range.

3 Digital Phantom Experiment

To initially validate the motion processing pipeline and to guide later design decisions, a digital phantom with an *a priori* known motion profile was generated. This phantom consisted of a flat plane with a protruding half cylinder representing the artery. A 512×320 patch showing a small artery was taken from an HD endoscopic image inside the abdominal cavity during a radical prostatectomy procedure. This image was mapped to the phantom to provide realistic texture. A 3D deformation field can then be applied to the artery to simulate a wide range of motions such as translations in lateral and depth directions, expansion of the artery or more complex motion patterns (Fig. 2).

Two virtual cameras are placed 6 cm from the target with a camera separation of 8 mm and the stereoscope video was simulated through surface ray-casting. For the phantom experiment we simulated two motion paterns, 0.025 mm translation in the lateral direction and a 0.025 mm expansion/contraction in diameter. These motion patterns corresponded to 2D displacements of approximately half a pixel in the simulated video and were subtle enough to be difficult for a human to

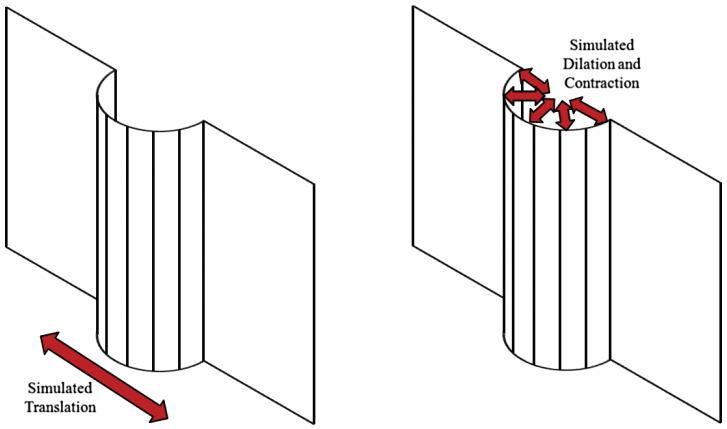


Fig. 2. Implementation of translation and expansion/contraction in the digital phantom.

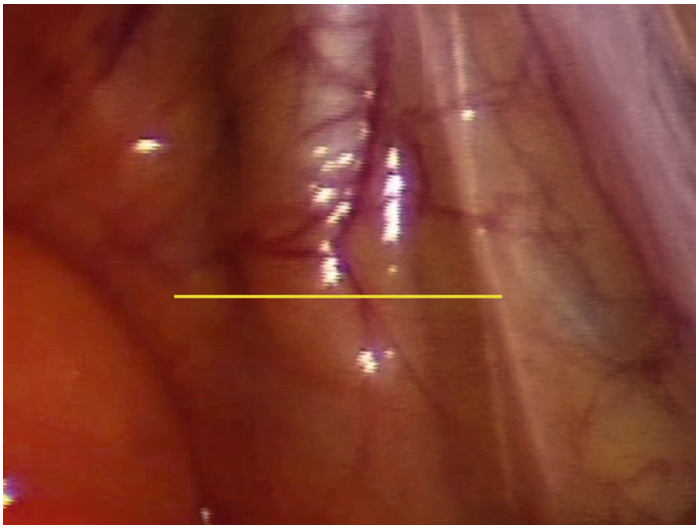


Fig. 3. First frame with yellow line indicating location of subsequent time profile visualization

observe. Using the same model we also simulated videos with 2, 4, 6, 8, and 10 times the original motion to serve as gold standards. Both the intensity-based and phase-based motion magnification pipelines were applied to the original profiles in the digital phantom. The resulting motion amplified videos were then compared against the gold standard videos in terms of root mean square error (RMSE) on the video intensities.

3.1 Results

Both intensity and phase pipelines produced visually convincing results. The pulsation, barely visible in the original videos, was greatly enhanced and very prominent in the processed videos. Visual results for a translation and expansion/contraction are given in Figs. 4 and 5 respectively. The time-profile for the gold standard and processed videos clearly show translational motion of the artery in the first case and expansion and contraction of the artery in the second. Qualitatively, it appears that both the magnification pipelines slightly underestimate the amount to which the motion should be magnified to achieve the ground truth videos. This is likely due to some of the intensity and phase changes caused by the motion falling outside the pass-band of the spatial-temporal filtering and becoming attenuated. The RMSE results for the amplified video intensity are shown in Fig. 6a and b for translational motion and Fig. 6c and d for expansion/contraction. The RMSE minima for the intensity-based pipeline correspond closely to the correct magnification factor. In all cases, the artifact reduction technique based on min/max filtering reduced the RMSE but this improvement was most prominent with the intensity-based pipeline. The minima of the RMSE curves for the intensity-based pipeline correspond to the correct magnification factor, however this property was not observed in the phase-based pipeline or after artifact reduction. This may be due to the increased processing required for these methods. The Riesz pyramid also results in some reconstruction error. While this error slightly increases the RMSE metric for phase-based pipelines, especially at low magnification factors where the RMSE from other sources is quite low, they were not visually perceptible and have minimal impact on the quality of the final video.

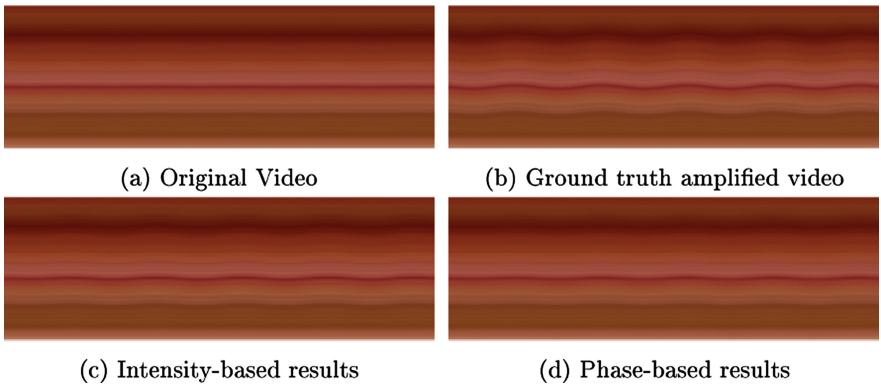


Fig. 4. Time profile visualization for translation motion pattern at 10x magnification. The yellow line in Fig. 3 indicates the location where the time profile is taken. The pixel intensities along the vertical direction correspond to pixel intensities along the yellow line in the video. The horizontal direction corresponds to time with the left side of the profile being the beginning of the video.

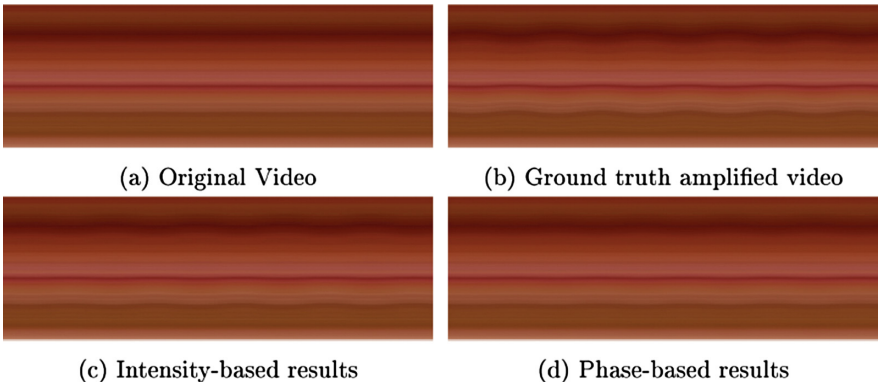


Fig. 5. Time profile visualization for expansion/contraction motion pattern at 10x magnification. The yellow line in Fig. 3 indicates the location where the time profile is taken.

4 Retrospective Video Analysis

The intensity and phase-based motion magnification pipelines were applied retrospectively to human video from a robotic radical prostatectomy using the daVinci surgical robot. Time profiles of the results are shown in Fig. 7. A magnification factor of $\alpha = 5$ was chosen based on the small magnitude of the motion being sought out, the pulsation of the neurovascular bundles, and the desired perceptibility of the motion in the amplified video. The challenge posed by these videos was to selectively magnify arterial pulsation in the presence of background motion and endoscopic lighting conditions.

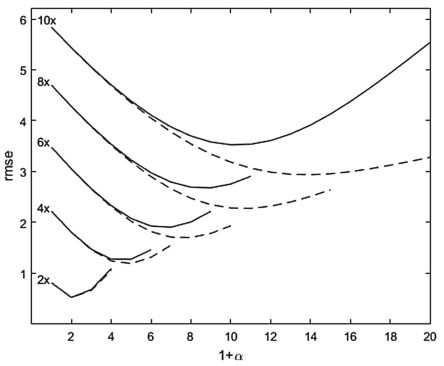
4.1 Results

As shown in the time profiles, both intensity-based and phase-based pipelines were able to selectively magnify motion at the heart-rate while being impervious to gross, non-periodic shifts in the scene as shown approximately 40 % of the way through the time profiles. Thus, both pipelines were readily able to increase the perceptibly of small arteries without excessive perceptual interference from gross prostate shift and the motion of the surgical tools.

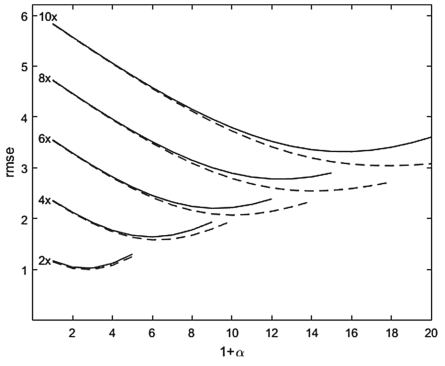
Qualitatively, the phase-based pipeline produced slightly higher quality amplified videos than the intensity-based pipeline, even considering artifact correction, however; the saliency of the arterial pulsation is excellent in both. The advantage of the intensity based pipeline is the reduced computational cost, especially when using only a bandpass spatial filter.

5 Discussion and Future Work

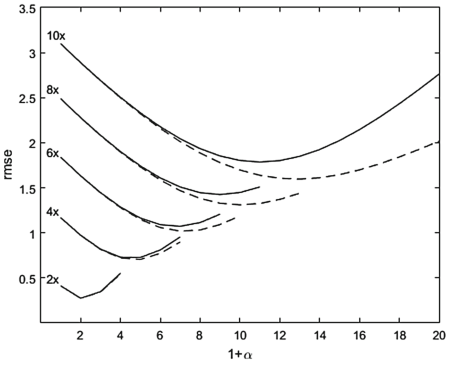
Stereoscopic motion magnification using both the intensity-based and phase-based Eulerian motion magnification pipeline has been shown to work for the



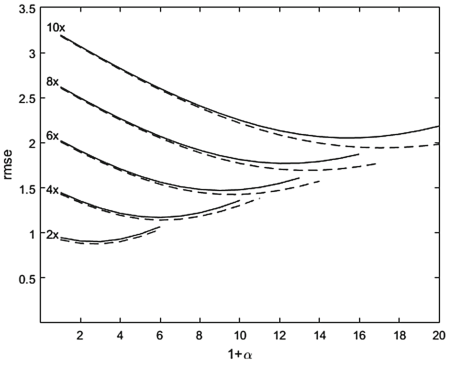
(a) Intensity-based pipeline for translation



(b) Phase-based pipeline for translation



(c) Intensity-based pipeline for expansion

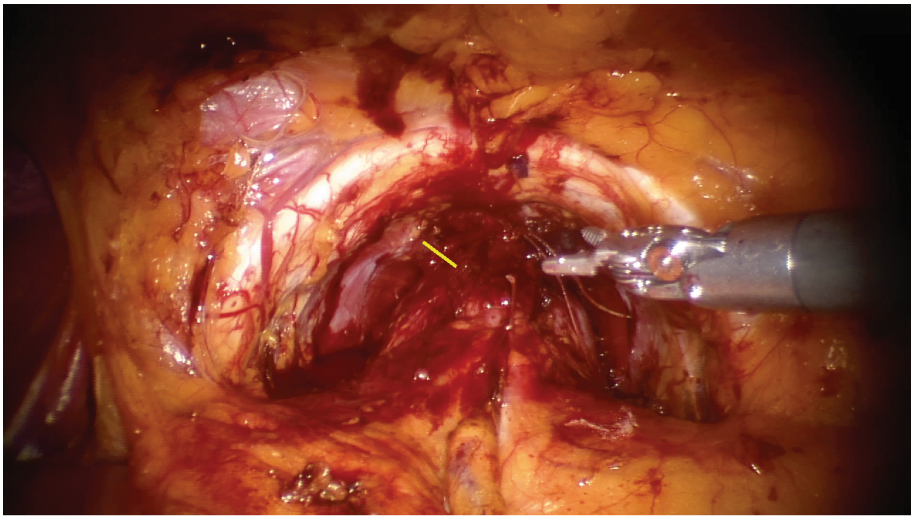


(d) Phase-based pipeline for expansion

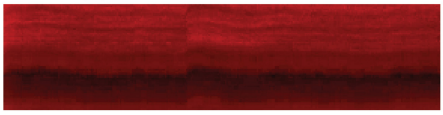
Fig. 6. RMSE error between each motion magnification pipeline and the digital phantom gold standard for magnification values varying from 1 to 10. The solid line shows the original pipeline while the dashed line shows the same results after artifact reduction through min/max filtering.

amplification of small motions relative to the depth of the moving object. This framework may be useful in laproscopic interventions in which a stereoscopic video source is available, such as in robot-assisted prostatectomy. Both pipelines have been characterized on a digital phantom with realistic motion and texture properties mimicking those seen in human prostatectomy videos. Eulerian motion magnification was then performed retrospectively on the stereoscopic human video showing a distinct increase in the perceptibly of vasculature exhibiting subtle pulsation.

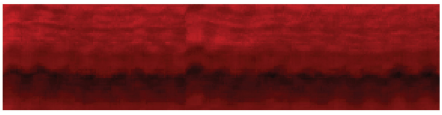
Future work for this project includes the integration of this framework into the daVinci robotic prostatectomy system with associated software modifications to ensure real-time performance. This will include work in accelerating the algorithm using general purpose graphics card programming, taking advantage of the inherently parallelization capabilities of the framework.



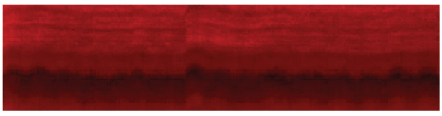
(a) First frame with yellow line indicating location for the time profile



(b) Original Video



(c) Intensity-Based Results



(d) Phase-Based Results

Fig. 7. Time profile visualization of intensity-based and phase-based motion magnification pipelines at $\alpha = 5$ magnification. The yellow line in Fig. 7a indicates the location from which the time profile is taken.

6 Conclusion

This article presents an Eulerian motion magnification framework tailored for use in robot-assisted prostatectomy for vessel sparing. This framework is computationally efficient, using only causal filtering, and is designed to operate real-time on streaming stereoscopic high-definition video.

Acknowledgments. We would like to thank Elvis Chen for his invaluable discussion and editing. Funding for this project was received from Intuitive Surgical, Canadian Institute for Health Research and Canadian Foundation for Innovation. Graduate student funding for Jonathan McLeod was received from the Vanier Canadian Graduate Scholarship program.

References

1. Amir-Khalili, A., Hamarneh, G., Peyrat, J.M., Abinahed, J., Al-Alao, O., Al-Ansari, A., Abugharbieh, R.: Automatic segmentation of occluded vasculature via pulsatile motion analysis in endoscopic robot-assisted partial nephrectomy video. Med. Image Anal. **25**(1), 103–110 (2015)
2. Anderson, J.E., Chang, D.C., Parsons, J.K., Talamini, M.A.: The first national examination of outcomes and trends in robotic surgery in the United States. J. Am. Coll. Surg. **215**(1), 107–114 (2012)
3. Delto, J.C., Wayne, G., Yanes, R., Nieder, A.M., Bhandari, A.: Reducing robotic prostatectomy costs by minimizing instrumentation. J. Endourol. **29**(5), 556–560 (2015)
4. Kowalczyk, K.J., Levy, J.M., Caplan, C.F., Lipsitz, S.R., Yu, H.Y., Gu, X., Hu, J.C.: Temporal national trends of minimally invasive and retropubic radical prostatectomy outcomes from 2003 to 2007: results from the 100% medicare sample. Eur. Urol. **61**(4), 803–809 (2012)
5. Liu, J.J., Maxwell, B.G., Panousis, P., Chung, B.I.: Perioperative outcomes for laparoscopic and robotic compared with open prostatectomy using the national surgical quality improvement program (nsqip) database. Urology **82**(3), 579–583 (2013)
6. Manny, T.B., Patel, M., Hemal, A.K.: Fluorescence-enhanced robotic radical prostatectomy using real-time lymphangiography and tissue marking with percutaneous injection of unconjugated indocyanine green: the initial clinical experience in 50 patients. Eur. Urol. **65**(6), 1162–1168 (2014)
7. McLeod, A.J., Baxter, J.S., Ameri, G., Ganapathy, S., Peters, T.M., Chen, E.C.: Detection and visualization of dural pulsation for spine needle interventions. Int. J. Comput. Assist. Radiol. Surg. **10**, 947–958 (2015)
8. McLeod, A.J., Baxter, J.S., de Ribaupierre, S., Peters, T.M.: Motion magnification for endoscopic surgery. In: SPIE Medical Imaging, pp. 90360C–90360C. International Society for Optics and Photonics (2014)
9. Sanda, M.G., Dunn, R.L., Michalski, J., Sandler, H.M., Northouse, L., Hembroff, L., Lin, X., Greenfield, T.K., Litwin, M.S., Saigal, C.S., et al.: Quality of life and satisfaction with outcome among prostate-cancer survivors. N. Engl. J. Med. **358**(12), 1250–1261 (2008)
10. Siegel, R., Ma, J., Zou, Z., Jemal, A.: Cancer statistics, 2014. CA Cancer J. Clin. **64**(1), 9–29 (2014)
11. Smith, J.O.: Introduction to Digital Filters: With Audio Applications. W3K Publishing (2007)
12. Tewari, A., Peabody, J.O., Fischer, M., Sarle, R., Vallancien, G., Delmas, V., Hassan, M., Bansal, A., Hemal, A.K., Guillonneau, B., et al.: An operative and anatomic study to help in nerve sparing during laparoscopic and robotic radical prostatectomy. Eur. Urol. **43**(5), 444–454 (2003)
13. Wadhwa, N., Rubinstein, M., Durand, F., Freeman, W.T.: Phase-based video motion processing. ACM Trans. Graph. (TOG) **32**(4), 80 (2013)
14. Wadhwa, N., Rubinstein, M., Durand, F., Freeman, W.T.: Riesz pyramids for fast phase-based video magnification. In: 2014 IEEE International Conference on Computational Photography (ICCP), pp. 1–10. IEEE (2014)
15. Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J.V., Durand, F., Freeman, W.T.: Eulerian video magnification for revealing subtle changes in the world. ACM Trans. Graph. **31**(4), 65 (2012)

Tissue Shape Acquisition with a Hybrid Structured Light and Photometric Stereo Endoscopic System

Marco Visentini-Scarzanella[1(✉)], Tatsuya Hanayama[1], Ryunosuke Masutani[2], Shigeto Yoshida[3], Yoko Kominami[3], Yoji Sanomura[3], Shinji Tanaka[3], Ryo Furukawa[2], and Hiroshi Kawasaki[1]

[1] Department of Information Systems and Biomedical Engineering, Kagoshima University, Kagoshima, Japan
marco.visentiniscarzanella@gmail.com
[2] Department of Endoscopy and Medicine, Hiroshima University, Hiroshima, Japan
[3] Department of Intelligent Systems, Hiroshima City University, Hiroshima, Japan

Abstract. *In situ* 3D reconstruction from endoscopic images is important to determine the correct course of action for, e.g., treatment of abnormal growths. Currently, the endoscopist has to rely solely on visual cues in order to infer the growth's shape and size and determine an appropriate treatment. However, tissue uniformity and scale ambiguity from traditional monocular endoscopes make this visual assessment prone to errors and time consuming. We propose a practical system to densely reconstruct both shape and size of tissues with minimal modifications to a standard endoscope. We present a custom single-fiber structured light probe projecting a wave pattern on the tissue surface that allows semi-dense reconstruction with few ambiguities. Based on the coarse reconstruction, we retrieve the surface reflectance parameters according to a hybrid diffuse/specular model which are used to initialise a close-range Photometric Stereo reconstruction. By taking into account the tissue characteristics and the light fall-off, our Photometric Stereo formulation provides dense metric 3D shape information without the need for surface normal integration. A preliminary study was carried out both on phantoms and *ex vivo* samples of human tissue.

1 Introduction

In recent development of endoscopy technology, diagnosis and treatment using endoscopes on digestive tracts have been widely performed [1]. As for the treatment of early stage gastric cancers, treatment methods differ depending on the size of the tumours. For this reason, accurate measurement of the size of neoplasias is important. Currently, forceps and 2D visual cues are used by the endoscopist to assess the size of polyps, but this is error-prone and time consuming. Therefore, techniques for objective measurements are desirable.

Intraoperative 3D reconstruction from endoscopic images has been the focus of extensive research in recent years, and a comprehensive review of the state-of-the-art can be found in [2,3]. However, as many of the systems mentioned in the

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 46–58, 2016.
DOI: 10.1007/978-3-319-29965-5_5

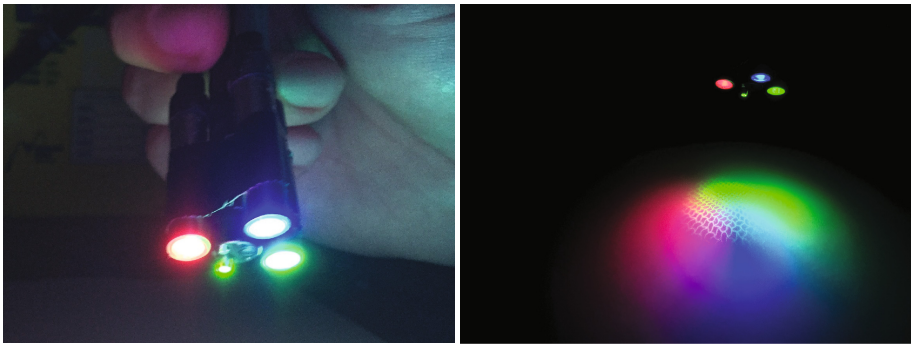
review article such as stereo or Structure-from-Motion require either costly stereo cameras or multiple images, solution to the one-shot reconstruction problem remains elusive. Indeed, in the recent evaluation of one-shot 3D reconstruction technique [3], 6 out of 8 methods require a stereo camera, whereas the remaining two are Structured Light and Time-of-Flight systems. Structured light (SL) is one of the systems that can solve the one-shot reconstruction problem that has seen recent applications in endoscopy: in [4], a micro pattern projector was mounted outside the endoscope for 3D reconstruction, in [5] a SL device for tubular structures was proposed, while sparse reconstruction with spectral encoding was studied in [6]. However, generally a sparse reconstruction of a limited area can be obtained, and noise as well as tissue texture can prevent large areas from being reconstructed and obtain reliable size information.

Earlier works in depth cue fusion [7] suggested combining sparse, reliable feature-based methods such as structured light with dense photometric-based techniques that can be initialised with sparse information. Photometric-based techniques such as Shape-from-Shading have been applied to endoscopy [8–10], but they require either pre-operative data for registration, prior calibration procedures, or intra-operative calibration to resolve the scaling ambiguity and recover absolute depth. Photometric Stereo (PS) is a technique that has been applied to endoscopy [11] on Lambertian surfaces, but required external markers for initialisation of the illumination response matrix. Recent developments in PS [12] allow direct computation of the depth without integration of the normal field after sparse depth initialisation.

In this work, we contribute by proposing a SL endoscope with an integrated projector in its instrument channel. We improve the PS formulation for endoscopes by using the sparse reconstruction to initialise a PS technique that is independent of the surface albedo and explicitly takes into account light intensity distribution and position, while being robust to non-Lambertian areas. Importantly, our system does not require significant alterations to standard equipment. To the best of our knowledge, this is the first work combining structured light and photometric stereo applied to endoscopy. Preliminary results on phantoms and *ex vivo* human tissue samples show interesting possibilities for further research.

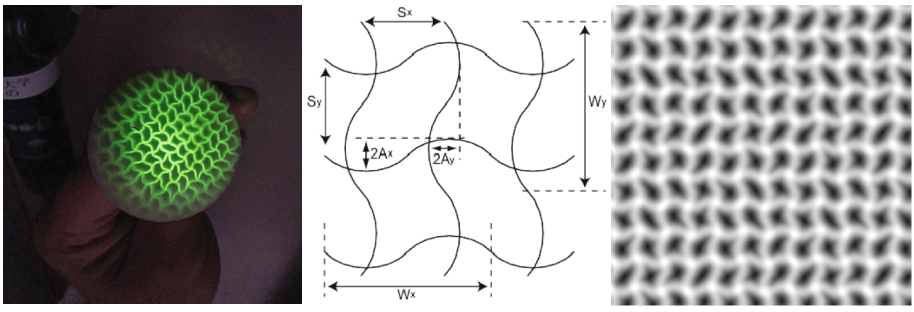
2 Method

Our proposed system consists of two main modules: first, we miniaturised a laser pattern projector consisting of a single optical fiber that can be fed through the instrument channel of the endoscope for SL projection. Second, the PS module consists of three externally mounted LEDs. While our prototype does not satisfy the endoscope size requirements due to the external mount, in the final product stage it will be possible to include three internal LEDs, or to apply colour filters to the on-board lighting as done in [11]. The overall system with the projected lights and patterns is shown in Fig. 1a and b. Reconstruction with the SL module is discussed in Sect. 2.2, while the final PS-based reconstruction is presented in Sect. 2.3. Prior to the reconstruction, however, together with the standard



(a) (b)

Fig. 1. (a) Close-up of the prototype. The structured light projector protrudes from the tool channel, while the three red, green and blue LEDs are mounted around the scope. (b) Projected patterns from LEDs and structured light. (Color figure online)



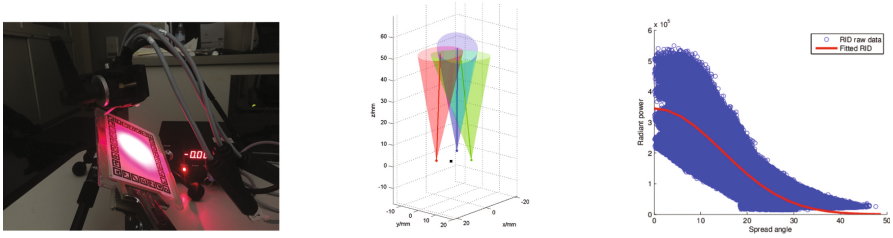
(a) (b) (c)

Fig. 2. (a) Calibration sphere with projected SL pattern. (b) Design of the wave pattern. (c) Projected SL pattern.

camera and photometric calibration, it is necessary to calibrate the SL and LED setup. This is discussed in the next section.

2.1 LED and Structured Light Calibration

Following camera and photometric calibration, calibration of extrinsic matrix relating the SL projector with the endoscope camera is performed following the steps outlined in [4]: images of a spherical object with known dimensions are taken while projecting the SL pattern as shown in Fig. 2a. A known pixel is used for initial matching between the projected and visualised patterns, while the distinctive crossing points of the wave pattern allow to find unique matches for calibration refinement. The wave pattern used is adapted from [13] and is shown in Fig. 2b. The wave lines are sinusoidal patterns, with equal wavelengths between vertical and horizontal lines. However, since the vertical spacing is not



(a) (b) (c)

Fig. 3. (a) RID calibration for red LED. (b) Estimated position and orientation of LEDs. (c) Estimated RID. (Color figure online)

equal to an integer multiple of the horizontal wavelength, the intersection points (or grid points) appear at the different phases on the wave patterns (Fig. 2c). This implies that the local pattern around an intersection point has a local uniqueness, which can be used as a helpful discriminative feature both during calibration as well as reconstruction. While our current prototype assumes the SL projector to be fixed relative to the camera, this is seldom the case during endoscopy since the tool channel might be needed for other purposes. We therefore proposed a system for auto-update of the calibration parameters, to appear in [14].

One of the crucial stage is accurate calibration of the LED setup. While this traditionally only involves positional calibration of the LEDs, which is achieved usually through triangulation of specular highlights, it is also necessary to estimate the Radiance Intensity Distribution (RID) of the light sources. The RID is a function $g(\phi)$ describing the power emitted by a light source according to the angle ϕ from its main axis. This is important since virtually all models used in SFS and PS assume an 'omnilight' model where the light emits radiance isotropically in all directions, and the incident light energy on the surface is only attenuated by the light fall-off inversely proportional to the square of the distance. However, as verified also in [11], this model is inadequate to the illumination types found on scopes, where the highly focused light is normally emitted up to $15°-30°$ from the main direction. To this end, we have recently proposed a system (to appear in [15]) for joint practical calibration of light position and RID. The technique leverages the fact that the projection of a light beam on a Lambertian plane will be symmetrical about an axis related to its position and orientation. Also, we prove that the point of maximum intensity lies on said axis, hence it is possible to recover the light position and orientation just by looking at its points of local maxima. In Fig. 3a, we show a frame from the red LED calibration, done with a matte calibration plane with AR markers for positional information around a blank space for RID calibration. The estimated position of the LEDs is shown in Fig. 3b, reflecting the triangle formation around the central camera (black point), while the fitted RID to the information is shown in Fig. 3c. While the absolute amplitude of the RID is only of relative importance since it depends on the material reflectance, the important aspect to calibrate is

the RID fall-off, estimated in our experiment to get to 10 % of its amplitude at approximately 25°. This corresponds to an analytical expression for the RID of $g(\phi) = \cos(\phi)^{15.58}$, which will be used in our PS reconstruction.

2.2 Structured Light Shape Acquisition

Our structured light system configuration is shown in Fig. 4(a). This consists of a FujiFilm VP-4450HD system coupled with a EG-590WR scope. The pattern projector is inserted through the instrument channel of the scope, with the projector lens slightly protruding from the head. The light source of the projector is a green laser module with a wavelength of 532 nm. The laser light is transmitted through a plastic optical fiber with a diameter of 1.8 mm to the tip of the projector. A micro pattern chip with the printed pattern is set at the tip of the fiber. The transmitted light passes through the micro pattern chip and then through the aspherical lens, with a field of view of 30°.

For the structured light shape acquisition, our method is based on template matching using grid-point features in [16]. In this system, given a wave intersection template dictionary, matches are searched along epipolar lines. Since the matching process can be affected by the pattern distortion as it is projected on the surface, this distortion is represented as an affine transformation with two DOFs. Finally, the overall template matching cost is estimated using the optimal surface normal directions given the affine transformation matrix. The

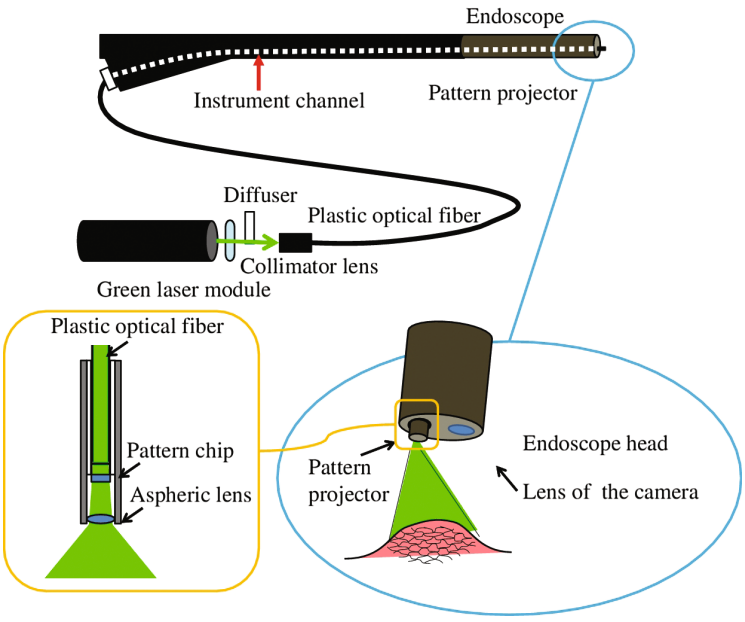


Fig. 4. (a) Proposed structured light endoscopic setup

patch-based template matching cost is then regularized by Belief Propagation according to the cost function:

$$E(T) = \sum_{p \in V} D(p, t_p) + \sum_{(p,q) \in U} W(t_p, t_q), \tag{1}$$

where $T = \{t_p | p \in V\}$ is a set of correspondences with t_p, t_q being the corresponding grid point of pattern points p and q respectively, V is the set of the grid points, D is the correlation cost between the grid points of the image and the pattern, and W is a distance cost which is 0 if p and q are neighbouring grid points and positive otherwise. This cost function is small if p and t_p have small matching cost D, and the connections between the grid points are coherent for the image and the pattern.

2.3 Photometric Stereo Reconstruction

Multispectral Photometric Stereo (MPS) [17] traditionally aims to recover surface normals from a single image of a Lambertian surface simultaneously illuminated by a red, blue and green light. In this way, the intensity recorded by each colour channel roughly corresponds to the reflected light from one of the light sources. More formally:

$$I = \rho(P \cdot L)\mathbf{n}, \tag{2}$$

where ρ is the monochromatic albedo, P is the matrix representing the spectral crosstalk between CCD sensors, L the concatenated light directions and $\mathbf{n}$ the target surface normals. Traditionally, methods involve a sparse reconstruction using Structure-from-Motion to estimate P, after which the normals can be found from a simple matrix inversion given the known light positions [17]. While PS in the general Computer Vision community can benefit from a number of simplifying assumptions such as a larger number of inputs and directional lighting, in MIS the problem is highly challenging. Only recently Collins and Bartoli [11] adapted the MPS problem to close-range lighting, recovering depth through a 2-stage local/global approach. However, they maintain the assumptions of uniform albedo, Lambertian reflectance through the use of polarising filters, and assume the presence of external tools/markers to estimate the unknown matrix P. More recently, in [12] a Fast Marching (FM) based procedure was proposed for close-range PS, which starting from a single known point, or 'seed', propagates the information while explicitly taking into account the light RID and distance attenuation.

We build on these past improvements by proposing a novel technique that considers a more realistic reflectance model with specularities. Moreover, we adapt the propagation method in [12], for explicit depth reconstruction independent of the albedo and integrated with our structured light initialisation. First, similarly to [17] with Structure-from-Motion and PS, we use our SL reconstruction to estimate the matrix P and calculate $\tilde{I} = P^{-1}I$, the

crosstalk-compensated image. Then, we consider the Blinn-Phong model for diffuse/specular reflection:

$$\tilde{I} = \frac{g\,(\mathbf{v}\cdot\mathbf{a})}{\|\mathbf{l}\|^2}\left(\rho_1\frac{\mathbf{l}\cdot\mathbf{n}}{\|\mathbf{l}\|\|\mathbf{n}\|} + \rho_2\left(\frac{(\mathbf{l}+\mathbf{v})\cdot\mathbf{n}}{\|\mathbf{l}+\mathbf{v}\|\|\mathbf{n}\|}\right)^{\alpha}\right), \tag{3}$$

where $g()$ is the light RID estimated through calibration [15], $\mathbf{l}$ the light vector at a point, $\mathbf{v}$ the view vector, (ρ_1, ρ_2) the diffuse and specular albedos and α the exponent determining the sharpness of specularities. One of the nice characteristics of the Blinn-Phong model, apart from approximating well surfaces with specularities, is the additive relationship between diffuse and specular components, which we exploit in our method.

We notice that in Minimally Invasive Surgery (MIS) specularities are very sharp due to tissue characteristics and focused lighting. This implies a very high α, which in turn implies that the specular term will be essentially zero for $\frac{(\mathbf{l}+\mathbf{v})\cdot\mathbf{n}}{\|\mathbf{l}+\mathbf{v}\|\|\mathbf{n}\|} < 0.95$, meaning that most of the image will be essentially Lambertian apart from specularities that can be identified through intensity/saturation thresholding.

We therefore proceed first to reconstruct the diffuse portion of the surface. In [12] it is shown that the PS problem involving light sources with RIDs of the form $g(\phi) = \cos(\phi)^{\beta}$ amounts to solving the PDE:

$$\begin{cases} \mathbf{b}_{ij}(x,y,z)\cdot\nabla z(u,v) = s_{ij}(x,y,z), & (u,v)\in\Omega_p \\ z(u,v) = p(u,v), & (u,v)\in\partial\Omega_p \end{cases}, \tag{4}$$

where (x, y, z) are the 3D coordinates, (u, v) the pixel coordinates, $z(u, v)$ the function that maps a pixel value with its depth, $p(u, v)$ are Dirichlet boundary conditions and:

$$\mathbf{b}_{ij} = \begin{pmatrix} \tilde{I}_i(u,v)q_i^{\beta+3}(x,y,z)L_i^x - \tilde{I}_j(u,v)q_j^{\beta+3}(x,y,z)L_j^x \\ \tilde{I}_i(u,v)q_i^{\beta+3}(x,y,z)L_i^y - \tilde{I}_j(u,v)q_j^{\beta+3}(x,y,z)L_j^y \end{pmatrix}. \tag{5}$$

The vector $\mathbf{b}$ can be calculated from any pair of image channels (i, j) from the image intensity, the corresponding light source position (L^x, L^y), the distance q from the light source to a pixel and the RID parameter β. The scalar function $s(x, y, z)$ is:

$$s_{ij}(x,y,z) = \left(\tilde{I}_i(u,v)q_i^{\beta+3}(x,y,z) - \tilde{I}_j(u,v)q_j^{\beta+3}(x,y,z)\right)\frac{z(u,v)^2}{f}, \tag{6}$$

where f is the focal length. The solution to Eq. (4) can be found through the Fast Marching algorithm with the following forward upwind scheme, for the k^{th} iteration:

$$z^{k+1} = \frac{\|b_{ij}^1(z^k)\|z^k_{u-\mathrm{sgn}\left(b_{i,j}^1(z^k)\right),v} + \|b_{ij}^2(z^k)\|z^k_{u,v-\mathrm{sgn}\left(b_{i,j}^2(z^k)\right)} + s_{ij}(z^k)}{\|b_{ij}^1(z^k)\| + \|b_{ij}^2(z^k)\|}. \tag{7}$$

Finally, starting from our SL seeds, we can outline the Fast Marching algorithm:

1. Initialise SL seed values and add all the points' neighbours to a list of points to be visited. Initialise all other points to the depth of their nearest seed.
2. Traverse the list of points to be visited and calculate $\mathbf{b}$ and s for each point. Update the value of z according to a standard forward upwind scheme.
3. Add the point's neighbours to the list of points to visit and repeat until convergence.

The algorithm above allows us to reconstruct Lambertian areas. However, it is adversely affected by specularities. In [12], the authors are able to 'steer' the propagation direction to propagate around cast shadows and avoid their interference in the propagation process. We propose the same strategy around specular highlights detected via standard intensity thresholding, so that the Lambertian portion of the image can be reconstructed error-free. Details on steering the Fast Marching propagation are in [12].

3 Results

We tested the performance of our proposed method on simulated data, phantom models and *ex vivo* human tissue samples. We first tested the PS and SL modules individually on simulations and phantom models respectively, while the full system was deployed for the *ex vivo* tissue samples. All C++ code has been executed on a standard consumer grade laptop and is available on the author's site. In terms of computational complexity, a single iteration of the serial, unoptimised code for our algorithm was found to execute in 280 ms for a 400×400 image on the CPU, with 3–4 iterations usually required until convergence. Given these timings, real-time operation is deemed to be feasible upon parallelisation of the fast marching algorithm.

3.1 Photometric Stereo Evaluation

For our simulations, we used the *AbsPeaks* and *Sphere* datasets. The models were virtually placed at around 20 mm and 60 mm respectively, and were chosen to evaluate the performance of the algorithm on both smooth and irregular surfaces with discontinuities. The rendering was done by placing a virtual camera at the origin and three light sources with similar intrinsic characteristics as those founds on the endoscope used for our experiments. For each dataset, an artificial albedo with normalised values ranging from 0.6 to 1 was generated with a Perlin noise process in order to show the algorithm's independence to the object texture. This can be seen in the sample renderings in Fig. 5. The Blinn-Phong exponent for specularities was set to 100. To further evaluate the performance of our algorithm in noisy environments, we performed two tests: first, the input images were injected with i.i.d. noise drawn from a uniform distribution with ranges $[0, 0.05I_{max}]$ and $[0, 0.1I_{max}]$ respectively, where I_{max} is the maximum intensity of the image, corresponding to 5 % and 10 % noise respecitively. These results are shown in Table 1(a). Finally, we tested the sensitivity of the algorithm to errors

Table 1. Reconstruction accuracy with absolute and relative errors for synthetic datasets against **(a)** image noise and **(b)** seed noise.

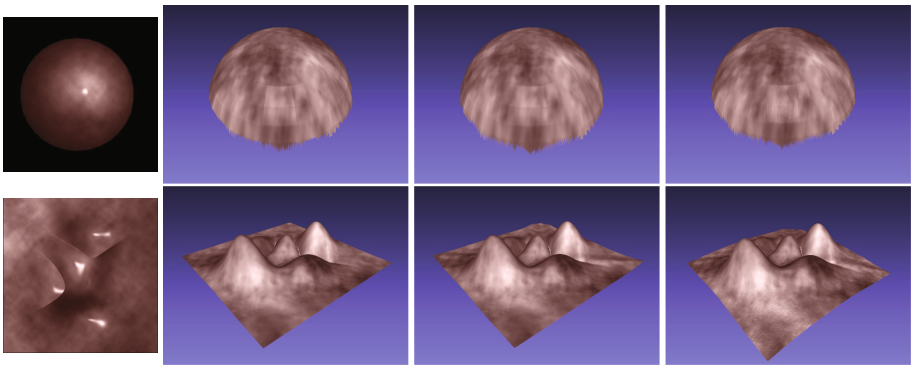
	Accuracy vs. Image noise		
	0%	*5%*	*10%*
Peaks */%*	0.04%	0.19%	0.38%
Peaks */mm*	0.01mm	0.04mm	0.08mm
Sphere */%*	0.08%	0.08%	0.23%
Sphere */mm*	0.05mm	0.05mm	0.14mm

(a)

	Accuracy vs. Seed noise		
	0%	*5%*	*10%*
Peaks - *Single seed*	0.04%	5.77%	11.65%
Peaks - *Multiple seeds*	0.04%	0.03%	0.02%
Sphere - *Single seed*	0.08%	5.618%	10.38%
Sphere - *Multiple seeds*	0.08%	0.12%	0.11%

(b)

in the initial seeds, by adding 5 % and 10 % noise to the initial seed values. Whenever a single input seed was used, this was placed in the middle of the image, and exactly 5 % and 10 % was added to its value. Whenever multiple seeds were used, the noise was drawn from a zero-mean uniform distribution and the seeds randomly placed across the dataset. These results are shown in Table 1(b). All results show a good performance of the algorithm under both noiseless and noisy conditions.



(a) (b) (c) (d)

Fig. 5. Reconstruction results for the Sphere (first row) and Peaks (second row) datasets. Columns (a–d) show an example rendered image from the dataset, the ground truth depth, and reconstruction results with no noise and 10 % noise added respectively.

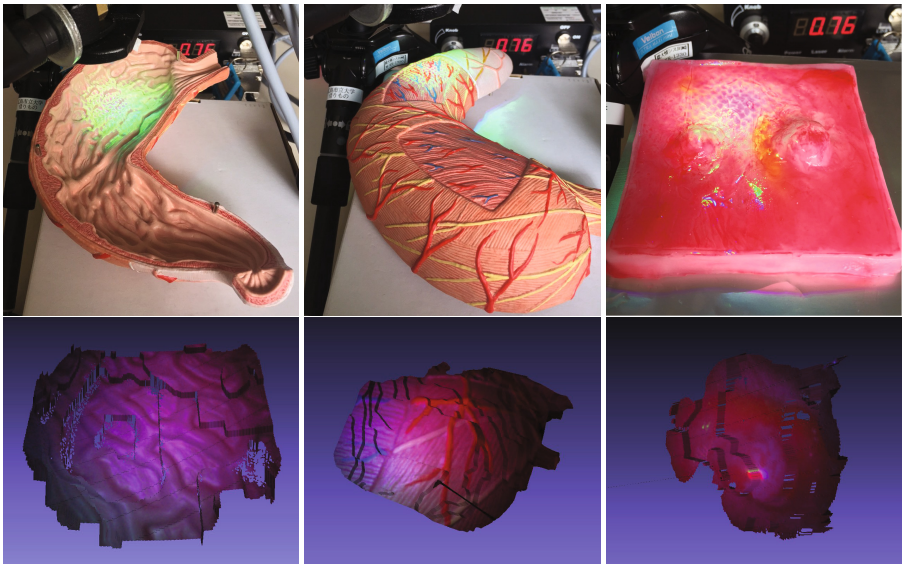
3.2 Structured Light Evaluation

To evaluate the performance of our SL system, we test it first on an anatomical stomach model (Kyoto Kagaku) and then on a custom tumour phantom created

to reproduce the reflectance characteristics of live tissue (Wetlab). The models were placed at approximately 7 cm from the scope, reconstructed and their depth value at the point perpendicular to the scope compared with our manual measurements. Phantoms and renderings of their reconstructions are shown in Fig. 6. The SL reconstruction is able to successfully reconstruct the general trend of the surfaces, like the convexity of the phantom stomach in models (a) and (b), and the distance was measured to be within 5 mm of its actual value. However, due to the lack of subpixel accuracy and the discreteness of disparity values, the reconstructions lack detail and exhibit some staircasing. In the tumour phantom for example, while there is a rise in the reconstructed volume corresponding to the lump, we are unable to clearly capture its boundaries. This aspect will be investigated as part of our future work, since a good initial reconstruction is crucial for correctly estimating the reflectance parameters needed by the PS module.

3.3 *Ex vivo* samples

In our experiments, *ex vivo* human tissue samples were collected and scanned at Hiroshima Hospital for qualitative validation. Three samples in total were collected during endoscopic resections of esophageal, gastric and colonic tumours. The samples were then fixed on a rigid support and scanned with our system. The input images to the system under multispectral and structured lighting



(a) (b) (c)

Fig. 6. SL evaluation. **Top:** images of (a), (b) anatomical stomach model and (c) tumour phantom. **Bottom:** their corresponding reconstruction renderings

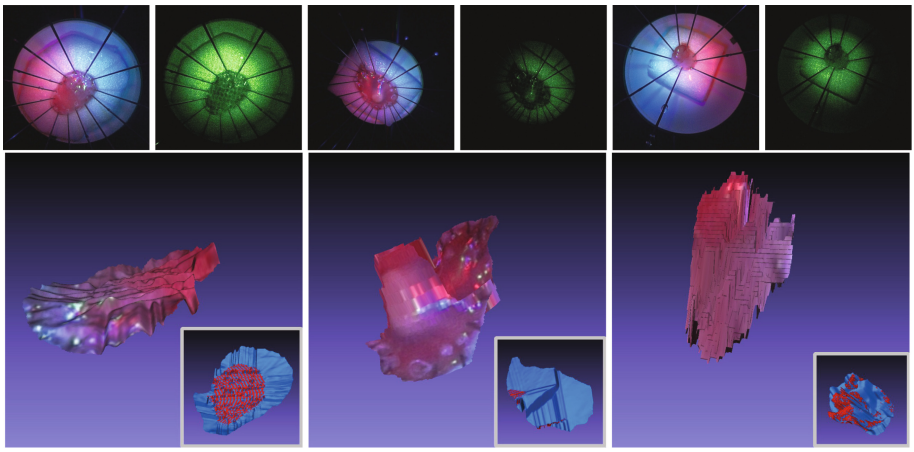


Fig. 7. *Ex vivo* samples. **Top:** SL and PS images of the collected samples. **Bottom**: reconstruction renderings with the real tissue texture. Inset, from the camera perspective, red points are the SL initialisation, the blue surface shows the final PS reconstruction. (Color figure online)

are shown in Fig. 7(top). The reconstruction renderings (bottom) are shown both with their true texture, while the inset images shown from the camera perspective are colour-coded to denote the reconstruction source: the red points represent the initial structured light reconstruction, and the blue surface the final recovered shape.

In the first sample (Fig. 7(a)) the structured light pattern was clearly visible and a large central portion of the sample was reconstructed with structured light. The final propagated surface as seen from the camera perspective exhibits a largely flat trend, with a gentle downward slope towards the upper right, which is confirmed by visual assessment of the camera images. The oscillations on the side of the surface are due to the fact that in our current formulation each point exclusively considers information propagated along the shortest path, which means that it is affected by three seed points at most. Hence, any noise in neighbouring initial seeds will give rise to slight ripples in the surface. With our approach, it was possible to obtain the metric size of the complete sample rather than the smaller area that could be reconstructed with structured light.

In the second sample, the darker images only allowed to reconstruct a small portion of the surface. While the final propagated surface is piecewise flat, it clearly exhibits a protruding tip in correspondence of the white mass in the center of the sample. This was confirmed by additional images of the same sample, where while the high frequency details were lost, the overall trend of a flat surface with a central mass was successfully recovered. The flatness of the reconstruction is due mostly to the lack of subpixel accuracy in our SL initialisation, leading to a staircasing effect. We expect that for applications such as endoscopic navigation,

with larger areas imaged, the method can benefit from a clearer contribution from the light fall-off and RID terms.

Finally, we show our third sample as a failure case, with the sample standing vertically on its base. The angle caused a noisy estimate of the seeds, leaving only a generic surface trend pointing towards the camera. Future work will aim to filter out noisy seeds from the process. While imaging conditions as of now do not allow the recovery of smaller details, but can still be used for to assess the extent of the area.

4 Conclusions

We propose a hybrid Structured Light/Multispectral Photometric Stereo system for tissue size and shape acquisition in endoscopy. Our SL probe is small enough to fit through a standard instrument port, and the wave pattern employed allows reconstruction of a sharp monochromatic laser pattern with few ambiguities. We have adapted a state-of-the-art Fast Marching PS technique to the challenging MIS environment and showed its performance under noisy and textured environments with specularities. Promising preliminary results have been shown on simulated data and *ex vivo* human tissue samples showing a good ability to recover the overall size and general surface trend even in challenging conditions. Future work will focus on detailed reconstruction and integrating the LEDs inside the scope head in order to allow *in vivo* operation.

Acknowledgments. This work was supported by The Japanese Foundation for the Promotion of Science, Grant-in-Aid for JSPS Fellows no. 26.04041.

References

1. Chadebecq, F., Tilmant, C., Bartoli, A.: How big is this neoplasia? live colonoscopic size measurement using the infocus-breakpoint. Med. Image Anal. **19**, 58–74 (2015)
2. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., Stoyanov, D.: Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. Med. Image Anal. **17**, 974–996 (2013)
3. Maier-Hein, L., Groch, A., Bartoli, A., Bodenstedt, S., Boissonnat, G., Chang, P.L., Clancy, N., Elson, D., Haase, S., Heim, E., Hornegger, J., Jannin, P., Kenngott, H., Kilgus, T., Muller-Stich, B., Oladokun, D., Rohl, S., dos Santos, T., Schlemmer, H.P., Seitel, A., Speidel, S., Wagner, M., Stoyanov, D.: Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction. IEEE Trans. Med. Imaging **33**, 1913–1930 (2014)
4. Furukawa, R., Aoyama, M., Hiura, S., Aoki, H., Kominami, Y., Sanomura, Y., Yoshida, S., Tanaka, S., Sagawa, R., Kawasaki, H.: Calibration of a 3D endoscopic system based on active stereo method for shape measurement of biological tissues and specimen. In: IEEE International Engineering in Medicine and Biology Conference (EMBC), pp. 4991–4994 (2014)

5. Schmalz, C., Forster, F., Schick, A., Angelopoulou, E.: An endoscopic 3D scanner based on structured light. Med. Image Anal. **16**, 1063–1072 (2012)
6. Clancy, N.T., Stoyanov, D., Maier-Hein, L., Groch, A., Yang, G.Z., Elson, D.: Spectrally encoded fiber-based structured lighting probe for intraoperative 3D imaging. Biomed. Opt. Express **2**, 3119–3128 (2011)
7. Visentini-Scarzanella, M., Mylonas, G.P., Stoyanov, D., Yang, G.-Z.: *i*-BRUSH: a gaze-contingent virtual paintbrush for dense 3D reconstruction in robotic assisted surgery. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 353–360. Springer, Heidelberg (2009)
8. Wu, C., Narasimhan, S., Jaramaz, B.: A multi-image shape-from-shading framework for near-lighting perspective endoscopes. Int. J. Comput. Vis. **86**, 211–228 (2010)
9. Malti, A., Bartoli, A.: Estimating the cook-torrance BRDF parameters in-vivo from laparoscopic images. In: Workshop on Augmented Environment in Medical Image Computing and Computer Assisted Intervention (MICCAI), Nice, France (2012)
10. Ciuti, G., Visentini-Scarzanella, M., Dore, A., Menciassi, A., Dario, P., Yang, G.Z.: Intra-operative monocular 3D reconstruction for image-guided navigation in active locomotion capsule endoscopy. In: IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), pp. 768–774 (2012)
11. Collins, T., Bartoli, A.: 3D reconstruction in laparoscopy with close-range photometric stereo. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 634–642. Springer, Heidelberg (2012)
12. Mecca, R., Wetzler, A., Bruckstein, A.M., Kimmel, R.: Near field photometric stereo with point light sources. SIAM J. Imaging Sci. **7**, 2732–2770 (2014)
13. Sagawa, R., Sakashita, K., Kasuya, N., Kawasaki, H., Furukawa, R., Yagi, Y.: Grid-based active stereo with single-colored wave pattern for dense one-shot 3D scan. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), pp. 363–370 (2012)
14. Furukawa, R., Masutani, R., Miyazaki, D., Baba, M., Hiura, S., Visentini-Scarzanella, M., Morinaga, H., Kawasaki, H., Sagawa, R.: 2-DOF auto-calibration for a 3D endoscope system based on active stereo. In: IEEE International Engineering in Medicine and Biology Conference (EMBC) (2015)
15. Visentini-Scarzanella, M., Kawasaki, H.: Simultaneous camera, light position and radiant intensity distribution calibration. In: Pacific Rim Symposium on Image and Video Technology (PSIVT) (2015)
16. Kawasaki, H., Masuyama, H., Sagawa, R., Furukawa, R.: Single colour one-shot scan using modified penrose tiling pattern. IET Comput. Vis. **7**, 293–301 (2013)
17. Vogiatzis, G., Hernandez, C.: Self-calibrated, multi-spectral photometric stereo for 3D face capture. Int. J. Comput. Vis. **97**, 91–103 (2012)

Using Shading to Register an Intraoperative CT Scan to a Laparoscopic Image

Sylvain Bernhardt[1,2(✉)], Stéphane A. Nicolau[2], Adrien Bartoli[4], Vincent Agnus[2], Luc Soler[1,3], and Christophe Doignon[2]

[1] IHU, Institut de Chirurgie Guidée par l'Image de Strasbourg, Strasbourg, France
sylvain.bernhardt@ihu-strasbourg.eu
[2] ICube, Université de Strasbourg, Strasbourg, France
[3] IRCAD, Virtual Surg, Strasbourg, France
[4] ALCoV-ISIT, Université d'Auvergne, Clermont-Ferrand, France

Abstract. In abdominal surgery, augmented reality has been attempted by registering preoperative 3D data onto the intraoperative laparoscopic view. The registration may be aided by an interventional 3D imaging system such as a rotational C-arm. It has been shown that one can determine the transformation between an intraoperative 3D volume and the laparoscopic view by letting the laparoscope tip enter the C-arm acquisition field. However, the transformation estimation was up to a 1D rotation and a 2D translation. We propose to complete this registration by using local shading constraints with a piecewise constant albedo hypothesis on the surface of the surgical scene. Thus, the registration becomes fully automatic with no extra apparatus required. Results from experiments on in vivo data show a millimetric registration accuracy.

Keywords: Registration · Abdominal imaging · Minimally invasive procedure · Intraoperative imaging · Endoscopic imaging

1 Introduction

With the advent of minimally invasive surgery and digital endoscopic cameras over the past few decades, intraoperative augmented reality has fostered much research in computer vision [1]. The general goal is to improve the surgeon's perceptions by augmenting the video feedback with a high definition 3D model provided by a preoperative CT or MRI [2–4]. Applications include revealing hidden vessels or tumors. Accurately performing this augmentation remains a challenge as the patient's anatomy may significantly change between the preoperative scanning and the intervention. Notably, in abdominal surgery, the cavity is insufflated with gas which applies pressure on the organs, thus hindering an accurate registration [5]. To compensate for this deformation, a solution may be to introduce a 3D rotational C-arm as an intermediary step in the augmentation process, as this type of apparatus is becoming increasingly popular. Given the non-rigid transformation of the organs of interest between the preoperative and intraoperative 3D scans ([6,7]), all that is left is to determine the relationship between the intraoperative volume and the laparoscopic camera.

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 59–68, 2016.
DOI: 10.1007/978-3-319-29965-5_6

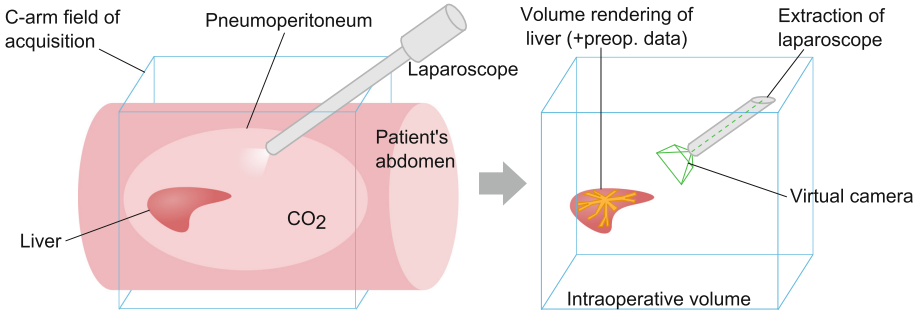


Fig. 1. Context of our method. As shown on the left, the laparoscope purposefully enters the acquisition field of the 3D rotational C-arm. Thus, as depicted on the right, our method allows us to extract the laparoscope from the resulting intraoperative volume, place a virtual camera accordingly and generate a virtual view of the organ of interest (usually with volume rendering) possibly including preoperative data. The laparoscopic image can then be augmented by superimposition with the virtual view.

In [8], we presented a method to tackle this problem without external tracking. First, after a classic camera calibration using a checkerboard, the laparoscope is blocked so that it sees the organ of interest and also enters the acquisition field of the 3D rotational C-arm (Fig. 1). As shown in [8], the metallic presence of the laparoscope does not produce artifacts affecting the region of interest. Then, an intraoperative 3D scan is performed and the laparoscope's body is extracted from the intraoperative volume. This allows us to estimate directly the rigid registration between the laparoscopic camera and the intraoperative 3D imaging system (Artis Zeego, Siemens). This relationship is valid only as long as the camera remains static, which already routinely occurs at several stages of an intervention like a liver segmentectomy.

The method [8] is very appealing but suffers from two main drawbacks. First, due to the tubular shape of the laparoscope, its roll angle cannot be determined from the intraoperative volume. This degree of freedom is estimated thanks to an accelerometer included in the camera, but this is *not* featured in most laparoscopes. Second, [8] assumes that the optical axis coincides with the revolution axis of the laparoscope, which may be violated depending on the model used, as illustrated in Fig. 2(a). This difference results in a 2D shift ϵ in the image plane. Though small at the scale of the device, it can yet result in up to several tens of pixels of registration error in the augmentation. Other parameters such as the zoom and focus also influence the position of the optical axis and thereby ϵ. A calibration dedicated to estimate ϵ is possible, but not relevant before the intervention. Indeed, many endoscopes are separable from the camera (Fig. 2(b)) and the surgeon may make it spin during the intervention to place the light cable upon desire (Fig. 2(c)), which changes ϵ. Likewise, the zoom and focus of the endoscopic camera may also be changed intraoperatively and invalidate the preoperative estimation of ϵ. While performing a supplementary calibration

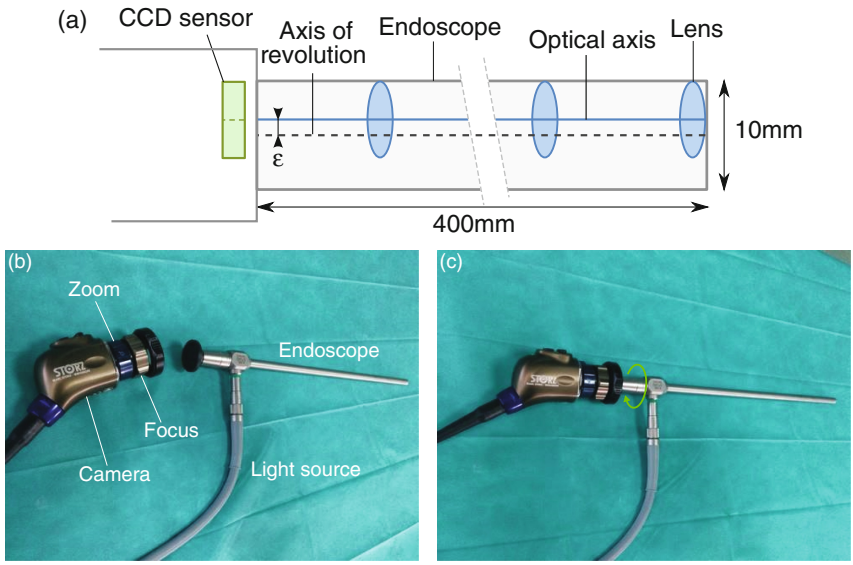


Fig. 2. Anatomy of a laparoscope. (a) A simplified illustration of the structure of a laparoscope presents an exaggerated misalignment ϵ between the optical axis and the axis of revolution. (b) Changing the zoom and focus influence ϵ. Also, the camera and the endoscope are separable. (c) As a result, both can rotate around each other once attached and ϵ varies.

intraoperatively might be feasible, the sake of preserving the workflow compels for a method purely based on image processing.

In this paper, we present a novel method to complete [8]. It solves the previously mentioned registration issues using only information from the intraoperative volume and the laparoscopic image. As discussed, three degrees of freedom are to determine – the roll angle and the translation ϵ along the image axes. We propose to obtain these by optimizing a dissimilarity metric between the laparoscopic image and the view from the virtual camera upon the content of the intraoperative volume (Fig. 1). Given the relatively poor contrast between the different organs in an intraoperative CT image, the surface of the abdominal cavity is one of the most relevant information we could extract from the volume for the virtual camera. Since the cavity is insufflated with carbon dioxide, it presents a good contrast with the surrounding tissues and therefore extracting its surface is trivial, using for instance marching cubes.

Related Work. There are three main ways to register an intraoperative surface to the laparoscopic image. One way is to use Shape-from-Shading (SfS), which reconstructs a surface from a single image based on the pixels' intensity and the reflectance function [9,10]. The reconstructed surface can then be registered to the surface extracted from the intraoperative volume using a method such as Iterative Closest Point (ICP). However, it has been established that SfS should

not be used on its own in laparoscopic surface reconstruction [11], notably due to the falseness of the hypothesis of constant albedo throughout the scene. In our case, SfS would be overachieving since we do not need to reconstruct the surface from the laparoscopic image, but rather to design a dissimilarity metric between the intraoperative volume and the image. This enables us to use a local approach to shading and thus to alleviate the hypothesis of constant albedo (Sect. 2.2). Another means to relate a surface with its image is simply to perform a correlation between their luminance using Mutual Information or an equivalent. However, the surface extracted from the intraoperative volume is textureless. There is thus no color information and approaches based purely on luminance are likely to fail (see Sect. 3 for experimental results supporting this assertion). We also cannot consider methods based on photo-consistency [12,13] which has been successfully applied to endoscopic scenes [14], as two or more images are required.

Our proposed method to complete the registration uses a local formulation of the shading constraints. In the next section, we present the shading model and the formulation of the dissimilarity metric between the two inputs.

2 Methodology

This section describes the shading model used to determine the received light intensity. This model is simple because it is applied locally on the surface and uses piecewise constant albedo and piecewise constant light intensity hypotheses.

2.1 Shading Model

As illustrated by Fig. 3, the only light source inside the abdominal cavity is the one from the laparoscope, modeled as a point light source of position $S \in \mathbb{R}^3$ and intensity $l \in \mathbb{R}$ supposed constant locally. We consider Σ the surface extracted from the intraoperative volume and $\varphi \in C^2(\mathbb{R}^2, \mathbb{R}^3)$ the embedding of Σ which provides the surface point for each pixel $q \in \mathbb{R}^2$ in the laparoscopic image I. φ is known up to the sought pose of the virtual camera. The normal to Σ at φ is given by $\mathcal{N} \in C^2(\mathbb{R}^2, \mathbb{R}^3)$. In a typical laparoscopic image, there are often specularities and poorly lit areas. If we discard those (see Sect. 2.2), it is reasonable to assume that the camera response is linear and therefore a quantity of light k is converted by the sensor into a pixel intensity given by $\tau(k) = ak, a > 0$. The albedo $\zeta \in C^0(\mathbb{R}^2, \mathbb{R})$, or surface reflection coefficient, is supposed constant on the surface locally for a same tissue and therefore $\zeta(q) = b, b > 0$. This is the classic limiting hypothesis in SfS, which we relax in Sect. 2.2.

In a laparoscopic setting, the effect of illumination fall-off may be strong. We model this by dividing the amount of received light by the squared surface-to-light source distance. Assuming S and the origin O coincide, the illumination vector $\mathcal{L} \in C(\mathbb{R}^2, \mathbb{R}^3)$ at φ is thus given by:

$$\mathcal{L} = l \frac{\overrightarrow{\varphi S}}{\|\overrightarrow{\varphi S}\|^2} = l \frac{S - \varphi}{\|S - \varphi\|^2} = -l \frac{\varphi}{\|\varphi\|^2} \tag{1}$$

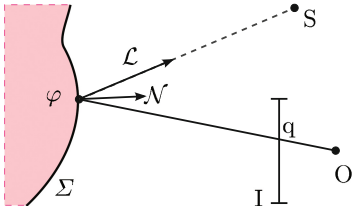


Fig. 3. Shading model. The point light source S emits a ray $-\mathcal{L}$ that hits the surface at φ. Assuming a Lambertian surface, the light is reflected with respect to the normal $\mathcal{N}$ and the illumination vector $\mathcal{L}$. This reflection is projected onto the image plane I at q with O being the optical center and the origin of the world space.

Assuming that the surface is Lambertian, the reflectance $\mathcal{R} \in C^2(\mathbb{R}^2, \mathbb{R})$ is given by $\mathcal{R} = \mathcal{L} \cdot \mathcal{N}$. Finally, using the camera response function τ, the intensity I of a pixel q is predicted by:

$$I = \tau \circ (\zeta \mathcal{R}) = ab(\mathcal{L} \cdot \mathcal{N}) = -c\frac{\varphi^\top \mathcal{N}}{\|\varphi\|^2} \qquad \text{with} \quad c = abl \tag{2}$$

Thus, based on reasonable assumptions about shading in the abdominal cavity, Eq. (2) is a simple solution to relating the surface to the luminance in the laparoscopic image. The coefficient c would ideally be a function of space as both albedo and light intensity vary in the scene. Therefore, we assume c to be constant only locally. The next section explains how this piecewise relationship between the surface and the laparoscopic image can be used in order to determine the three unknown registration degrees of freedom.

2.2 Shading-Based Surface-Image Dissimilarity

Equation (2) is valid for areas in the scene that are not extremely lit (specularities), unlit and for which the albedo is approximately constant. Therefore, we first apply a simple large median filter (23×23) on the 1080p laparoscopic image in order to robustly remove high frequencies (texture and specularities) while preserving the edges. Dark areas are discarded with a simple threshold on luminance. Satisfying the locally constant c requirement is equivalent to locally enforcing constancy for both albedo and intensity.

Therefore, we divide the image into a set $\mathcal{P}$ of homogeneous patches using the watershed algorithm (Fig. 4). The distance between the watershed seeds is related to the size of the image and the kind of its content. In a typical laparoscopic scene filmed at 1080p, the size of the different organs is commonly above 100 pixels, due to the close-up view. Setting the seeds too coarsely would result in missing small structures, while patches not large enough would not contain enough shading information and thus would fail at constraining the dissimilarity measurement. From our experience, a distance between the seeds of 150–200 pixels is ideal for 1080p laparoscopic images.

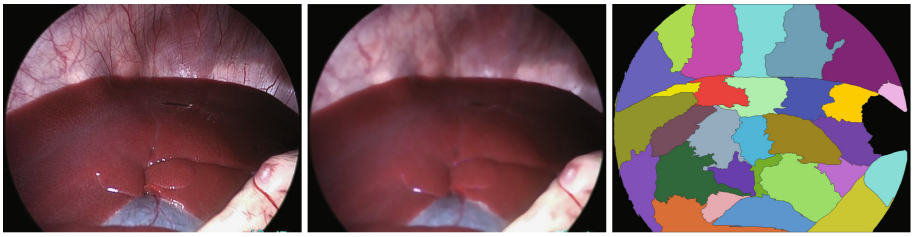


Fig. 4. Image processing and clusterization. The input laparoscopic image is undistorted (left), applied a median filter (middle) and divided into homogeneous patches by watershed (right). Dark areas are discarded (middle right in image).

For each patch $p \in \mathcal{P}$, we use Eq. (2) at each pixel $q \in p$ to estimate c by linear regression. The resulting residuals constitute a least-squares cost function f_p that measures how well the laparoscopic image and the virtual view of the cavity surface concur for a patch p. The variable is the camera pose ω, which affects both φ and $\mathcal{N}$ through the location of the coinciding points O and S.

$$f_p(\omega) = \arg\min_{c \in \mathbb{R}} \sum_{q \in p} \left\| I(q) + c \frac{\varphi_\omega(q)^\top \mathcal{N}_\omega(q)}{\|\varphi_\omega(q)\|^2} \right\|^2 \tag{3}$$

Finally, we obtain the transformation $\hat{\omega}$ composed of the three sought degrees of freedom by minimizing the residuals for each patch $p \in \mathcal{P}$ in the global cost function F:

$$F(\omega) = \sum_{p \in \mathcal{P}} f_p(\omega) = \sum_{p \in \mathcal{P}} \left(\arg\min_{c \in \mathbb{R}} \sum_{q \in p} \left\| I(q) + c \frac{\varphi_\omega(q)^\top \mathcal{N}_\omega(q)}{\|\varphi_\omega(q)\|^2} \right\|^2 \right) \tag{4}$$

We solve $\arg\min_{\omega \in \mathbb{R}^3} F(\omega)$ by using a continuous numerical optimization algorithm (Powell's conjugate direction search in our case). The registration between the laparoscopic image and its virtual equivalent can thus be completed in rotation and translation, allowing an accurate augmentation of the surgical scene.

3 Experiments and Results

In the previous section, we proposed to minimize the cost function (4) in order to accurately register the laparoscopic image and the intraoperative volume. Therefore, the success of our method also depends on the difficulty that optimization algorithms may have to find the global minimum in the search space. A couple of considerations ensure that an initialization at (0,0,0) is close to the global optimum. First, the surgeon is very unlikely to rotate the laparoscope so much that the scene would be upside down. Second, the sensor cannot diverge too much from the laparoscope axis without hindering the completeness of the

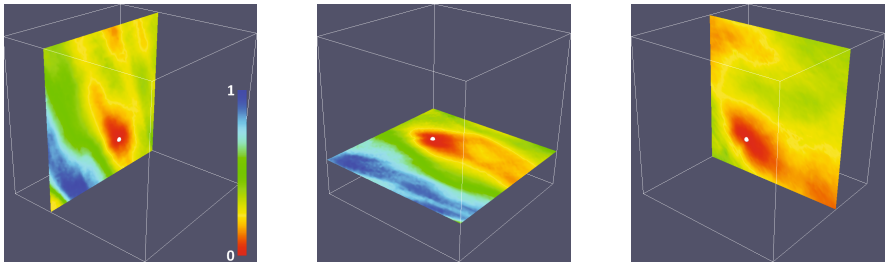


Fig. 5. Example of search space typically ranging $\pm 30^\circ$ and $(\pm 150)^2$ pixels. The cost issued by $F(\omega)$ is here normalized and colored from blue (high) to red (low). Sections are displayed along each of the three dimensions and passing by the global optimum (white dot) (Color figure online).

image captured out of the optics. An example of a clear global optimum in such a 3-dimensional search space around the initialization is illustrated by Fig. 5. These data originate from an *in vivo* acquisition of a pig's liver, for which we applied our method. A total of three different acquisitions on three different pigs were performed. Each time, the intraoperative images were taken during breathhold. Results are displayed in Fig. 6.

For these experiments, one can notice the very good accuracy in registration achieved by our method. Over the three data sets, we performed manual measurements of the Target Visualization Error (TVE) by pointing 15 visual cues such as edges or corners in both images (Table 1). Our method proved to be more than twice as accurate than [8], with an average TVE of 11.3 ± 4.7 pixels in the image. This corresponds to less than a millimeter in the scene at nominal distance (around 70 mm). Thus, the remaining three degrees of freedom are accurately determined and so is the complete relationship between the laparoscopic image and the intraoperative 3D data, without additional apparatus or calibration. Typical optimization computation times range from 15 to 30 s on a standard PC. Added to the initialization, the complete augmentation process takes between 25 to 55 s.

Table 1. TVE (in pixels) manually measured across the three datasets at initialization at (0,0,0), after performing [8] and after the proposed method.

	Initialization	Method from [8]	Proposed method
Case 1	123	13	6
Case 2	59	21	13
Case 3	>300	44	15
Average	>161 ± 124	26 ± 16.1	11.3 ± 4.7

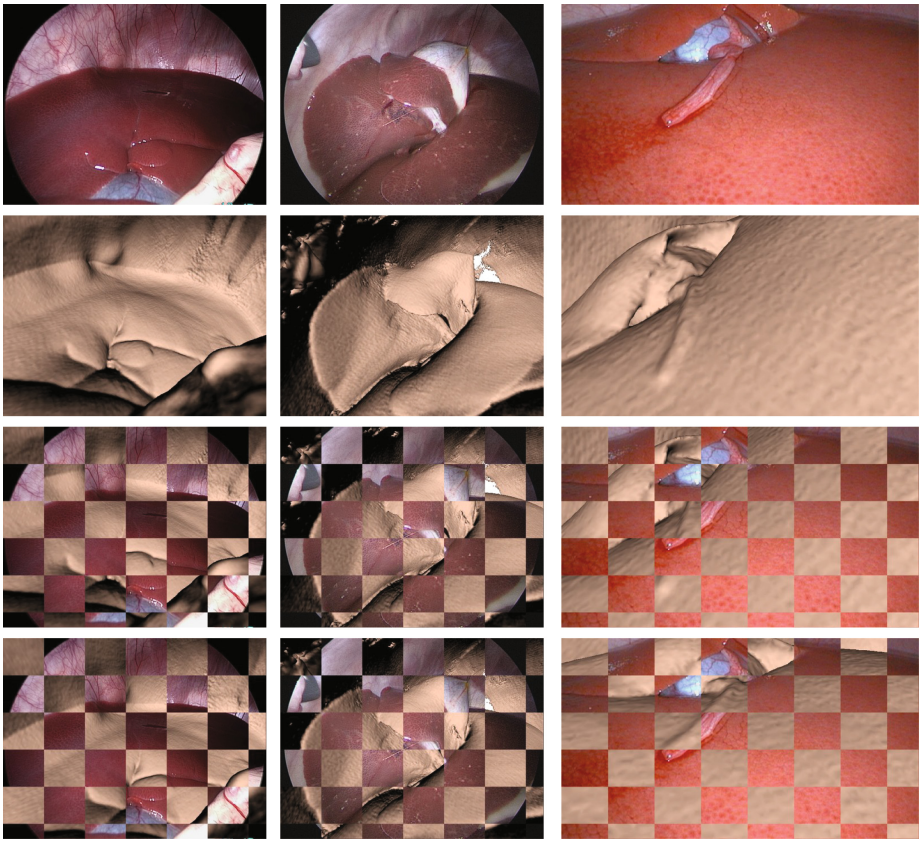


Fig. 6. The laparoscopic image (top) is registered with the view from the virtual camera upon the surface extracted from the intraoperative 3D data and rendered in VTK (middle top). A mosaic of the two shows the alignment before the proposed optimization (middle bottom) and after (bottom).

Finally, in the introduction we asserted that classic 2D image-to-image registration methods such as Mutual Information would fail with such data. For the sake of verification, we calculated for each case the Normalized Mutual Information (NMI) between the endoscopic image and the surface view, while setting the translation to its correct value and varying only the angle. Similarly, to demonstrate the importance of a piecewise approach to shading, we calculated the proposed cost function $F(\omega)$ with globally constant c and piecewise constant c. These three cost functions are compared against each other in Fig. 7. One can notice that NMI does not show a global optimum for any of the three *in vivo* data sets. Moreover, our method with a globally constant c performs well only in Case 3, for which most of the laparoscopic image displays mostly only one organ and thus a same albedo.

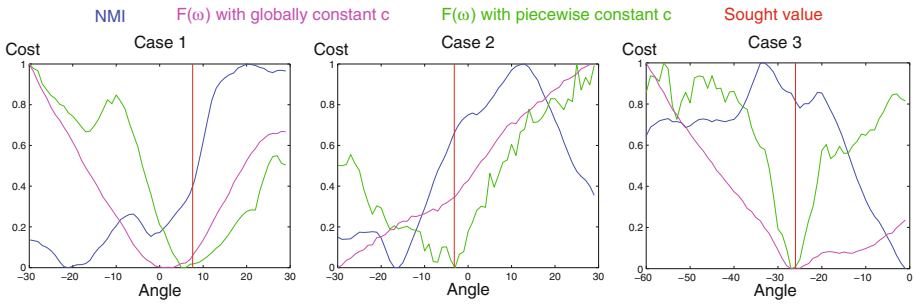


Fig. 7. Display of the normalized cost function in rotation only for NMI (blue), $F(\omega)$ with globally constant c (pink) and $F(\omega)$ with piecewise constant c (green). The graphs show that only our piecewise approach clearly displays a global optimum at the correct angle value (red) in all three cases (Color figure online).

4 Conclusion and Discussion

We have presented a novel method to complete a partial registration between a laparoscopic image and a surface extracted from intraoperative 3D data. When combined with [8], we can provide a millimetric registration between the laparoscopic view and the intraoperative referential frame, using only standard hybrid operating room equipment and requiring no extra calibration process. This facilitates a fast and reliable augmentation of the scene with relevant information coming either from the intraoperative or the preoperative acquisitions.

So, while most shading methods aim at recovering the structure of the scene, we seek the camera pose. Thus, we do Pose-From-Shading rather than Shape-From-Shading. The concept of using shading to estimate the camera pose with respect to a known model is new. Moreover, most existing work on shading assumes a constant albedo over the whole image. It is obviously wrong in a typical intra-abdominal scene where different organs and tissues have different albedo and reflectance. This is why we propose this novel piecewise approach to shading, making it compatible with such scenes.

However, there is still room for improvement. First, the piecewise approach of our method makes it highly parallelizable and a GPU implementation would allow it to reach a shorter processing time. This would make our application more suitable for clinical applications, but also could compensate for breathing if real-time processing is achieved. Second, our approach obviously requires that the laparoscope tip has to show in the intraoperative scan. Although various experiments with surgeons have proved that doing so is not problematic for them, we plan to investigate the possibility of extrapolating our work and determining all the six registration degrees of freedom only from the shading constraints. If not feasible in real time, and for the sake of providing a dynamic augmented reality solution in the hybrid operating rooms, we could also look into updating the augmentation with laparoscope tracking techniques such as SLAM or a robotic arm.

References

1. Sielhorst, T., Feuerstein, M., Navab, N.: Advanced medical displays: a literature review of augmented reality. J. Display Technol. **4**(4), 451–467 (2008)
2. Baumhauer, M., Feuerstein, M., Meinzer, H.-P., Rassweiler, J.: Navigation in endoscopic soft tissue surgery: perspectives and limitations. J. Endourol./Endourological Soc. **22**(4), 751–766 (2008)
3. Nicolau, S.A., Soler, L., Mutter, D., Marescaux, J.: Augmented reality in laparoscopic surgical oncology. Surg. Oncol. **20**(3), 189–201 (2011)
4. Mountney, P., Fallert, J., Nicolau, S., Soler, L., Mewes, P.W.: An augmented reality framework for soft tissue surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 423–431. Springer, Heidelberg (2014)
5. Sánchez-Margallo, F.M., Moyano-Cuevas, J.L., Latorre, R., Maestre, J., et al.: Anatomical changes due to pneumoperitoneum analyzed by MRI: an experimental study in pigs. Surg. Radiol. Anat. **33**(5), 389–396 (2011)
6. Bano, J., Nicolau, S.A., Hostettler, A., Doignon, C., Marescaux, J., Soler, L.: Registration of preoperative liver model for laparoscopic surgery from intraoperative 3D acquisition. In: Liao, H., Linte, C.A., Masamune, K., Peters, T.M., Zheng, G. (eds.) MIAR 2013 and AE-CAI 2013. LNCS, vol. 8090, pp. 201–210. Springer, Heidelberg (2013)
7. Oktay, O., Zhang, L., Mansi, T., Mountney, P., Mewes, P., Nicolau, S., Soler, L., Chefd'hotel, C.: Biomechanically driven registration of pre- to intra-operative 3D images for laparoscopic surgery. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 1–9. Springer, Heidelberg (2013)
8. Bernhardt, S., Nicolau, S.A., Agnus, V., Soler, L., Doignon, C., Marescaux, J.: Automatic detection of endoscope in intraoperative ct image: application to AR guidance in laparoscopic surgery. In: IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 563–567. IEEE (2014)
9. Durou, J.-D., Falcone, M., Sagona, M.: Numerical methods for shape-from-shading: a new survey with benchmarks. Comput. Vis. Image Underst. **109**(1), 22–43 (2008)
10. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., et al.: Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. Med. Image Anal. **17**(8), 974–996 (2013)
11. Collins, T., Bartoli, A.: Towards live monocular 3D laparoscopy using shading and specularity information. In: Abolmaesumi, P., Joskowicz, L., Navab, N., Jannin, P. (eds.) IPCAI 2012. LNCS, vol. 7330, pp. 11–21. Springer, Heidelberg (2012)
12. Clarkson, M.J., Rueckert, D., Hill, D.L.G., Hawkes, D.J.: Using photo-consistency to register 2D optical images of the human face to a 3D surface model. Trans. Pattern Anal. Mach. Intell. **23**(11), 1266–1280 (2001)
13. Jankó, Z., Chetverikov, D.: Photo-consistency based registration of an uncalibrated image pair to a 3D surface model using genetic algorithm. In: Proceedings of 3D Data Processing, Visualization and Transmission, pp. 616–622 (2004)
14. Figl, M., Rueckert, D., Hawkes, D., Casula, R., Hu, M., Pedro, O., Zhang, D.P., et al.: Image guidance for robotic minimally invasive coronary artery bypass. Comput. Med. Imaging Graph. **34**(1), 61–68 (2010)

Surgical Simulation Robot with Haptics and Friction Compensation

Tao Yang[1(✉)], Weimin Huang[1], Kyaw Kyar Toe[1], Jiayin Zhou[1], Yuping Duan[1], Yanling Chi[1], and Loong Ee Loh[2]

[1] Institute for Infocomm Research, Singapore 138632, Singapore
{tyang,wmhuang,kktoe,jzhou,duany,ylchi}@i2r.a-star.edu.sg
[2] School of Mechanical and Aerospace Engineering,
Nanyang Technological University, Singapore 639798, Singapore
leloh1@e.ntu.edu.sg

Abstract. Haptic feedback brings a surgical simulator closer to real surgery. However, friction in surgical simulator's hardware affects its performance significantly. We introduce a surgical simulation robot with roller mechanism for laparoscopic surgical simulation. Roller mechanism is implemented in a constrained space to reduce the friction. Motion based friction cancellation method is also applied to further mitigate the friction effects. Comparing with the same surgical simulation robot without roller mechanism, the one with roller mechanism reduces friction by 32.86 % and 38.87 % on two motion directions, and the motion based friction cancellation method can mitigate the friction effect by 49.46 % and 62.08 % on the two motion directions.

Keywords: Laparoscopic surigcal simulator · Haptics · Friction compensation

1 Introduction

In a laparoscopic surgery, the surgeon has limited access, i.e. visual and haptic only, to the pathological site. The tactile feeling provides the information not only on anatomy, but also on the pathology and the insertion depth of the MIS (Minimally Invasive Surgery) instruments. The tactile information conveys the tool-tissue interaction status to the surgeon through the sense of touch. It always plays an important role in decision making during the surgery [1]. The training instructor also teaches the medical residents to perceive the tactile information during training. Nowadays, as the advent of computer, robotics and virtual reality technologies, various types of simulators and robot assisted devices have been developed for the purpose of laparoscopic surgical training. Most of the surgical simulators or robot assisted surgery tools [2–5] are designed with haptic output capability that enables the system to give the user tactile feelings.

The haptic function built in the surgical simulator or robot is a force output function of the system that simulates the tool-tissue interaction, although there is

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 69–80, 2016.
DOI: 10.1007/978-3-319-29965-5_7

no real tool-tissue interaction under the handheld devices. In our previous work [6], a laparoscopic surgical simulation robot was studied. We applied a semi-spherical mechanism to execute trajectory and haptics for a virtual laparoscopic surgery. There are lots of moving parts contacting with each other in the robot. Hence, friction is inevitable in such systems, which affects the performance of the robot in moving, positioning and torque delivery etc. It needs to be taken care of when considering a stable haptic output, especially for high haptic output at low velocity. Two basic methodologies are commonly applied to deal with friction, i.e. minimize the friction by design and mitigate the friction by compensation. Unfortunately, friction forces are highly non-linear. It is difficult to compensate. Therefore, it is important to reduce the friction forces by designing the system mechanism and apply appropriate compensation technology to mitigate the effect of remaining friction. Friction compensation methods have been studied thoroughly in the past decades [7,8], such as fixed friction compensation, model-based compensation, and neural fuzzy techniques. Neural network method is one of the good methods for friction compensation in practical engineering, as the neural network is capable to handle highly non-linear scenarioes [9]. However, neural network solution is not a physical based method where the parameters do not relate to the physical phenomena directly. Various friction models have been proposed and tested to understand and compensate the frictional force. Most friction models could not match with the real friction scenarios well after a long service period due to wear and tear.

In this paper, a new robot design for laparoscopic surgical simulation is presented. We apply both design and compensation techniques to reduce friction and mitigate its effects respectively. Friction forces in between the moving parts are reduced by introducing teh roller mechanisms. A motion based friction cancellation method with friction model is applied to mitigate the friction effect for stable haptic output. The paper is organized as follow: Sect. 2 describes the low friction design, finite element analysis and system modelling of the semi-spherical mechanism for the surgical simulation robot. Section 3 describes the motion based friction compensation method and its application on the robot. The work is concluded in Sect. 4.

2 Surgical Simulation Robot

2.1 Friction in Haptics

A surgical simulation robot, as shown in Fig. 1 was designed for image guided robot assisted surgical training in our previous work [6]. The replicated surgical tools are driven by the robot to allow the user to operate on virtual patient with haptic feedback, and provide haptic guidance for surgical training purpose as well. Each of the replicated laparoscopic surgical tools has five Degree-of-Freedom (DOF), namely pitch, yaw, translation, roll and handle grasping. It mimics the DOF of surgical tools in real laparoscopic surgery. A semi-spherical mechanism is the major component to achieve the DOFs mentioned above [6]. Friction in the semi-spherical mechanism affects the performances of haptics, especially the friction on pitch and yaw axes.

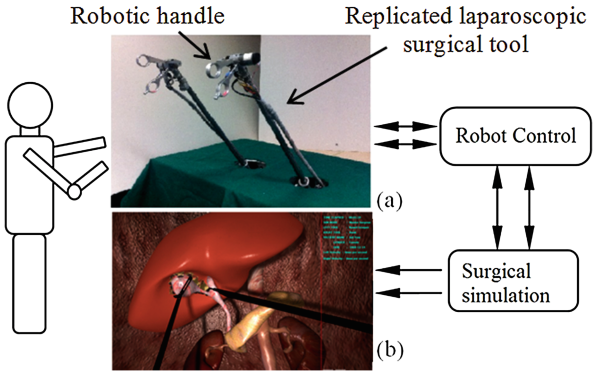


Fig. 1. Overview of the surgical training system: (a) robotic surgical trainer and (b) virtual surgical simulation platform.

As experienced in our robot, friction is inevitable, and it is highly non-linear and difficult to model. It can be categorized as two basic categories, the rolling friction and the sliding friction. The rolling friction is expressed as

$$F_r = C_{rr} F_N, \tag{1}$$

and the sliding friction is expressed as

$$F_s = \mu F_N, \tag{2}$$

where C_{rr} is the rolling resistance coefficient which depends on material elasticity, μ is the sliding friction coefficient which depends on material pair and surface condition. μ is usually much larger than C_{rr}, F_N is the normal force acting on the contact surface. In a haptic device, it can be expressed as a function of haptic output.

2.2 Design Considerations

We introduce a bearing-like mechanism to create rolling motion on the semi-spherical mechanism to reduce friction and enhance the haptic performance. Figure 2(a) shows the overall design of a semi-spherical mechanism with rollers that reduces the frictional force. The semi-spherical mechanism can be divided into four parts as shown in Fig. 2(b). Part I and Part II contain guiding blocks where the rollers are hosted. Part III includes two arches clamped in between of Part I and Part II. Part IV applies and maintains appropriate pressure in between of Parts I and III, Parts II and III. Hence, the mechanical gap between the rollers and Part III, and the motion precision of Part I could be controlled.

Rollers were placed at all possible places, as shown in Fig. 2(a) where relative motion exists. Due to space constraint, the rollers were supported by bushings instead of ball bearings. The relative motion between the roller and its hosting bush still introduces sliding friction. Polytetrafluoroethylene (PTFE) was

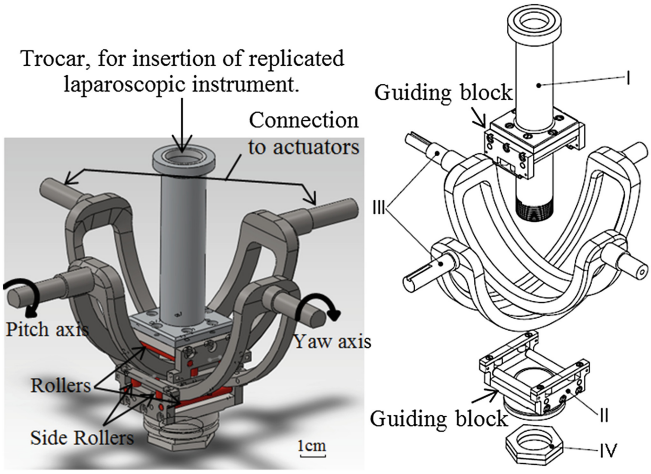


Fig. 2. (a) Overall view of semi-spherical mechanism that provides lower friction for haptic output. (b) Four major components. I: upper guiding block with rollers, II: lower guiding block with rollers. III: two arches which work as haptic input/output interface for pitch and yaw axes. IV: locking nuts for adjusting proper pressure between part I, II and III.

selected to work as bushing for its low friction coefficient and high wear resistance. Although PTFE has very good wearing resistance, it would still be worn off and the size of the hole will be changed where the roller is hosted. However, the contacting profile between the roller and the bushing would not be altered significantly, and hence the friction profile. With this design, wear and tear on the rollers and their contacting surfaces are minimized. The contact profile between the roller and the contact surface could be maintained consistent even after a long service period.

2.3 FE Analysis

The finite element analysis by Abaqus/Explicit 6.13 was adopted to investigate the stress distribution of the semi-spherical mechanism design under loading along pitch and yaw axes (x and y directions in Cartesian coordinate, as shown in Fig. 3). Fifty Newton was applied at a reference point (RP) which is on the trocar's longitudinal axis and 300 mm (equivalent to 15 Nm) above rotational origin of the axis. The translational DOFs of each node on the trocar were coupled with those of the RP to simulate a surgical tool passing through the trocar. Quasi-static loading procedure was applied in the analysis. The two arches were modeled as rigid body and were fully fixed at their reference points as shown in Fig. 3.

The material properties of the components in the mechanism are listed in Table 1. The friction coefficient was 0.16 between steels and 0.05 between steel

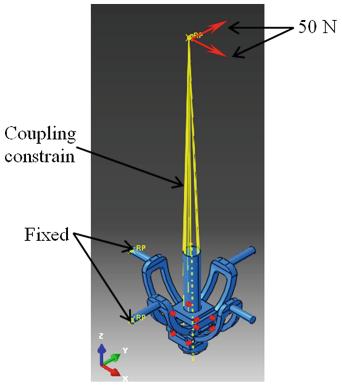


Fig. 3. FE modelling of the semi-spherical mechanism.

Table 1. Material properties used in FE model

	Density(kg/m^3)	Young's modulus(GPa)	Poisson's ratio
Stainless steel	7850	200	0.3
PTFE	2600	0.55	0.46

and PTFE. The trocar was meshed with shell element S4R and the rest components were meshed with solid element C3D4.

Under loading along yaw axis, the maximum Von Mises stress is around 15 MPa at the roller, as shown in Fig. 4. When loading along pitch axis, the rollers also have similar stress distribution with a maximum stress around 12 MPa. Under loading along pitch axis, the maximum Von Mises stress, around 70 MPa, occurs at two side rollers. Under both pitch and yaw loadings, the two guiding blocks of the rollers have small stress level. Figure 5 shows the stress distribution of the guiding block under loading along yaw axis, the Von Mises stress is around 1–2 MPa for the shaft of the side rollers.

The simulation results show that the maximum stress is far below the yield stress of the stainless steel when it is loaded at 15 Nm torque. It suggests that the strength of the semi-spherical mechanism is sufficient for haptic output. The semi-spherical mechanism was fabricated as shown in Fig. 6.

2.4 System Modelling

The semi-spherical mechanism was installed in an existing robot control system [6] to connect with actuators and inserted with a replicated laparoscopic surgical tool. A frequency response experiment was conducted to measure the system response. We assume that the data acquisition speed of a force sensor in the haptic feedback loop is infinitely high, hence the haptic output force is proportional to the acceleration of the surgical tool by $F = ma$, where m is a

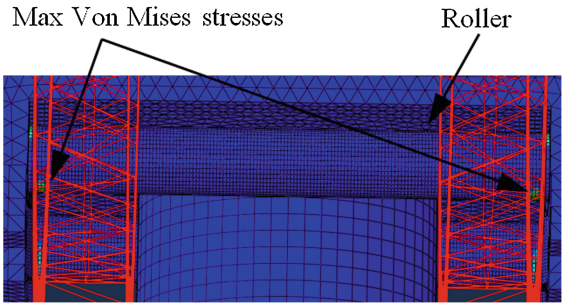


Fig. 4. Locations of maximum Von Mises stresses when the mechanism is under loading along yaw axis.

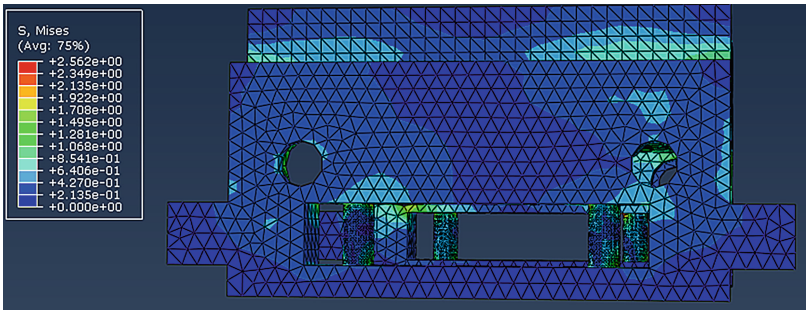


Fig. 5. Stress distribution of the guiding block under loading along yaw axis.

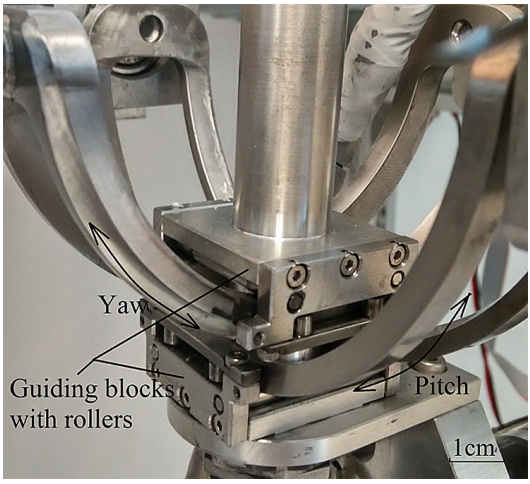


Fig. 6. Fabricated semi-spherical mechanism with rollers, and relative velocity at the contacting area of each axis.

mass constant. Therefore we can use the transfer function of system acceleration to represent the transfer function in haptic force. The transfer function for the robot can be obtained by dividing the measured acceleration $\alpha_m$with the commanded acceleration α_c,

$$G = \frac{\alpha_m}{\alpha_c}. \tag{3}$$

Sinusoid signal with frequency up to 20 Hz was input to the robot. Figure 7 shows the bode plot obtained from the experiment data.

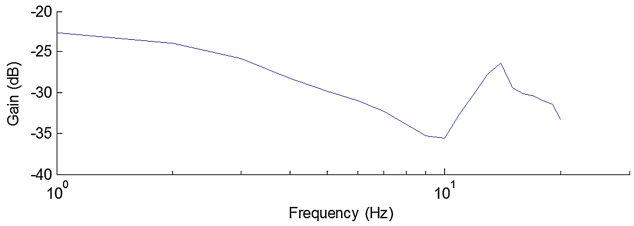


Fig. 7. Open loop bode plot of the semi-spherical mechanism with a replicated surgical tool.

The robot system was taken as second order system. Matlab system identification toolbox was applied to model the system based on the open loop bode plot shown in Fig. 7. The system transfer function $G(s)$ is estimated as

$$G(s) = \frac{0.03s + 0.06}{s^2 + 0.18s + 0.0064}. \tag{4}$$

3 Friction Control Model for Haptics

Despite of the design considerations for friction reduction, friction compensation is still required as high haptic output will lead to high frictional force between the moving parts. The resultant frictional force in the design is a combination of sliding friction from the bushings and rolling friction from the rollers. Hence, stribeck phenomena would affect the performance of haptic output, especially when the haptic output force is large and moving velocity is low.

Experiments were conducted to measure the frictional force with respect to the velocity and haptic output. The robot was set to output a series of haptic force exerting on a user. The haptic output was set from 1 N to 7 N with 1 N increment for each experiment. The user pushed the robotic handle (as shown in Fig. 1) to move against the direction of haptic output. The guiding block with rollers moved from one end of the arch to the other end as shown in Fig. 6. The force applied to execute the motion was measured while the robotic handle was moving. Frictional force was obtained by subtracting the desired haptic output from the measurement. This procedure was repeated 50 times at each haptic

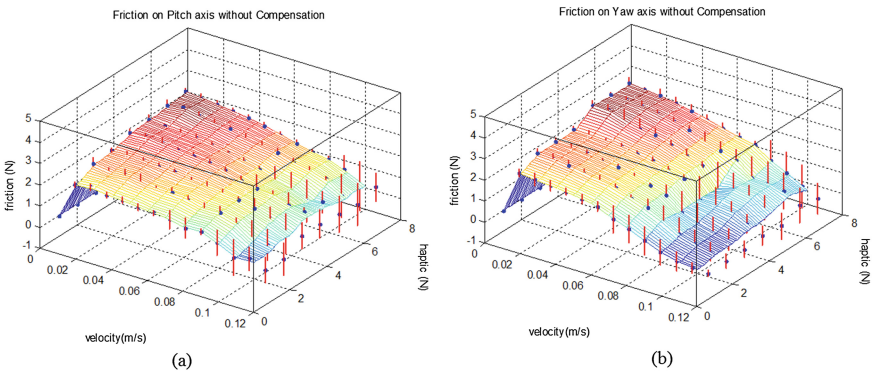


Fig. 8. Mean frictional force from current design with haptic output from 1 N to 7N. The frictional force is larger when the components are just to move, and it is reduced significantly and tends to stabilize when the components moving at higher velocity. The frictional forces are generally higher when the robot outputs a higher haptic force. Vertical bars are the standard deviations at the specific velocity and haptic output. (a) Frictional force for pitch axis. (b) Frictional force for yaw axis.

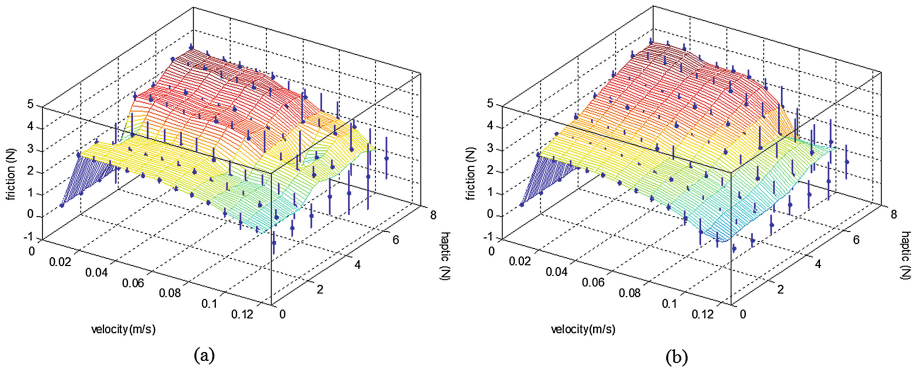


Fig. 9. Mean frictional force measured from the design in [6]. Vertical bars are the standard deviations at the specific velocity and haptic output. (a) Frictional force for pitch axis. (b) Frictional force for yaw axis.

level. The velocity span covered from 0 to 0.125 m/s. The maximum velocity in the experiment was relatively low. Therefore, viscous friction was not taken into consideration during modelling.

Figure 8 shows the measured frictional force on both moving axis. The overall haptic output is smooth, and the maximum friction forces are 2.79 N and 2.98 N for pitch and yaw axes respectively. These measurements will be used in fitting with friction model for compensation. The same experiment was conducted on our previous design in [6] to measure the frictional force. The design has similar overall structure, but no roller mechanism. All moving components create sliding friction. The measured friction forces are shown in Fig. 9. The maximally friction

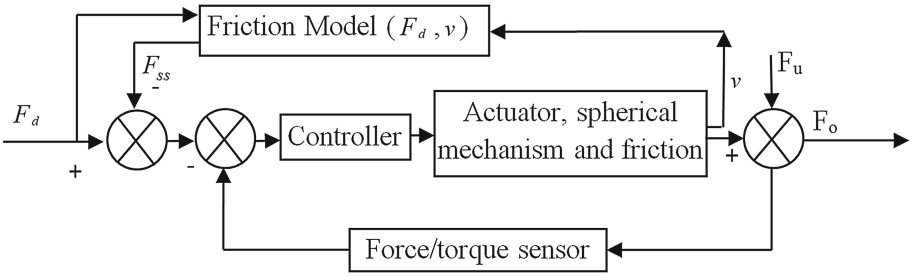


Fig. 10. Control diagram for friction compensation and haptic output. F_d is the haptic output reference force or the desired haptic output, F_o is the haptic output force, F_u is the user's interaction force.

forces are 3.96 N and 4.13 N for the pitch and yaw axes respectively. Comparing Figs. 8 and 9, we notice that the overall friction forces with the design are reduced by 32.86 % and 38.87 % on pitch and yaw axes respectively when comparing with the mechanism without rollers in the same velocity and haptic output span.

Here, we applied a motion based friction cancellation method to compensate the effect of friction and the stribeck phenomena for stable haptic output. The control diagram of such haptic output system is shown in Fig. 10.

Various friction models have been proposed by researchers [10]. A basic friction model was employed in this study. The friction model is written as

$$F_{ss} = (F_c + (F_s - F_c)e^{-(v/v_s)^2})sgn(v), \tag{5}$$

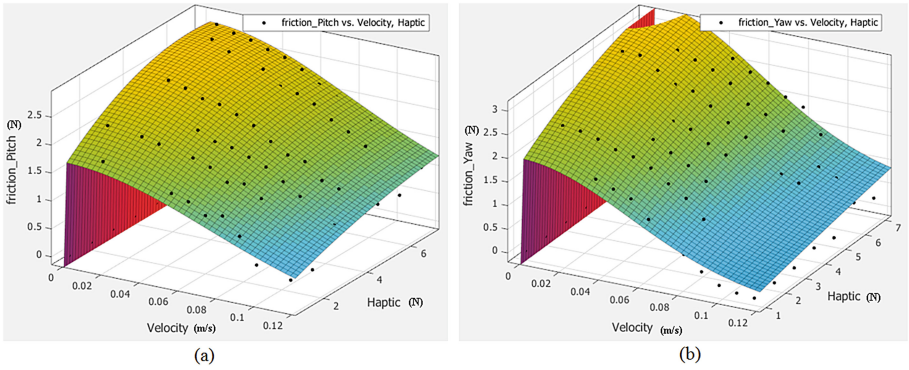


Fig. 11. Surface fitting result with Eq. 5. Experimental results shown in Fig. 8 were fitted with Eq. 5 using Matlab curve fitting toolbox. Black dots are the down sampled experimental measurements. (a) Frictional force fitting for pitch axis with $\mu_e = 0.08$, $a_2 = -0.032$, $a_1 = 0.403$, $a_0 = 1.476$. (b) Frictional force fitting for yaw axis with $\mu_e = 0.086$, $a_2 = -0.019$, $a_1 = 0.351$, $a_0 = 1.82$.

where F_{ss} is the steady state friction, F_c is the Coulomb frictional force, F_s is the stribeck force, v_s is the relative velocity at stribeck, v and is the relative velocity of two moving components. F_c and F_s are dependent on the magnitude of haptic output. They can be written as a function of the desired haptic output F_d, i.e. $F_c = f_c(F_d)$, $F_s = f_s(F_d)$. The functions need to be determined experimentally as different system configuration results in different friction profile. For the semi-spherical mechanism with rollers presented in this paper, the Coulomb frictional force was taken as

$$F_c = \mu_e F_d, \tag{6}$$

where $\mu_e$is an equivalent friction coefficient for the system. A second order polynomial function is taken to represent the stribeck force as

$$F_s = a_2 F_d^2 + a_1 F_d + a_0 \tag{7}$$

Curve fitting was applied on the experiment data (shown in Fig. 8) to identify the parameters in Eqs. (5), (6) and (7). Figure 11 and Table 2 show the surface fitting results and the estimated parameters.

The motion based cancellation method was tested by the same experiment method described in the beginning of Sect. 3. Figure 12 shows the mean frictional force measured from pitch and yaw axes. Comparing with frictional force shown in Fig. 8 in which has no compensation, the frictional force and the stribeck

Table 2. Frictional force fitting results with Eq. (5).

	R^2	Adjusted R^2	RMSE
Pitch	96.75 %	96.65 %	0.14
Yaw	93.92 %	93.72 %	0.24

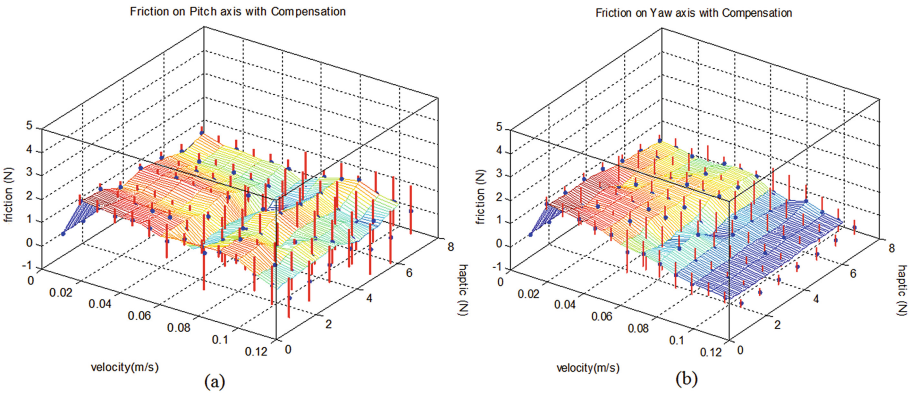


Fig. 12. Mean frictional force measured with friction compensation. Vertical bars are the standard deviations at the specific velocity and haptic output. (a) Frictional force for pitch axis. (b) Frictional force for yaw axis.

phenomena have been mitigated significantly. The total volume covered under the surface (Fig. 12) were reduced by 49.46 % and 62.08 % for pitch and yaw axes respectively. The max measured friction is about 1 N whereas it could reach to 2.5 N when there is no compensation. Figure 12 suggests that the motion based cancellation method is able to work well with a wide range of velocity on our designed mechanism, from 0.02 m/s to 0.12 m/s. It is noticed that the standard derivations increase slightly when the velocity is higher. Comprehensive system model and friction model are required to further improve the performance of this motion based cancellation method.

4 Discussion and Conclusion

A design for laparoscopic surgical simulation robot was presented in this work. Design consideration, FEM verification, system modelling, friction identification and compensation were studied. The semi-spherical mechanism with rollers reduces frictional force significantly comparing to the mechanism without rollers. The motion based cancellation method is capable to mitigate the effect of friction well. Although the frictional force has been mitigated significantly, some residual friction forces are still not removed. It is due to the limitation of motion based cancellation method, which needs a velocity input before the friction model can estimate the frictional force for compensation, but the frictional force is already there before velocity is detected. The compensation is therefore always delayed. Advanced friction compensation methods thus need to be explored to further improve the performance.

Acknowledgments. This work is supported by Agency for Science, Technology and Research, Singapore.

References

1. Ottermo, M.V., Ovstedal, M., Lango, T., Stavdahl, O., Yavuz, Y., Johansen, T.A., Marvik, R.: The role of tactile feedback in laparoscopic surgery. Surg. Laparosc. Endosc. Percutan. Tech. **16**, 390–400 (2006)
2. Hyosig, K., Wen, J.T.: Robotic assistants aid surgeons during minimally invasive procedures. IEEE Eng. Med. Biol. Mag. **20**, 94–104 (2001)
3. van den Bedem, L.J.M., Hendrix, R., Naus, G.J.L., van der Aalst, R., Rosielle, P.C.J.N., Nijmeijer, H., Maessen, J.G., Broeders, I.A.M.J., Steinbuch, M.: Sofie, a robotic system for minimally invasive surgery. In: The 6th International MIRA Congress, Athens, Greece, p. 056 (2011)
4. van den Bedem, L.J.M.: Realization of a demonstrator slave for robotic minimally invasive surgery/door Linda Jacoba Martina van den Bedem. In: Department of Mechanical Engineering. Doctoral degree Technische Universiteit Eindhoven, p. 199 (2010)
5. Simbionix (2014). http://simbionix.com/simulators/lap-mentor/

6. Yang, T., Liu, J., Huang, W., Su, Y., Yang, L., Chui, C.K., Ang Jr., M.H., Chang, S.K.Y.: Mechanism of a learning robot manipulator for laparoscopic surgical training. In: Lee, S., Cho, H., Yoon, K.-J., Lee, J. (eds.) Intelligent Autonomous Systems 12. AISC, vol. 194, pp. 17–26. Springer, Heidelberg (2013)
7. Olsson, H., Astrom, K.J., Canudas de Wit, C., Gafvert, M., Lischinsky, P.: Friction models and friction compensation. Eur. J. Control **4**, 176–195 (1998)
8. Bona, B., Indri, M.: Friction compensation in robotics: an overview. In: 44th IEEE Conference on Decision and Control, and 2005 European Control Conference, CDC-ECC 2005, pp. 4360–4367 (2005)
9. Nguyen, D.H., Widrow, B.: Neural networks for self-learning control systems. IEEE Control Syst. Mag. **10**, 18–23 (1990)
10. Armstrong-Hlouvry, B., Dupont, P., De Wit, C.C.: A survey of models, analysis tools and compensation methods for the control of machines with friction. Automatica **30**, 1083–1138 (1994)

A Real-Time Target Tracking Algorithm for a Robotic Flexible Endoscopy Platform

Nanda van der Stap[1(✉)], Luuk Voskuilen[1], Guido de Jong[1], Hendrikus J.M. Pullens[2], Matthijs P. Schwartz[2], Ivo Broeders[3], and Ferdi van der Heijden[1]

[1] MIRA Institute for BMT and TM, Carré 3.526, PO box 217, 7500 AE Enschede, The Netherlands
{n.stap,f.vanderheijden}@utwente.nl
[2] Department of Gastroenterology and Hepatology, Meander Medical Center, Amersfoort, The Netherlands
[3] Department of Surgery, Meander Medical Center, PO Box 1502, 3800 BM Amersfoort, The Netherlands

Abstract. Complex endoscopic interventions require a new generation of devices and instruments. A robotic platform for flexible endoscopy through telemanipulation was developed to meet this demand. The concept of telemanipulation allows the development of software for computer-aided surgery. Intelligent navigation such as automated target centralization could assist the endoscopist during procedures.

A real-time algorithm was designed for tracking a target region that is of specific interest for the surgeon. Therefore, the physician needs to indicate the region to be tracked, which then will be centralized (locked). The goal of this research is to investigate the robustness and accuracy of the tracking algorithm during endoscopic interventions. The region of interest can be a polyp for polypectomy, Vater's ampulla for Endoscopic Retrograde CholangioPancreatography (ERCP), Barrett's epithelia for gastroscopic biopsy or any area in more complex procedures. The algorithm was tested in vitro on image sequences obtained during real endoscopic interventions.

The indicated area of interest could be tracked in all image sequences, with an accuracy of 91.6 % (Q1–Q3 77.7 %–99.0 %, intraclass correlation). The algorithm was robust against instruments or smoke in the field of view. Tracking was less robust against very large camera movements.

The developed target lock worked robustly, in real-time and was found to be accurate. Improvements include improving the robustness of the algorithm against motion blur and drift.

1 Introduction

A trend towards minimizing the invasiveness of surgical procedures exists. Instead of open surgery, endoscopic or keyhole surgery is more often performed. Endoscopic surgery, where rigid cameras or endoscopes are used, decreases the amount of scarring and blood loss in the patient, leading to less pain and faster

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 81–89, 2016.
DOI: 10.1007/978-3-319-29965-5_8

recovery times. With the transition from open to endoscopic surgery, flexible endoscopes are looked at for more complex procedures as well. Originally, only diagnostics and small therapeutic interventions (the removal of polyps in the colon for example) were performed with flexible endoscopes. Nowadays, increasingly large tumors are being removed, with the aim to spare the patient from a more invasive surgical procedure.

However, flexible endoscopes have vital drawbacks that make them difficult to handle, especially during complex procedures [1]. Endoscope handling is not intuitive and far from ergonomic. The instruments that can be inserted through the working channel of the endoscope are not capable of triangulation. Mostly, only one instrument channel is present, causing many instrument changes and a significant loss of dexterity.

Robotization is thought to improve the handling properties and dexterity of flexible endoscopes. Additionally, navigation can be automated (e.g. [2–4]). Our research is specifically aimed at automating endoscope navigation and fixation during interventions. Ultimate aim is to use image-based control to correct the endoscope tip once the endoscope is at the intervention site. The endoscopist can indicate a focus area manually, and the system described here will track this area continuously in the image. The tracked area will be kept in the center of the screen as much as possible, resulting in a so-called 'target lock'. The target should be centralized despite movement of the endoscope or the environment.

Others have investigated re-targeting of endoscopes. A useful application is for instance the optical biopsy, a visualization of cellular structures using optical instruments in the working channel of an endoscope. SLAM (Simultaneous Localization and Mapping) techniques combined with probe tracking, video manifolds for the patient-specific clustering of images and epipolar geometry recovery are examples of solutions for the retargeting problem in this application [5–8]. Chu et al. describe a flexible tip for a rigid endoscope and target tracking, but it is unclear which tracking approach they use [9].

We are interested in developing a clinically successful target locking system for robotized flexible endoscopy. This leads to the following requirements:

1. Real clinical added value: the procedure of interest should not be hindered or take longer due to the automation.
2. Robust to low texture frames, large movements, varying illumination conditions, instrument interference and occlusions (fluids, tissue deformation).
3. Accurate enough to enable small tip corrections in the order of a few millimeters.
4. Real-time functionality, such that the endoscopist can direct all his/her attention towards the surgical target instead of controlling the camera.
5. Easy correction functionality: if the target changes, for instance if tissue is taken from it, it should be easy to re-localize the target region.

Feedback for the control will be obtained from the images by visual motion tracking and correcting for this. Motion tracking has been employed in flexible endoscopy [2]. A key issue is robustness and accuracy of tracking, implying an accurate outlier detection mechanism. Visual motion tracking is challenging due

to the nature of the images (often low in texture and suffering from the artifacts named above in 2.). There is a trade-off between accuracy and computational effort with feature tracking algorithms. For our application, the algorithm must run fast enough to accurately correct the tip before the endoscopic image has changed too much.

The contribution of this paper lies in the real-time, accurate, feature-based target tracking aimed at optimal system performance with real clinical value. Therefore, real clinical data is used for evaluation. The described application is developed for the robotized flexible endoscopy system Teleflex [10], but minor changes can make it suitable for other robotized endoscopes.

2 Materials and Methods

System requirements were established in close collaboration with several expert endoscopists. A thorough clinical evaluation among the endoscopists led to the conclusion that routine colonoscopies, ERCPs and EMRs (Endoscopic Mucosal Resections) were the interventions most likely to benefit from a target lock. These procedures are known for their clinical complexity with respect to specific sub-interventions. The most difficult part during an ERCP is the insertion of a probe in Vater's papilla. Once the papilla is in the proper position in the view, the endoscope should remain fixed so that the probe can be manipulated properly. Similar situations were indicated for colonoscopies and EMRs. Twelve image sequences were selected and contained the various sub-interventions of interest for the target lock (Table 1). A sub-intervention was estimated to last for 4 s on average; this is therefore the length of the sequences. An exception forms the papilla insertion during an ERCP. Therefore, the length of sequence 3 was doubled.

To use images as control feedback, the *bandwidth of the motion* should be an order of magnitude less than the *sample frequency*. The Nyquist frequency for 25 fps is 12.5 Hz. For control purposes, the rule of thumb is to have a 5–10 times higher sample frequency than the motion bandwidth. In this case, most motions have a frequency of 0–5 Hz. With an effective frame rate (sample frequency) of 24–25 fps this requirement was met in our system.

Real-time optimization can be done by cropping or down-scaling of the images, frame-skipping, code- and platform-optimizations or heterogeneous computing. The latter will result in the most significant improvement without data loss. To enable heterogeneous computing, the feature tracking algorithms were implemented using OpenCV with OpenCL.

Image sequences had a resolution of 768×576 and a frame rate of 25 frames per second (fps). All results were generated using a HP Elitebook 8570 w mobile workstation running on a 64 bit operating system (Windows 8.1) with an Intel Core i7-3630QM processor, 8 GB DDR3 RAM and an AMD FirePro M4000 graphics card. Programming was done using Microsoft Visual Studio Express 2013, with libraries from OpenCV 2.4.9 and OpenCL 1.1. A colonoscope, an ERCP-scope and a pediatric colonoscope (for EMR) were used to record the procedures. The properties of each of them are listed in Table 2.

Table 1. Image sequence properties. APC: Argon Plasma Coagulation. Note: in all sequences, instruments are present in the field of view (FOV).

Number	Procedure	Intervention	Target	Disturbing factor
1	Colonoscopy	Polypectomy	Polyp	Coarse movements
2	Colonoscopy	Polypectomy	Polyp	Poor illumination; Occlusions
3	ERCP	Cannulation	Vater's papilla	Target near edge of the FOV; Occlusions
4	ERCP	Sphincterotomy	Vater's papilla	Target near edge of FOV; Smoke; Occlusions
5	EMR	Injection	Primary tumor	Large target; Color change due to dye injection
6	EMR	APC	Residual lesion	Large area of removed mucosa; Small target; Sparks
7	Colonoscopy	Injection	Polyp	Color change due to dye
8	Colonoscopy	Polypectomy	Polyp	Large polyp; Dirt on lens
9	ERCP	Cannulation	Vaters ampulla	Endoscope motion
10	ERCP	Stent removal	Vaters ampulla	Multiple instruments in view
11	EMR	Partial resection	Primary tumour	Large target; Coarse movements; Dye injection
12	Colonoscopy	APC	Polyp	Sparks; Bubbles; Small target

Table 2. Properties of each endoscope used to record the image sequences. FOV: Field of View. DoF: Depth of Field. DoV: Direction of View.

Procedure	Endoscope	Properties
Colonoscopy	Olympus CF-H180AL	FOV: 170°; DoF: 2–100 mm; Length: 1680 mm
ERCP	Olympus TJF-160VR	FOV: 100°; DoF: 5–60 mm; Length: 1240 mm
EMR	Olympus PCF-PH190L	FOV: 140°; DoF: 2–100 mm; Length: 1680 mm

2.1 Algorithm

For accurate and robust feature tracking, SIFT (Scale Invariant Feature Transform [11]) will be suitable, because blob-like features are abundantly present in the image sequences that were used (Fig. 1). However, detecting and matching

these features takes considerable computational effort. For our real-time application, we therefore chose to use SURF features (Speeded-Up Robust Features [12]). These are nearly as accurate as SIFT features, but decrease computational effort considerably [13].

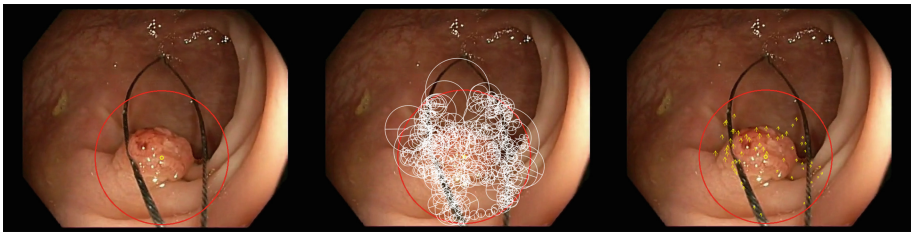


Fig. 1. Example image of polypectomy with polyp and instrument present. Left: scene with ROI indication. Middle: SURF features withing ROI (indicated as white circles). Right: final motion vectors (indicated as arrows).

As stated in the Introduction, an accurate outlier detection mechanism is key to robust system performance. In the original SURF algorithm brute force matching was selected for accuracy reasons. In an attempt to reduce computational time and improve reliability of the matches, we added the second-to-nearest-neighbor (SNN) distance ratio check, as proposed by Lowe [11]. To increase the number of features, even in low-textured areas, preprocessing (grayscale conversion and histogram equalization) was applied on the images. Outlier removal based on vector magnitude was added to the algorithm to improve robustness. The complete algorithm works as follows:

1. Initialization:
   (a) Acquire and visualize the first image, reference image I_{ref}.
   (b) Select the target ROI in I_{ref}: a circle with position $\mathbf{x}_{ref}$ and radius R.
   (c) Set $\mathbf{x}_{target} = \mathbf{x}_{ref}$.
   (d) $\{\mathbf{y}_{ref}(n), \mathbf{f}_{ref}(n)\} = \text{get\_surf\_features}(I_{ref}, \mathbf{x}_{ref})$.
2. Acquire current image I_{cur}.
3. $\{\mathbf{y}_{cur}(m), \mathbf{f}_{cur}(m)\} = \text{get\_surf\_features}(I_{cur}, \mathbf{x}_{target})$.
4. Match $\{\mathbf{f}_{ref}(n)\}$ to $\{\mathbf{f}_{cur}(m)\}$ with SNN distance ratio check, yielding matched indices $\{n(k), m(k)\}$, with $k = 1, ..., K$.
5. Get displacement vectors $\{\mathbf{d}(k) = \mathbf{y}_{cur}(m(k)) - \mathbf{y}_{ref}(n(k))\}$
6. Remove outliers. Condition: $\|\mathbf{d}(k)\| > 2 * median(\{\|\mathbf{d}(k)\|\})$.
7. $\mathbf{x}_{target} = \mathbf{x}_{ref} + mean(\{\mathbf{d}(k)\})$.
8. Repeat till end from 2.

Procedure $\{\mathbf{y}(n), \mathbf{f}(n)\} = \text{get\_surf\_features}(I, \mathbf{x})$

1. Convert image I from RGB to grayscale.
2. Apply histogram equalization to I to increase feature number.
3. Detect SURF key point positions $\{\mathbf{y}(n)\}$ and key point descriptors $\{\mathbf{f}(n)\}$.

2.2 Analysis

Targets to be tracked (Table 1) were manually annotated throughout the image sequences by an expert interventional endoscopist (>2000 endoscopies). The automatically found location was compared for accuracy to the manual results using intra-class correlation analysis (ICC, [14]). Tracking error was given by the Root Mean Square Error (RMSE) of the distance in pixels between the two targets. Computational times were recorded to measure real-time performance of the system. Robustness was measured by counting the number of feature matches and inlying matches per frame.

3 Results

In all sequences, the manually indicated target could be tracked with an high accuracy of 91.6 % (Q1–Q3 77.7 %–99.0 %, see Table 3). Sequence 7 had the best tracking results with a correlation of 99.9 % and a RMSE of 9.4 pixels. Median tracking error was 38.9 pixels (Q1–Q3: 28.7–76.3, see Table 3). The algorithm proves robust against smoke, fluid and instrument interference, color changes, occlusions and poor illumination. Sequences that suffered from these artifacts nonetheless led to the best results (Fig. 2). The text boxes and show that large motions and motion blur are the cause for the biggest tracking errors.

Table 3. Results per image sequence.

Sequence	ICC (%)	Median matches	Median inliers	RMSE (pixels)
1	81.9	77	34	149.1
2	89.4	165	56	32.4
3	99.7	154	67	25.5
4	98.8	123	32	34.2
5	96.6	201	73	27.4
6	93.8	239	73	28.0
7	99.9	184	62	9.4
8	73.6	205	61	116.0
9	99.8	88	32	16.6
10	75.7	236	38	43.7
11	78.4	131	24	123.0
12	63.5	223	58	57.1

The median number of matches and inliers per frame was at most 239 and 73, respectively (Table 3). Note that the lowest number of matches and inliers correspond to the lowest ICC. Our matching approach was more accurate and slightly faster than the original brute force matching, with an average of 43.79

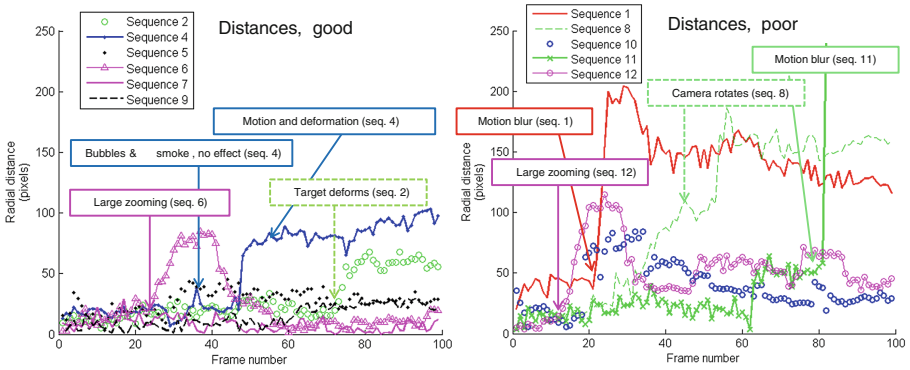


Fig. 2. Left: sequences with best ICC and smallest RMSE. Graphs show distance in pixels between the center of the two (automatic and annotated) targets. Note the text boxes that explain the larger shifts. Right: sequences with the poorest outcome. All tracking errors are caused by large motions.

(±4.37) ms against 48.55 (±4.33) ms per frame. We have uploaded the image sequences in the additional conference materials to illustrate the disturbing factors more clearly.

4 Discussion

In this study, the accuracy and robustness of a real-time tracking algorithm for automated target centralization in a robotized flexible endoscopy system was evaluated. This algorithm can be used for a variety of interventions. Here, we evaluated the algorithm using six image sequences of three different interventions. The achieved median accuracy was 91.6 %, which is an excellent result.

Robustness of the algorithm was shown by the continuous ability to track the target throughout the sequences, independent from procedure or tissue type, although several disturbing factors were present (Table 1). Inlying vectors mostly remained present and tracking was kept accurate, even with occlusions, color and illumination changes, surgical instruments, smoke and fluids present. Large and fast movements still form a problem; this caused most errors. If such an error occurred, the tracking was disturbed. For longer tracking periods of the same region this means re-initialization of the algorithm in its current form is sometimes necessary. However, we expect robustness to be improved with system implementation (see below).

Computations took a mean total time of 43.79 ms per frame. The algorithm could theoretically track the target every $\frac{1sec}{25frames} = 40$ ms. However, when using a newer computer with a better CPU (Intel Core i5-3570K) and a better GPU (AMD Sapphire Tri-X R9 290), computational time was below 40 ms and real-time system performance was ensured.

A limitation of this research is that zooming motion is assumed to be absent, although small zooming motions were present in the used sequences. Therefore, our current focus is on the implementation of the algorithm in the robotic system, complete with zooming functionality.

Current research further includes optimizing the robotic control based on the feedback that is generated from this algorithm. When combining the algorithm with the robotic control, large tracking errors (such as these over 100 pixels) will be diminished by the integral action that is present in the robotic controller. This action effectively smoothes the feedback signal because of the limited displacement possibility of the motors within a certain time frame. If this smoothing will not be enough to obtain the desired robustness, smart filtering with which previous information (key frames) is employed will be added to the system.

Finally, we will focus on establishing clinical relevance and patient safety. The algorithms are integrated in the Teleflex system, which is currently being evaluated in a phase II clinical trial, and we expect good results from this evaluation.

5 Conclusion

A target lock was designed for complex flexible endoscopic interventions. The algorithm performed accurately, robustly and worked in real-time. Intelligent navigation in robotized systems could assist the endoscopist during complex and time-consuming procedures. Clinical added value for the patient still needs to be objectively evaluated, but preliminary evaluation results seem promising.

References

1. Valdastri, P., Simi, M., Webster III, R.J.: Advanced technologies for gastrointestinal endoscopy. Ann. Rev. Biomed. Eng. **14**, 397–429 (2012). http://www.ncbi.nlm.nih.gov/pubmed/22655598
2. Van der Stap, N., Reilink, R., Misra, S., Broeders, I., van der Heijden, F.: The use of the focus of expansion for automated steering of flexible endoscopes. In: IEEE BioRob, pp. 13–18. IEEE, Rome (2012). http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6290804
3. van der Stap, N., Slump, C.H., Broeders, I.A.M.J., van der Heijden, F.: Image-based navigation for a robotized flexible endoscope. In: Luo, X., Reich, T., Mirota, D., Soper, T. (eds.) CARE 2014. LNCS, vol. 8899, pp. 77–87. Springer, Heidelberg (2014)
4. Ott, L., Nageotte, F., Zanne, P., Mathelin, M.D., Member, S.: Robotic assistance to flexible endoscopy by physiological-motion tracking. IEEE Trans. Robot. **27**(2), 346–359 (2011)
5. Atasoy, S., Mateus, D., Meining, A., Yang, G.-Z., Navab, N.: Endoscopic video manifolds for targeted optical biopsy. IEEE Trans. Med. Imaging **31**(3), 637–653 (2012). http://www.ncbi.nlm.nih.gov/pubmed/22057050
6. Allain, B., Hu, M., Lovat, L., Cook, R., Vercauteren, T., Ourselin, S., Hawkes, D.: Re-localisation of a biopsy site in endoscopic images and characterisation of its uncertainty. Med. Image Anal. **16**(2), 482–496 (2012). http://www.ncbi.nlm.nih.gov/pubmed/22197442

7. Ye, M., Giannarou, S., Patel, N., Teare, J., Yang, G.-Z.: Pathological site retargeting under tissue deformation using geometrical association and tracking. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 67–74. Springer, Heidelberg (2013)
8. Liu, J., Wang, B., Hu, W., Zong, Y., Si, J., Duan, H.: A non-invasive navigation system for retargeting gastroscopic lesions. Biomed. Mater. Eng. **24**(6), 2673–2679 (2014). http://www.ncbi.nlm.nih.gov/pubmed/25226971
9. Chu, Y.-J., Liu, S.-P., Luo, R.C., Hu, R.-H., Yeh, C.-C., Peng, Y.W., Yen, P.-L.: Dynamic tracking of anatomical object for a steerable endoscope. In: International Conference on Advanced Intelligent Mechatronics (2012)
10. Ruiter, J., Rozeboom, E., Voort, M.V.D., Bonnema, M., Broeders, I.: Design and evaluation of robotic steering of a flexible endoscope. In: IEEE BioRob, pp. 761–767. IEEE, Roma (2012)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 1–28 (2004)
12. Bay, H., Ess, A., Tuytelaars, T., Vangool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008). http://linkinghub.elsevier.com/retrieve/pii/S1077314207001555
13. Speidel, S., Krappe, S.: Robust feature tracking for endoscopic pose estimation and structure recovery. In: Proceedings of the SPIE. SPIE digital library, vol. 8671, pp. 1–7 (2013). http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1663315
14. Jellinek, E.M.: On the use of the intra-class correlation coefficient in the testing of the difference of certain variance ratios. J. Educ. Psychol. **31**(1), 60–63 (1940). http://dx.doi.org/10.1037/h0062703

2D/3D Real-Time Tracking of Surgical Instruments Based on Endoscopic Image Processing

Anthony Agustinos[1(✉)] and Sandrine Voros[2]

[1] UJF-Grenoble 1, CNRS, TIMC-IMAG UMR 5525, Grenoble 38401, France
Anthony.Agustinos@imag.fr
[2] UJF-Grenoble 1, INSERM, TIMC-IMAG UMR 5525, Grenoble 38401, France
Sandrine.Voros@imag.fr

Abstract. This paper describes a simple and robust algorithm which permits to track surgical instruments without artificial markers in endoscopic images. Based on image processing, this algorithm can estimate the 2D/3D pose of all the instruments visible in the image, in real-time (30 Hz). The originality of the approach is based on the use of a Frangi filter for detecting edges and the tip of instruments. The accuracy of the instruments' location in the image is evaluated using an extensive dataset (1500 images, 3 laparoscopic surgeries). Pose estimation of instruments in space is quantitatively evaluated on a test bench through comparison with the ground truth positioning provided by a calibrated robotic instrument holder. This method opens perspectives in the real-time control of surgical robots and the intra-operative recognition of surgical gestures.

Keywords: Laparoscopy · Image processing · Surgical instruments · Real-time tracking

1 Introduction

Laparoscopic surgery is a minimally invasive procedure. This technique reproduces the principles of conventional surgery with minimal physical trauma. Compared to open surgery, this approach is more beneficial to the patient but significantly increases the complexity of the surgical gestures. The constraints for surgeons are mostly ergonomic with the manipulation of surgical instruments (reduction of instrument mobility due to fixed insertion points on the abdominal cavity, loss of tactile sense) and the visualization of the surgical scene (limited field of view, indirect view of the surgical scene, endoscope manipulation). The realization of a laparoscopy requires a large adaptability from surgeons and requires a long learning curve.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-29965-5_9) contains supplementary material, which is available to authorized users.

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 90–100, 2016.
DOI: 10.1007/978-3-319-29965-5_9

Automatic localization of instruments can be helpful to respond to several limitations of laparoscopy and to assist surgeons during an intervention. For instance, [1] propose to localize instruments in space in a surgical trainer, based on a projective model and gradient image processing. In [2], a similar approach is proposed (also in a surgical trainer), with the addition of an extended Kalman filter to extract the edges of instruments.

In [3], the authors use the instrument insertion point as a constraint and a probabilistic algorithm to find instruments with the aim of controlling a robotic endoscope holder to assist surgeons during surgery.

All these methods use a gradient approach to extract instrument edges in the image. However, such approaches are sensitive to noise, illumination and shadows that can lead to insufficient segmentation for robust localization of instruments in the image [4]. To overcome this problem, we propose to use a 2D Frangi filter [5] to obtain a robust instruments edge detection. We present an algorithm to localize and track surgical instruments in endoscopic images in real-time. Our algorithm also permits to estimate the 3D position and orientation of the instruments using 2D information in the images, knowing the camera and instrument models.

2 Instrument Localization and Tracking Framework

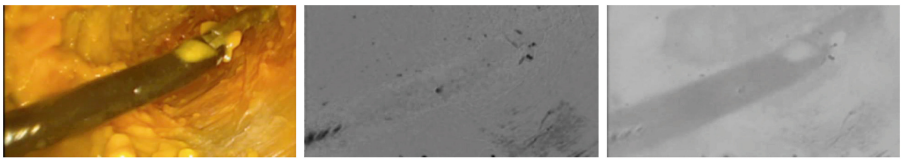
The principle of our instrument detection algorithm consists in:

- roughly identifying all regions corresponding to the location of an instrument in each laparoscopic image Sect. 2.1,
- refining the instruments detection within the identified regions Sect. 2.2,
- estimating the 3D pose of the instrument Sect. 2.3.

After an initial detection, the segmentation is constrained by the localization in the previous images to track the instrument.

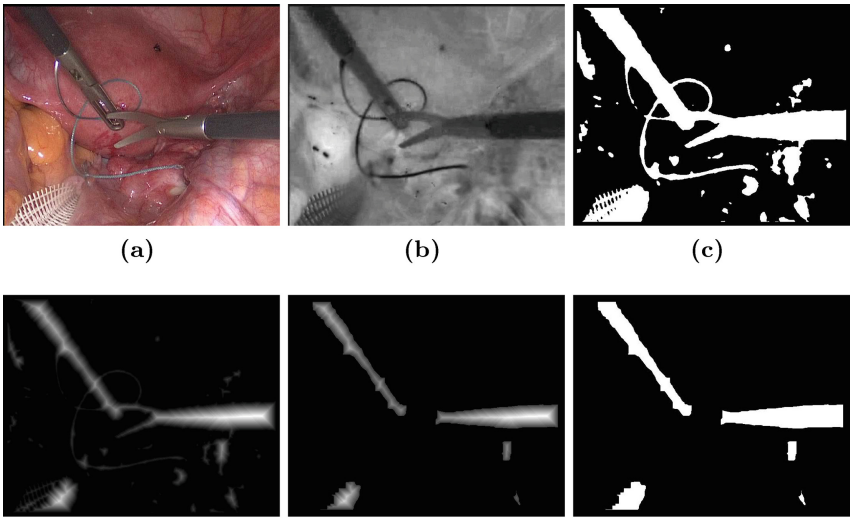
2.1 Rough Extraction of Instruments Regions

First, the laparoscopic color image (Fig. 2a) is converted from the RGB color space to the CIELab color space. The L channel, corresponding to the luminance is removed to free ourselves from variations of light inherent to laparoscopic surgery. We thus obtain a grayscale image composed of the a and b channels (Fig. 2b) corresponding to the chromaticity $C_{ab} = \sqrt{a^2 + b^2}$. Using this color space is more robust for challenging images than color spaces commonly used such as HSV [7] or RGB, see Fig. 1. We then binarize this grayscale image using an automatic Otsu thresholding approach [8]. Since the laparoscopic instruments have a color very distinct from the background (laparoscopic tools are usually black, metallic, or blue/green), instrument pixels will appear as white pixels whereas background pixels will appear as black (Fig. 2c). Of course, this pre-processing step is noisy, with background pixels appearing as white and tool pixels appearing as black (Fig. 2c). We disconnect the regions by skeletonizing



(a) (b) (c)

Fig. 1. Typical images obtained by color to grayscale conversion. (a) Original image (b) Saturation modified channels in HSV space [7] (c) Chromaticity C_{ab} of CIELab space (Color figure online)

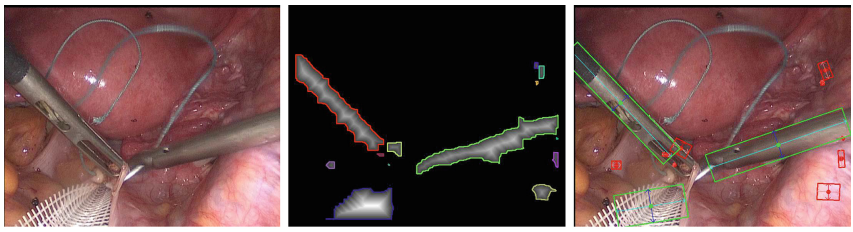


(a) (b) (c)

(d) (e) (f)

Fig. 2. Segmentation of a surgical instrument in 2D images. (a) Original image (b) Chromaticity C_{ab} of CIELab space (c) Segmentation using Otsu's thresholding (d) Conversion of the binary image using the distance transform (e) Disconnection of regions in the binary image using distance transform (f) Binarization of the distance transform image (Color figure online)

the image using a simple distance transform [9] and refine the separation by performing a simple erosion step on a cross-shaped kernel (Fig. 2d). Finally, we use a contour detection algorithm [10] to extract the extreme outer contour of each region as an oriented bounding box (see Fig. 3b). Based on the observation that laparoscopic instruments have a long and thin cylindrical shape, we eliminate bounding boxes with a width/length ratio inferior to 2 (red boxes in Fig. 3c).



(a) (b) (c)

Fig. 3. Edge detection of a surgical instrument in 2D images. (a) Original image (b) Edge detection (c) Potential instrument bounding boxes obtained from image b (green) and incompatible bounding boxes (red) (Color figure online)

2.2 Fine Extraction of Instrument Edges

Now that we have potential bounding boxes for the instruments, we search for instrument edges within each bounding box. To do so, we use a Frangi filter [5], which is the major contribution of this paper. We compared the Frangi filter to the classical Canny filter [6] to search instrument edges (see Fig. 4). The Canny filter is the most classical gradient approach based on the Sobel filter. This filter uses a hysteresis thresholding that requires to find two optimal thresholds for accurate extraction of the edges of an instrument. However, as shown in Fig. 4, the conditions of the surgical scene evolves during an intervention, thresholds initially determined may no longer be optimal and cause of false detections. The advantage of the approach based on the Frangi filter is that it can be applied to different surgery conditions without adjusting the filter parameters. This filter is classically used in vessel detection in medical images. It is based on the computation of the eigenvalues of the image's Hessian matrix λ_1, λ_2 such that $|\lambda_1| \leqslant |\lambda_2|$. The Hessian matrix is obtained by convolving the image with derivatives of a Gaussian kernel with standard deviation σ.

The Frangi filter function can be defined as:

$$\begin{cases} 0 & \text{if } \lambda_2 > 0, \\ V_0 = \exp(-\frac{R_B^2}{2\beta^2})(1 - \exp(-\frac{s^2}{2c^2})) \end{cases} \tag{1}$$

where, $R_B = \frac{\lambda_1}{\lambda_2}$ is the blobness measure, $s = \sqrt{\lambda_1^2 + \lambda_2^2}$ is the structureness measure and c, β are parameters to adjust the filter sensitivity. After applying the Frangi filter, each pixel value V_0 of the image indicates the pixel's probability of belonging to a tubular structure. Here, we do not use the Frangi filter to extract the whole cylindrical shape of the instrument. Indeed, the instrument's diameter in the image varies depending on its relative orientation with respect to the endoscope (i.e. we cannot fix the standard deviation σ). Instead, we apply the filter with a very low σ, in order to highlight the instrument edges (Fig. 5b). Finally, we identify the two borders of an instrument: the bounding box is extended and separated into two areas to search the top and bottom borders of the instrument separately using Hough transform [11] with a very low

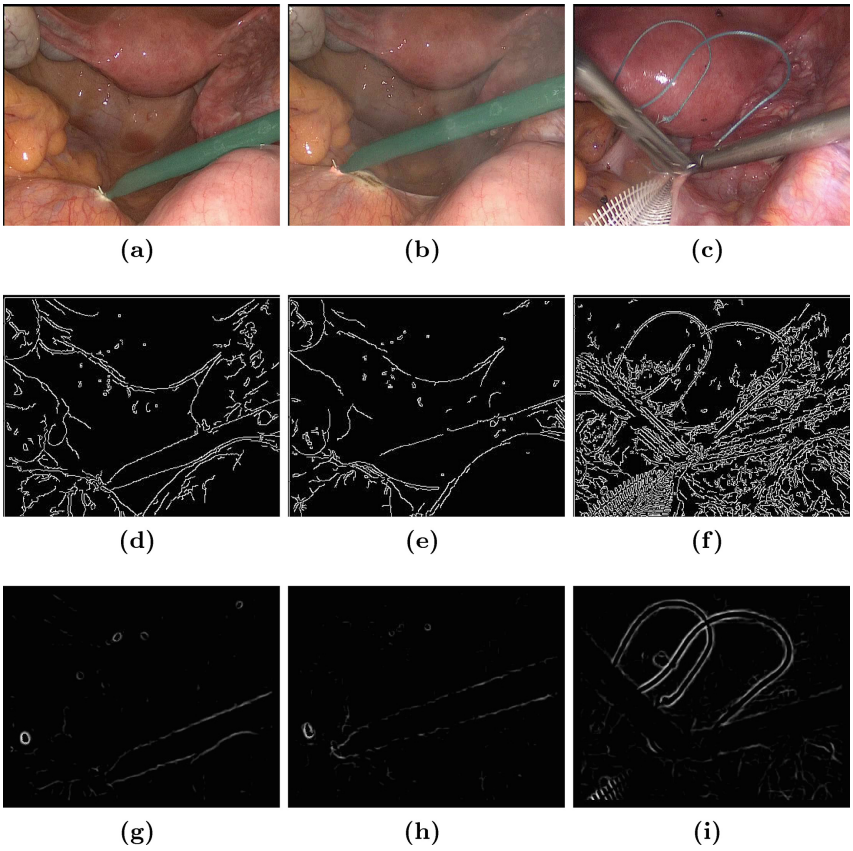


Fig. 4. Extraction of instrument edges. (a) and (b) Two images extracted from the same surgery at different time intervals (c) Image extracted from another surgery (d), (e) and (f) Edge detection in the images (a), (b) and (c) by the Canny filter with the thresholds $T_L = 30$ and $T_H = 90$ (h), (i) and (j) Edge detection by the Frangi filter in the images (a), (b) and (c) with the parameters $\sigma = 2$, $\beta = 0.5$ and $c = 0.5\max(S)$

threshold, as illustrated in Fig. 5b. At this step, we can eliminate lines that are incompatible with a surgical instrument based on the relative orientation and position of the detected lines (as illustrated by Fig. 5c).

2.3 Estimation of 3D Pose of the Instruments

The two borders of an instrument define two tangent planes $\sum_{\mathbf{i}}$ of normal $\mathbf{n_i}$ passing through the optical center of the camera $\mathbf{C}$ in space (see Fig. 5g). The camera calibration can be obtained with a classical chessboard calibration procedure such as [12]. The intersection of these two planes is a line $\mathbf{D} : (\mathbf{C}, \mathbf{e_1})$ parallel to the central axis of the instrument passing through the optical center of the camera with a direction vector $\mathbf{e_1}$. This line defines the instrument's central axis

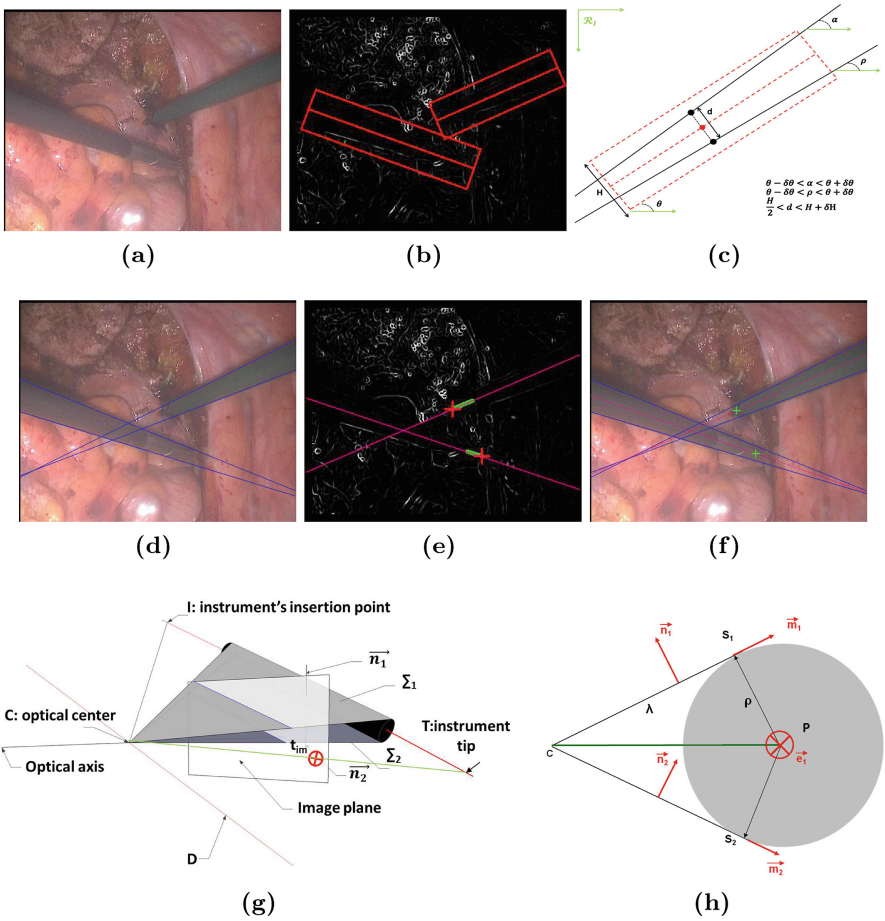


Fig. 5. Estimation of instruments poses in the image and in space. (a) Orignal image (b) Expansion and separation of a compatible bounding box in image filtered by the Frangi filter (c) Instruments' borders refinement process (d) Detection of the instruments borders (e) Instruments tips detection in the Frangi image (f) Instruments' pose in the image (g) Geometric representation of an instrument in space (h) Illustration of the compute instrument's position in space

direction in space. In order to fully describe the tool's orientation in space, we need to find a point $\mathbf{P}$ on the instrument's axis. To do so, we follow the approach proposed in [3]: the instrument is modeled as a finite cylinder of radius ρ (see Fig. 5g). Such a point $\mathbf{P}$ can be easily computed on the plane perpendicular to the instrument's axis (Fig. 5h). Indeed, $\mathbf{P}$ must respect the condition:

$$\lambda \mathbf{m_1} - \rho \mathbf{n_1} = \lambda \mathbf{m_2} + \rho \mathbf{n_2} \tag{2}$$

where $\mathbf{m_i} = \mathbf{e_1} \otimes \mathbf{n_i}$, λ is the distance from the optical center to tangent points $\mathbf{S_i}$ and $\mathbf{n_i}$ the normal to the plane $\mathbf{i}$. Using Eq. 2, we can compute λ and obtain:

$$\overrightarrow{\mathbf{CP}} = \lambda \mathbf{m_1} - \rho \mathbf{n_1} = \rho \frac{\|\mathbf{n_1} + \mathbf{n_2}\|^2}{(\mathbf{m_1} - \mathbf{m_2}).(\mathbf{n_1} + \mathbf{n_2})} \mathbf{m_1} - \rho \mathbf{n_1} \tag{3}$$

Then, we search the position of the instrument's tip $\mathbf{t_{im}}$, in the Frangi image along the projection of the instrument's axis $(\mathbf{P}, \mathbf{e_1})$ in the image (see Fig. 5e). The pixel along the line with maximum grey level in the Frangi image is considered as the tip. Finally, we find the 3D position of the instrument's tip $\mathbf{T}$ as the intersection of $(\mathbf{P}, \mathbf{e_1})$ and the projection line of the tool's tip $(\mathbf{C}, \mathbf{t_{im}})$.

2.4 Tracking of Surgical Instruments

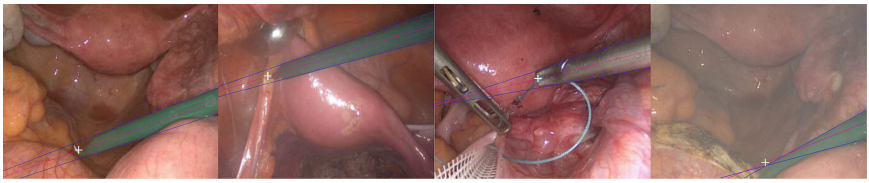
For our instrument tracking algorithm, we assume that between two successive images, an instrument does not undergo large displacements. In the initial step (first image), we find the instrument as described in Sects. 2.1 and 2.2. In the following images, we find the candidate bounding boxes, but we refine the instrument search only inside the bounding box best compatible with the position/orientation of the instrument in the previous image. If the instrument is not found in several images, we re-initialize the algorithm. In the case of several instruments, it is possible to track all the visible instruments or a particular one. Since only one instrument can be inserted at once through an insertion point $\mathbf{I}$ on the abdominal wall, we can identify an instrument thanks to its insertion point, which can be easily computed using a pivot algorithm on $(\mathbf{P}, \mathbf{e_1})$.

3 Experiments and Results

Our algorithm is implemented in C^{++} using OpenCV and OpenMP libraries. For the computations, we used an Intel Xeon PC 2.67 GHz, 3.48 GB RAM. The 2D evaluation was performed on real laparoscopic images (720×556). The 3D evaluation was performed on a laparoscopy test bench using an OLYMPUS OTV600 CCD and an IC Imaging Source grabber (720×480, 25 fps). To achieve a fast processing time the image resolution is divided by 2 for the region extraction and by 4 for the Frangi filter. We evaluated 2D tracking of our algorithm on three in-vivo video sequences of laparoscopic rectopexies obtained through the Digestive Departement of Grenoble Hospital with challenging situations (see Fig. 6).

In these images, the tip position and orientation of the instrument were compared to manual annotation. The results obtained for each sequence are presented in Table 1 with a mean error of 16.10 pixels (std. dev. of 28.98) for the tip position, a mean error of 0.90° (std. dev. 0.88°) for the 2D orientation and a frequency of 30 Hz. Videos of this evaluation are included in supplementary material.

To evaluate the accuracy of the 3D pose estimation, we performed experiments on a testbench (see Fig. 7) consisting of a surgery trainer box on which a



(a) (b) (c) (d)

Fig. 6. Results of the tracking on the laparoscopic image sequences. (a) Sequence 1: Monopolar hook instrument (b) Sequence 2: Monopolar hook instrument (c) Sequence 3: Needle holder instrument (d) Example of a bad tip detection

Table 1. Laparoscopic images statistics

Instruments	# Images	Fps	Position error (pixels) Mean (Std. Dev.)	Orientation error (°) Mean (Std. Dev.)
Sequence 1	500	31.47	10.02 (11.2)	0.71 (0.55)
Sequence 2	550	29.76	15.46 (9.72)	0.75 (0.63)
Sequence 3	525	28.87	22.83 (47.14)	1.25 (1.19)
All	1575	30.03	16.10 (28.98)	0.90 (0.88)

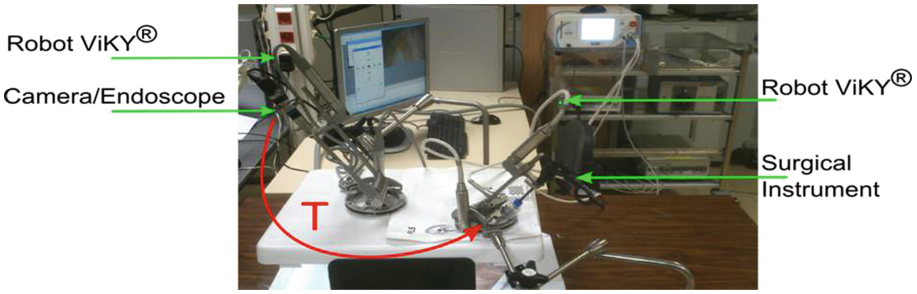
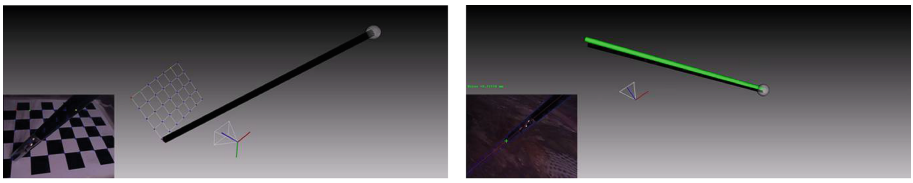


Fig. 7. Experimental test bench to evaluate the 3D pose estimation accuracy with a printout of a surgical scene as background.



(a) (b)

Fig. 8. Estimation of the instrument's pose in space. (a) Calibration step to find the rigid transformation **T** (b) Evaluation of the 3D pose estimation accuracy with in black, the reference pose obtained with the robot, in green, the pose computed with our method (Color figure online)

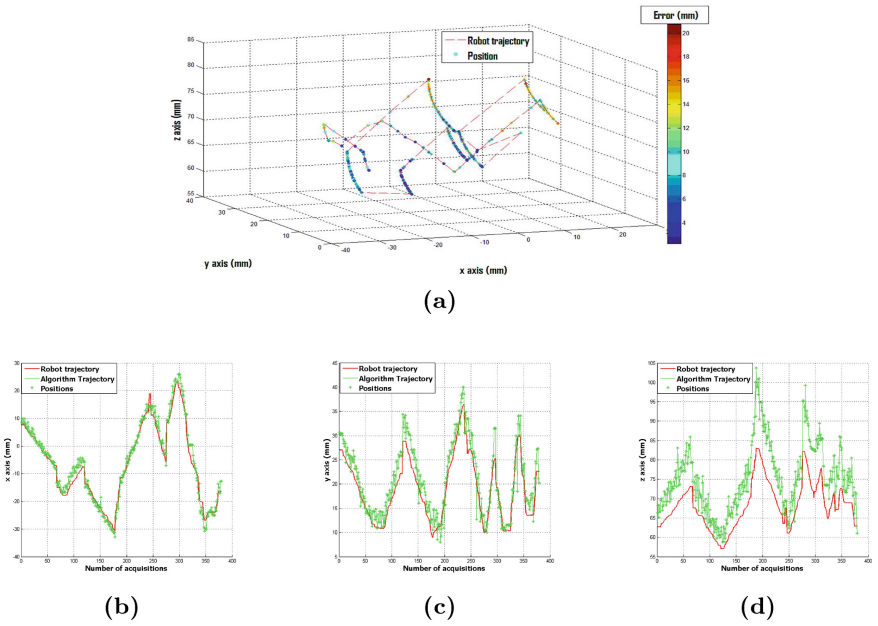


Fig. 9. Robot trajectory (red) and our tracking method trajectory (green) for 380 instrument positions. (a) 3D trajectories (b), (c) and (d) X, Y, Z trajectories with respect to the camera frame (of normal Z) (Color figure online).

commercial robotic instrument holder is directly positioned, and a printout of a surgical scene as background. We compared the 3D tip position of the instrument found by our algorithm to the 3D tip position given by the robot expressed in the camera referential. This required calibrating the system to find the rigid transformation $\mathbf{T}$ between the robot and camera frame such that: $\mathbf{p}_{\mathbf{cam}}^{\mathbf{frangi}} = \mathbf{T}\mathbf{p}_{\mathbf{cam}}^{\mathbf{robot}}$. $\mathbf{T}$ is obtained by pointing 12 points of a chessboard, for 6 chessboard positions, with the instrument carried by the robot (see Fig. 8). These 12 points can be expressed in the camera frame thanks to a standard extrinsic camera calibration procedure [12] and are also measured in the robot frame. We resolve a classical least squares system to find the rigid transformation between the two sets of 3D points coupled with a RANSAC to eliminate outliers. We obtain a camera calibration Root Mean Square (RMS) error of 0.25 pixels and $\mathbf{T}$ with a RMS error of 1.2 mm. Figure 9 shows an example of the robot trajectory and of our tracking method for a series of instrument movements. The results for 380 measurements are presented in Table 2. In all results presented, we fixed the Frangi filter parameters as $\sigma = 2$, $\beta = 0.5$ and $c = 0.5\max(s)$, according to recommandations from the literature.

Table 2. Error of the 3D pose estimation with our method compared to the position obtained with a robotic instrument holder

Axis	Mean 3D position error (mm)	Std. Dev. (mm)
x	1.79	1.45
y	2.42	1.60
z	7.24	3.51

4 Conclusion

We presented a surgical instrument tracking algorithm based on image processing. It permits to estimate the 2D/3D instruments pose in real-time without artificial fiducials. An extensive 2D evaluation on real surgical videos shows that our 2D pose estimation is accurate and robust on wide range of realistic cases. In difficult situations as a suture gesture, we can lose accuracy in the instrument's tip position but the orientation is still correct. A machine learning approach as [13], applied in the neighbourhood of our estimated tip position could increase the accuracy of the tip detection. Our approach for 3D pose estimation was validated on a test-bench using a printout of a surgery background. Although this might lack realism we estimated that the robustness of the proposed method on realistic images was already shown extensively on the 2D case. This 3D evaluation provides us with the precision range we can expect when the 2D detection works well. The greatest errors are found in the depth estimation along the z axis. This error could be reduced by using a stereoscopic endoscope.

Our 2D localization approach is robust and accurate enough to control a robotic endoscope holder. Even if the Frangi filter might not be the most obvious approach for edge detection, we showed that it works better than classical approaches. Other more sophisticated edge detection approaches could easily be compared on our image database. The 3D pose estimation could be useful for surgical gesture recognition or for co-manipulation, if we are able to increase the depth precision. Another application could be the online calibration of no rigidly-linked robotic endoscope and instrument holders, which could lead to less bulky surgical systems. Our next step will be to evaluate the 3D pose estimation more extensively in conditions closer to the clinical reality (cadaver experiments).

References

1. Cano, A.M., Gayá, F., Lamata, P., Sánchez-González, P., Gómez, E.J.: Laparoscopic tool tracking method for augmented reality surgical applications. In: Bello, F., Edwards, E. (eds.) ISBMS 2008. LNCS, vol. 5104, pp. 191–196. Springer, Heidelberg (2008)
2. Zhou, J., Payandeh, S.: Visual tracking of laparoscopic instruments. J. Autom. Control Eng. **2**(3), 234–241 (2014)

3. Wolf, R., Duchateau, J., Cinquin, P., Voros, S.: 3D tracking of laparoscopic instruments using statistical and geometric modeling. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 203–210. Springer, Heidelberg (2011)
4. Allan, M., Ourselin, S., Thompson, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: Toward detection and localization of instruments in minimally invasive surgery. IEEE Trans. Biomed. Eng. **60**(4), 1050–1058 (2013)
5. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
6. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 679–698 (1986)
7. Doignon, C., Graebling, P., de Mathelin, M.: Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. Real-Time Imaging **11**(5), 429–442 (2005)
8. Otsu, N.: A threshold selection method from gray-level histograms. Automatica **11**(285–296), 23–27 (1975)
9. Felzenszwalb, P., Huttenlocher, D.: Distance transforms of sampled functions. Cornell University. (2004)
10. Suzuki, S.: Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. **30**(1), 32–46 (1985)
11. Hough, V., Paul, C.: U.S. Patent No. 3,069,654. U.S. Patent and Trademark Office, Washington (1962)
12. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1330–1334 (2000)
13. Sznitman, R., Becker, C., Fua, P.: Fast part-based classification for instrument detection in minimally invasive surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 692–699. Springer, Heidelberg (2014)

Tracking Accuracy Evaluation of Electromagnetic Sensor-Based Colonoscope Tracking Method

Masahiro Oda[1(✉)], Hiroaki Kondo[1], Takayuki Kitasaka[2], Kazuhiro Furukawa[3], Ryoji Miyahara[3], Yoshiki Hirooka[4], Hidemi Goto[3], Nassir Navab[5], and Kensaku Mori[1,6]

[1] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan
moda@mori.m.is.nagoya-u.ac.jp
[2] School of Information Science, Aichi Institute of Technology, Toyota, Japan
[3] Department of Gastroenterology and Hepatology, Nagoya University Graduate School of Medicine, Nagoya, Japan
[4] Department of Endoscopy, Nagoya University Hospital, Nagoya, Japan
[5] Technische Universität, Munich, Germany
[6] Strategy Office, Information and Communications, Nagoya University, Nagoya, Japan
http://www.mori.m.is.nagoya-u.ac.jp/∼moda/index-e.html

Abstract. This paper reports a detailed evaluation results of a colonoscope tracking method. A colonoscope tracking method utilizing electromagnetic sensors and a CT volume has been proposed. Tracking accuracy of this method was evaluated by using a colon phantom. In the previously proposed paper, tracking errors were measured only at six points on the colon phantom for the accuracy evaluation. The point number is not enough to evaluate relationships between the tracking errors and positions in the colon. In this paper, we evaluated the colonoscope tracking method based on more detailed measurement results of the tracking errors. We measured tracking errors at 52 points on the colon phantom and visualized magnitudes of the tracking errors. From our experiments, tracking errors in the ascending and descending colons were enough small to perform colonoscope navigations. However, tracking errors in the transverse and descending colons were large due to colon deformations.

Keywords: Colon · Colonoscope tracking · CT image · Evaluation

1 Introduction

Colonoscopy is conventionally performed as a colon diagnosis or inspection method. However, colonoscopy may cause discomfort for patients while diagnosis. Also, colonoscopy has a risk of complication including perforations of the

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 101–108, 2016.
DOI: 10.1007/978-3-319-29965-5_10

colon. Success of colon diagnosis under a colonoscope is heavily depends on physician's skill.

CT colonography (CTC) is a colon diagnosis method that reduces discomfort and a risk of complication on patients. Diagnosis is performed by using CT images of a patient in CTC. Some computer aided diagnosis (CAD) systems for CTC are commercially available. These systems commonly display 2D or 3D or unfolded views of the colon for observation. Physicians diagnose the colon to find polyps or cancers.

Polyps or cancers of the early stages found by CTC CAD systems can be removed in colonoscopic polypectomy. Colonoscopic polypectomy is a surgery to remove polyps or cancers. In colonoscopic examinations including colonoscopic polypectomy, a physician controls a colonoscope based on his/her experience. Experienced physicians remove polyps or cancers minimizing patient discomfort. However, colonoscopic examinations performed by inexperienced physicians may painful for patients. Utilization of a navigation system for colonoscope is one solution for such problem. Colonoscope navigation systems indicate positions of the colonoscope tip and targets such as polyp positions while performing colonoscopic examinations. Colonoscope navigation systems can be used to reduce overlooking of polyps and to assist inexperienced physicians. Conventionally, information obtained from CT volumes of patients is utilized only for the diagnosis stage including diagnosis using CTC CAD systems. Information obtained from CT volumes contains the polyp positions and the colon shapes. Such information is useful for the treatment stage including colonoscopic examinations. The colon shapes obtained from CT volumes can be used as maps of the colon in colonoscope navigation systems. Also, the polyp positions can be used as target point in navigations. A colonoscope will be navigated to polyp positions while performing colonoscope examinations by utilizing information obtained from CT volumes of patients.

To achieve colonoscope navigation systems, tracking method for colonoscope is required. Tracking methods of endoscopes including colonoscope have been proposed by many research groups, which estimate an endoscope tip position in the organs. For bronchoscope tracking, image-based [1–4] and sensor-based [5,6] tracking methods were reported. Colonoscope tracking is difficult compared to the tracking for other hollow organs because the colon greatly deforms during an insertion of the colonoscope. Liu et al. [7] tried to estimate colonoscope tip movements from the optical flow of colonoscope videos. This method can track a colonoscope tip without using additional equipments for the tracking. However, tracking using colonoscope videos is easily fails when unclear video frames appear. Unclear video frames frequently appear in colonoscope video because fluid, feces, and bubbles exist in the colon. The colonoscope tip touches the wall of the colon many times while colonoscope examinations. It causes black video frames that make interruptions of tracking. A colonoscope shape tracking system, the Olympus ScopeGuide (UPD-3), is commercially available. The system detects colonoscope shape in the colon using electromagnetic (EM) position sensors. Clinical reports about utilization of the system in colonoscopy have been reported [8]. The system just displays

the shape of the colonoscope without combining CT volumes or CTC information. Oda et al. [9,10] and Kondo et al. [11] proposed colonoscope tracking method using EM position sensors with combination of CT volumes. They attach EM sensors to colonoscope to obtain the colonoscope shape. They obtain the colon shape of a patient from a CT volume. Correspondences between the colonoscope and colon shapes by applying two steps correspondence finding processes. Based on the correspondences, they find a point in the CT volume which corresponds to the colonoscope tip position. These methods can track the colonoscope tip position even if the viewing fields of the colonoscope are not clear. In their tracking error evaluations, they measured tracking errors only at six points in a colon phantom. Behaviors of the colon deformation are differ according to position. Therefore, tracking errors should evaluated at many positions in the colon.

In this paper, we perform detailed evaluations of the tracking errors of the colonoscope tracking method using EM sensors [11]. We measured tracking errors at 52 points in a colon phantom by using the tracking method. Based on the measurement results, we discuss relationships between the colon deformations and the tracking errors.

In the Sect. 2, we briefly introduce the colonoscope tracking method proposed in the reference [11]. Experimental results including tracking error measurement results are shown in the Sect. 3. Discussion about the experimental results are described in the Sect. 4.

2 Method

2.1 Colon Centerline and Colonoscope Line Generation

A colon centerline is obtained from a CT volume. We use a region growing method to extract a colon region from the CT volume. A thinning and a line smoothing processes are applied to generate a colon centerline.

A colonoscope line that represents the colonoscope shape is obtained by using EM position sensors. We insert an Aurora 5/6 DOF Shape Tool Type 1 (NDI) to the colonoscope working channel. The shape tool gives positions and directions of the colonoscope at seven points. The colonoscope line is calculated by applying the Hermite spline interpolation to positions and directions measured by the shape tool.

2.2 Coordinate System Registration

We generate a modified colon centerline that simulates the shape of the colon while an insertion of the colonoscope. This process is required because the colon largely deformed while colonoscope insertions. To generate the modified colon centerline, we detect sections on the colon centerline that corresponds the transverse and sigmoid colons. The transverse and sigmoid colon sections on the colon centerline are replaced with straight line sections. The transverse and sigmoid

colon sections are identified based on the positions and the shape of the colon centerline.

We register the CT and sensor coordinate systems by using the ICP algorithm [12]. The ICP algorithm finds a rigid transformation matrix that minimizes the Euclidean distance between the modified colon centerline and colonoscope line. The colonoscope line is transformed to the CT coordinate system by using the rigid transformation matrix.

2.3 Colonoscope Tip Position Finding

We find correspondences between each point on the colon centerline and colonoscope line. This process consists of two steps including a landmark-based coarse correspondence finding and a length-based fine correspondence finding. The landmark-based coarse correspondence finding process finds corresponding point pairs on the colon centerline and colonoscope line at five anatomical landmarks. The five anatomical landmarks are detected based on their positions and the shape of the colon. After performing the coarse correspondence finding, the length-based fine correspondence finding is applied. The length-based fine correspondence finding process finds corresponding point pairs on the colon centerline and colonoscope line at all points on them. This process finds correspondences by using lengths along the lines. Finally, a point on the colon centerline that corresponds to the tip of the colonoscope line is defined as the colonoscope tip position in the CT coordinate system.

3 Experiments

The proposed method was evaluated in experiments using a colon phantom. A colon phantom (KOKEN colonoscopy training model type I-B) (Fig. 1(a)) and its CT volume are utilized in our experiments. We used a colonoscope (Olympus CF-Q260AI). An Aurora 5/6 DOF Shape Tool Type 1 (NDI) is inserted to the working channel of the colonoscope. The Aurora 5/6 DOF Shape Tool Type 1 has seven EM sensors.

In colonoscopic examinations, physicians observe the colon by using a colonoscope while pulling back the colonoscope after insertion up to the cecum. To simulate this situation, we inserted the colonoscope up to the cecum of the colon phantom before performing the colonoscope tracking. After the insertion, we started the colonoscope tracking and measured tracking errors while pulling back the colonoscope. Definition of the tracking error is described below.

We evaluated the performance of the proposed method by using a tracking error. We defined evaluation points (EPs) on the surface of the colon phantom. Points on the colon phantom surface which have characteristic shapes (such as parts of the haustral folds or taeniae coli) were selected as EPs. The EPs are visually identifiable from both of the colon phantom and its CT volume. Positions of the EPs and indices of them are shown in the Fig. 1(b). The position of each EP was projected to the closest point on the colon centerline. The tracking error

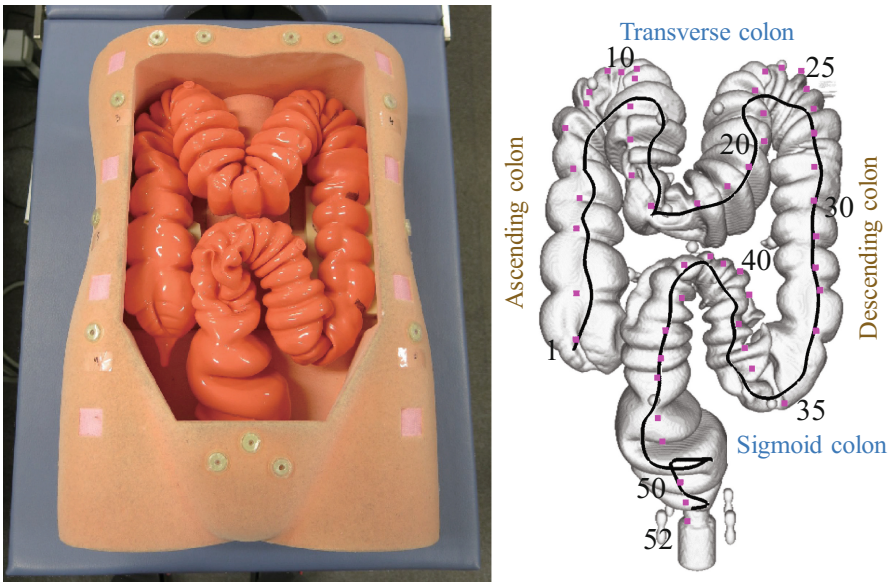


Fig. 1. (a) Colon phantom. (b) Positions of EPs placed on the surface of the colon phantom. Purple points are EPs and numbers are indices of EPs. Black line is colon centerline.

is the length along the colon centerline between an estimated position of the colonoscope tip (estimated by the colonoscope tracking method) and a position of a projected EP when the real colonoscope tip comes to the closest position to the marker.

In the reference [11], tracking errors were measured at six EPs. We performed a detailed evaluation of the colonoscope tracking method by using 52 EPs. We measured tracking errors of colonoscope insertions in three trials. Figure 2 shows the average tracking errors at each EPs. The segments of the colon (ascending, transverse, descending, and sigmoid colons) are also shown in this figure. Measurements were failed at the EPs of indices from 46 to 50 because we could not find these EPs due to large deformations of the colon phantom while pulling the colonoscope. We showed the average tracking errors by using colors on the colon phantom as Fig. 3. In this figure, blue and red colors indicate small and large average tracking errors. Small tracking errors were obtained in the ascending and descending colons. Large tracking errors were obtained in the transverse and sigmoid colons.

4 Discussion

From the experimental results, relations between regions in the colon and tracking errors were clearly shown. We measured tracking errors at 52 EPs. The

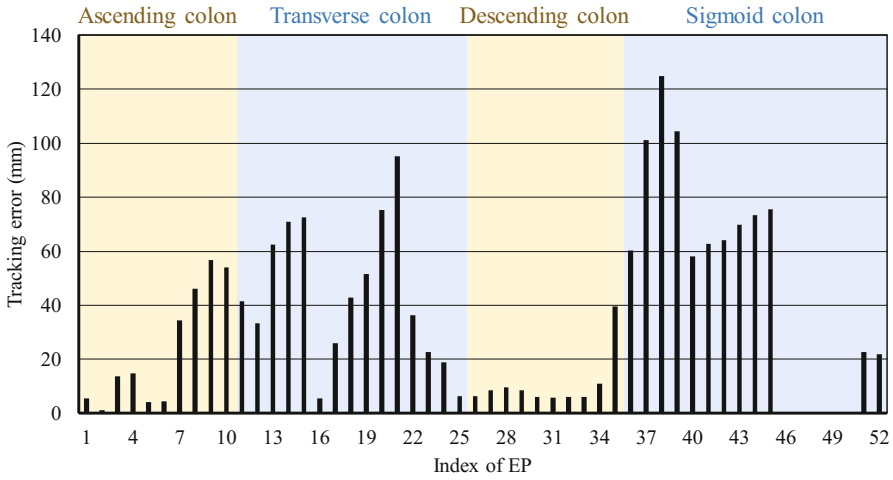


Fig. 2. Average tracking errors at EPs from three trials. Areas correspond to the ascending, transverse, descending, and sigmoid colons are indicated by yellow and blue colors. Average tracking errors of EPs having indices from 46 to 50 are not shown here because measurements were failed (Color figure online).

number of EPs was significantly larger than the reference [11]. Our result is useful to investigate causes of the tracking errors.

The tracking errors were quite small in the ascending and descending colons. Most of the tracking errors in the regions were smaller than 40 mm. Tracking errors in the regions were enough small to perform colonoscope navigations. A physician who specializes in gastroenterology said that tracking errors smaller than 50 mm are acceptable for navigations of the colonoscope tip to polyps. If a polyp comes to a position near the colonoscope tip (about 50 mm or closer), it is observable from the colonoscope camera. The tracking method is applicable for colon navigations to find polyps in the ascending and descending colons. The tracking errors were small in the ascending and descending colons because these regions not deform largely. The ascending and descending colons are fixed to the other tissues. It makes small tracking errors in these regions.

Unlike the ascending and descending colons, the transverse and sigmoid colons largely moves in the abdominal cavity. The transverse and sigmoid colons largely deform during colonoscope insertions. The shapes of the transverse and sigmoid colons during the colonoscope insertions are nearly straight. It caused the large tracking errors in the transverse and sigmoid colons. Estimation method of colon deformations during colonoscope insertions is required to reduce tracking errors.

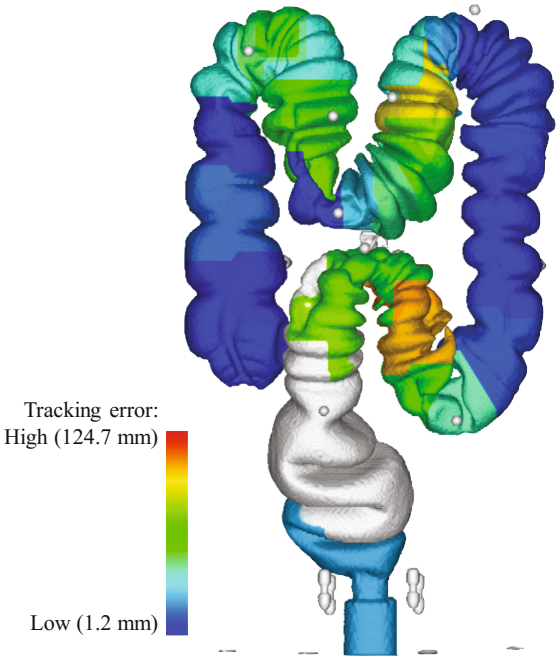


Fig. 3. Average tracking errors at EPs indicated by colors on the colon phantom. Blue and red colors indicate small and large average tracking errors. White color means measurement was failed at the region due to deformations of the colon phantom while pulling the colonoscope. This image is rendered from a CT volume of the colon phantom (Color figure online).

5 Conclusions

This paper reported a detailed evaluation results of the tracking errors of the colonoscope tracking method. The colonoscope tracking method estimates the colonoscope tip position in the colon by using EM sensors and a CT volume. We measured average tracking errors at 52 points in a colon phantom by using the tracking method. Three trials of measurements were performed. The average tracking errors in the ascending and descending colons were enough small to perform colonoscope navigations. However, the average tracking errors in the transverse and sigmoid colons were large due to deformations of the colon.

Acknowledgments. Parts of this research were supported by the MEXT, the JSPS KAKENHI Grant Numbers 24700494, 25242047, 26108006, 26560255, and the Kayamori Foundation of Informational Science Advancement.

References

1. Peters, T., Cleary, K.: Image-Guided Interventions: Technology and Applications. Springer, New York (2008)
2. Deligianni, F., Chung, A., Zhong, G.: Predictive camera tracking for bronchoscope simulation with CONDensation. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 910–916. Springer, Heidelberg (2005)
3. Rai, L., Helferty, J.P., Higgins, W.E.: Combined video tracking and image-video registration for continuous bronchoscopic guidance. Int. J. Comput. Assist. Radiol. Surg. **3**, 3–4 (2008)
4. Deguchi, D., Mori, K., Feuerstein, M., Kitasaka, T., Maurer, C.R., Suenaga, Y., Takabatake, H., Mori, M., Natori, H.: Selective image similarity measure for bronchoscope tracking based on image registration. Med. Image Anal. **3**(14), 621–633 (2009)
5. Gildea, T.R., Mazzone, P.J., Karnak, D., Meziane, M., Mehta, A.: Electromagnetic navigation diagnostic bronchoscopy: a prospective study. Am. J. Respir. Crit. Care Med. **174**(9), 982–989 (2006)
6. Schwarz, Y., Greif, J., Becker, H., Ernst, A., Metha, A.: Real-time electromagnetic navigation bronchoscopy to peripheral lung lesions using overlaid CT images: the first human study. Chest **129**(4), 988–994 (2006)
7. Liu, J., Subramanian, K.R., Yoo, T.S.: An optical flow approach to tracking colonoscopy video. Comput. Med. Imaging Graph. **37**(3), 207–223 (2013)
8. Fukuzawa, M., Uematsu, J., Kono, S., Suzuki, S., Sato, T., Yagi, N., Tsuji, Y., Yagi, K., Kusano, C., Gotoda, T., Kawai, T., Moriyasu, F.: Clinical impact of endoscopy position detecting unit (UPD-3) for a non-sedated colonoscopy. World J. Gastroenterol. **21**(16), 4903–4910 (2015)
9. Oda, M., Acar, B., Furukawa, K., Kitasaka, T., Suenaga, Y., Navab, N., Mori, K.: Colonoscope tracking method based on line registration using CT images and electromagnetic sensors. Int. J. Comput. Assist. Radiol. Surg. **8**(1), S349–S351 (2013)
10. Oda, M., Kondo, H., Kitasaka, T., Furukawa, K., Miyahara, R., Hirooka, Y., Goto, H., Navab, N., Mori, K.: Colonoscope navigation system using colonoscope tracking method based on line registration. In: Proceedings of SPIE 9036, 903626-1-7 (2014)
11. Kondo, H., Oda, M., Furukawa, K., Miyahara, R., Hirooka, Y., Goto, H., Kitasaka, T., Mori, K.: Development of marker-free estimation method of colonoscope tip position using electromagnetic sensors and CT volumes. Int. J. Comput. Assist. Radiol. Surg. **9**(1), S11–S12 (2014)
12. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**, 239–256 (1992)

Non Rigid Registration of 3D Images to Laparoscopic Video for Image Guided Surgery

Max Allan[1], Ankur Kapoor[1], Philip Mewes[2], and Peter Mountney[1(✉)]

[1] Siemens Corporate Research, Princeton, NJ, USA
peter.mountney@siemens.com
[2] Siemens Healthcare, Forchheim, Germany

Abstract. Image guidance and the visualization of sub surface structures during laparoscopic procedures have the potential to change the current capabilities of surgery. Increased target localization accuracy and the identification of critical structures can reduce resection margins, procedure time and tissue trauma while simplifying procedures and enabling new functional capabilities. Image guidance requires the registration of 3D images to the laparoscopic video. Tissue deformation and lack of cross modality landmarks make this challenging. Registration can be performed by aligning the 3D image to a surface reconstructed from stereo laparoscopic images. Current research is focused on creating more generic stereo reconstruction techniques and rigid registration methods. This paper proposes a novel stereo reconstruction approach which exploits prior knowledge of patient specific organ models and outlier robust non rigid registration. The approach is validated on phantom data and the practical application of the reconstruction is demonstrated on in vivo data.

1 Introduction

During procedures such as liver resection, the laparoscopic camera is used to visualize the surfaces of organs. The target anatomy (e.g. a tumor) can be hidden below the surface of the organ making surgery challenging. Image guidance, through the registration of 3D volumetric data (CT, MRI) and laparoscopic images, has the capability to display sub surface information directly on the laparoscopic video feed. Registration is challenging due to the lack of cross modality landmarks, the laparoscope's small field of view and tissue deformation caused by cardiac/respiratory motion, CO_2 insufflation and tissue tool interaction.

Registration can be performed manually [1], using robotic kinematics or with a calibrated tracking system [2]. These approaches do not require naturally occuring cross modality landmarks but do not account for tissue deformation. Inaccuracies of several mm can be introduced due to tracking and calibration error making their stand alone use for Image Guidance limited [2]. However, such methods can be used to initialize image based registrations approaches.

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 109–116, 2016.
DOI: 10.1007/978-3-319-29965-5_11

Image based surface registration approaches typically use two separate parts, (1) generating a 3D surface from intra-operative imaging (laparoscope) and (2) registering this surface to the 3D volumetric data. The intra-operative 3D surface can be generated using specialized hardware [3,4] or monocular laparoscopes using rigid [2] or non rigid [5] structure from motion or using a stereo laparoscope. The clinical use of stereo laparoscopes in steadily increasing due to robotics and the recent release of several commercially available stereo systems. Early work on surgical stereo surface reconstruction [6] used winner-takes-all matching approaches which can suffer from false matches. To reduce false matches and exploit local information, anchor points and 2D constraints have been introduced [7] that limit disparity estimates and propagate good matches. Assumptions about the prior shape of the tissue surface have been introduced to improve stereo reconstruction [8,9]; however, the stereo reconstruction cannot always be modelled as a generic spline and these methods do not make use of patient specific prior knowledge of the organ surface.

Registration of 3D images to intra-operative surface reconstructions has focused on rigid Iterative Closest Point (ICP) based algorithms [1,3]. Prior knowledge of sensor noise can be incorporated [3] to improve registration. Recently, a method [10] has been proposed for deformable surfaces that finds a sparse set of points which have close-to-rigid transformations. In [4] rigid registration is followed by a biologically plausible point to mesh minimization. This minimizes deformation but has huge computational cost.

This paper proposes a method for registering 3D images and laparoscopic video. Stereo reconstruction is proposed which exploits patient specific prior knowledge of the organ's surface. Combined rigid and non rigid registration based on coherent point drift [11] is used for registering the 3D image to the stereo reconstruction. Reconstruction is applied to in vivo data to demonstrate its clinical application.

2 Method

The proposed method for registering 3D images to laparoscopic video is outlined in Fig. 1. A new surgical workflow is proposed where 3D images are captured using cone beam CT after the patient is insufflated. The cone beam CT image is automatically segmented to extract an anatomical model. The stereo laparoscope is initially aligned to the cone beam CT coordinate system manually or using a calibrated tracking system. Stereo reconstruction is performed using prior knowledge of the patient anatomy. The stereo reconstruction is non rigidly registered to the 3D image to compensate for errors in the initial alignment and tissue deformation.

2.1 3D Surface Reconstruction

The goal of any 3D reconstruction is to rebuild a surface in 3D space. By defining this target surface as a mapping from pixels to depth values: $(u, v) \mapsto p(u, v)$

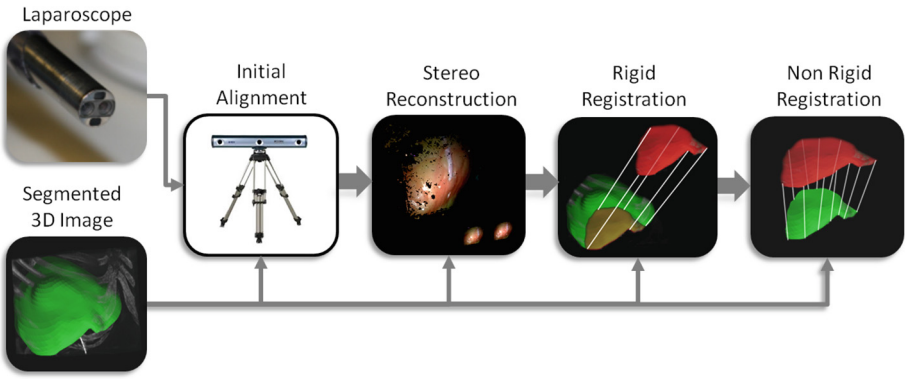


Fig. 1. Registration of 3D volumetric data to the laparoscopic video stream.

this problem can be posed within an energy minimization framework where the reconstruction problem is reduced to finding the minimum of a functional of the form

$$F = \int\int \left(c(p(u,v),u,v) + \kappa s\left(\frac{\partial p(u,v)}{du}, \frac{\partial p(u,v)}{dv} \right) \right) dudv \tag{1}$$

where first term $c(p(u,v),u,v)$ defines a cost for assigning a particular depth value $p(u,v)$ to a coordinate (u,v), and the second term $s(.)$ regularizes the solution to obtain a smooth surface reconstruction by minimizing the derivative of the depth function. κ is a weighting constant to affect the strength of the smoothness regularization.

This functional can be minimized by constructing it as a labeling problem on a Markov Random Field (MRF) for which graph-cut algorithms exist to find the optimal labeling. A 3D graph is embedded in a volume of interest that contains the target surface. Each voxel of space is represented as a vertex in the graph and is connected to its six nearest-neighbor vertices. The unary costs $c(p(u,v),u,v)$ of Eq. 1 are assigned to each edge that links a vertex (x,y,z) to its neighbor at $(x,y,z+1)$ and the pairwise costs $s(.)$ link it to its neighbors $(x+1,y,z)$ and $(x,y+1,z)$. More details regarding how to construct a graph where the minimum cut represents the minimum of the functional in Eq. 1 can be found in [12]. When the minimum cut is found, the target surface is defined along the transition between the voxels which are cut from their neighbors in the z direction.

The unary cost function $c(.)$ is chosen to be the normalized-cross-correlation (NCC) as [7] showed this gives good results in surgical surface reconstruction due to its ability to handle intensity changes that may be observed between images.

A significant challenge in reconstructing a surface is to achieve a good regularization of the solution. By exploiting knowledge of the surface shape from optically registered preoperative scans it is possible to bias the reconstruction to better represent the observed object. This prior is incorporated by weighting the unary matching cost between pixel patches with a modified robust Tukey

influence function of the residual between the estimated depth $p(u, v)$ of the correspondence and prior depth $d(u, v)$ of the registered 3D model:

$$w(x) = \begin{cases} \lambda + (1 - (1 - \frac{x^2}{\sigma^2})^3)(1 - \lambda) & \text{if } |x| \leq \sigma \\ 1 & \text{if } |x| > \sigma \end{cases} \quad (2)$$

where $x = p(u, v) - d(u, v)$ and the σ measure is obtained from an estimate of the inaccuracy of the initial optical registration, λ is a constant which shifts the values of the Tukey function to $\lambda \leq \text{Tukey(x)} \leq 1$, which was set at 0.3. This prevents overfitting to the prior which occurs when the disparity estimate exactly matches the prior depth estimate leading to a cost of zero for the node. The usual regularization technique of minimizing the first derivatives on the surface can still provide useful smoothness information.

As the most expensive step of the algorithm is building the graph, it is possible to optimize the performance of the algorithm by first building a low resolution version of the graph, reconstructing the surface and then refining this estimate. To achieve this the images are downscaled to half size and a low resolution voxel space is constructed. The surface is estimated and then a refinement graph is build using the full resolution images. This graph defines edge costs in the z direction to be the cost of shifting the disparity estimate by up to N pixels in either direction, where N is the subwindow size used in the NCC cost estimation. Here we define smoothness as before.

2.2 Registration of 3D Images to Laparoscopic Video

Registration of the 3D images to the stereo point cloud is performed using (1) A manual or calibration based registration. This contains inaccuracies caused by manual, calibration or tracking error (typically a rigid offset). (2) Rigid registration is used to compensate for errors in step one. (3) Non rigid registration is performed to account for tissue deformations.

Rigid and non rigid registration is performed using the Coherent Point Drift (CPD) [11] algorithm. The registration problem is posed as a probability density estimation problem where the reference point set is modeled as Gaussian Mixture Model (GMM) centroids which are forced to transform coherently to preserve topological structure. The registration parameters are optimized using the Expectation-Maximization (EM) algorithm. X is registered to Y, where X is the $N \times 3$ matrix corresponding to the $3D$ surface reconstructed from stereo images, as a realization of a GMM and Y is the $M \times 3$ matrix corresponding to the anatomical model extracted from the 3D volumetric image. The transformation T applied on Y is derived as $T(Y, R, t) = YR' + 1t'$ in the rigid case (with R a 3×3 rotation matrix and t a 3×1 translation vector), and as $T(Y, v) = Y + v(Y)$ in the deformable case (with $v(Y)$ corresponding to the displacement field).

The component density for the GMM $p(\mathbf{x}|m)$, which predicts the location of a target point $\mathbf{x}$ given a source point index for source point $\mathbf{y}_m$ is specified as an isotropic Gaussian with standard deviation σ, with the optimal σ. When combined, a single outlier model point matching all outlier points in the

target domain (with pre-specified outlier percentage w), this GMM negative log-likelihood yields the cost function $E(R, t, \sigma)$ in the rigid case. In the deformable case, an additional regularization term inducing spatial motion coherence via a regularization strength parameter λ and a regularization bandwidth parameter β is included in the cost function $E(v, \sigma)$. For rigid and non rigid, the EM framework is used to solve for the unknown transformation parameters and σ, collectively referred to by Θ.

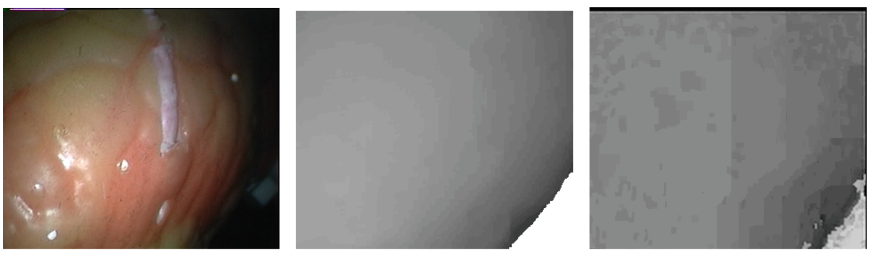
The E step provides a surrogate objective function $Q(\Theta; \Theta_{old})$ lying above $E(\Theta)$ and touching it at the previous estimate of the unknown parameters Θ_{old}. It computes a probabilistic match matrix P (incorporated within $Q(\Theta; \Theta_{old})$) between the model and target points:

$$p_{mn} = \frac{\exp\left(-\frac{1}{2\sigma^2_{old}}||x_n - T(y_m, \Theta_{old})||^2\right)}{\frac{w}{1-w}\frac{M(2\pi\sigma^2_{old})^{3/2}}{N} + \sum_{m=1}^{M}\exp\left(-\frac{1}{2\sigma^2_{old}}||x_n - T(y_m, \Theta_{old})||^2\right)} \tag{3}$$

In the M step of the algorithm, the surrogate $Q(\Theta; \Theta_{old})$ is minimized with respect to. Θ using an analytical formula in the rigid case and by solving the corresponding Euler-Lagrange equation in the deformable case, which in turn leads to a linear system with the various terms rapidly computable using the Fast Gauss transform.

3 Results

The proposed method is validated on a publicly available phantom dataset [7,13]; The silicon phantom is visually realistic and contains a mechanism for repeatable deformation. The ground truth was obtained from 16 CT scans of the phantom captured at various states of deformation. Two stereo video sequences of the phantom are used in this study. The CT data is aligned to the stereo camera via CT visible markers and provides disparity maps for stereo reconstruction validation and depth maps for registration validation. In vivo experiments are presented, however at the time of writing no 3D image data was available for validation (Fig. 2).



(a) Endoscopic Image (b) Ground Truth (c) Proposed Method

Fig. 2. An image of the phantom with the associated ground truth. The disparity maps have manually shifted histograms to improve visibility.

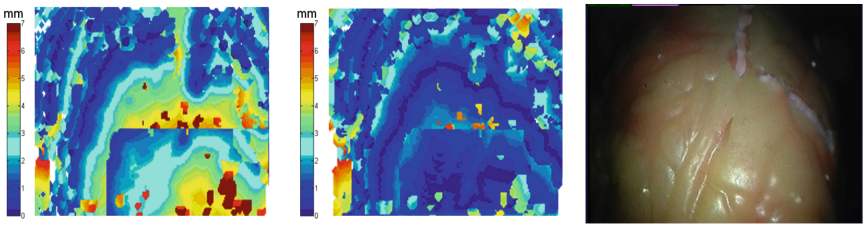
The stereo reconstruction is quantitatively evaluated with respect to the ground truth and compared to the state of the art [7]. An experiment was carried out to test the robustness of the proposed approach with respect to error in the prior derived from the 3D CT image. Offsets were applied to the prior to simulate inaccuracies. The translations range from 0–3.5 mm in the X, Y and Z axis and rotations around these axis from 0–18°. The offsets were applied to both video sequences creating a total of 32 datasets with known ground truth.

Quantitative reconstruction result are obtained using the RMS per pixel disparity error and percentage of pixels with a disparity error greater than a threshold of 5 pixels. The results from these experiments are shown in Table 1 and demonstrate that the reconstruction is improved through integration of prior knowledge and additionally provides significant improvement over using only the pre-aligned prior for reconstruction.

Table 1. The errors are shown using the proposed method, using the proposed method without using the prior and using the disparity map provided by the misaligned prior.

	RMS Error	% Error
Whole Image - Proposed method	2.89	5.45
Whole Image - Proposed method (Without Prior)	4.28	7.89
Whole Image - Prior only	7.11	9.53

The method of [7] provides highly accurate estimates of the disparity an enables us to benchmark our technique. It estimates disparity only over a subset of the pixels, achieving comparable accuracy (RMS error 1.96, % Error 1.41) to our own technique (RMS error 2.65, % Error 4.36) but does not provide complete surface coverage (estimating 96 % of pixels compared to our estimates of 99.5 %). To illustrate the stereo reconstruction accuracy Fig. 2 shows disparity maps of the stereo reconstruction algorithms.

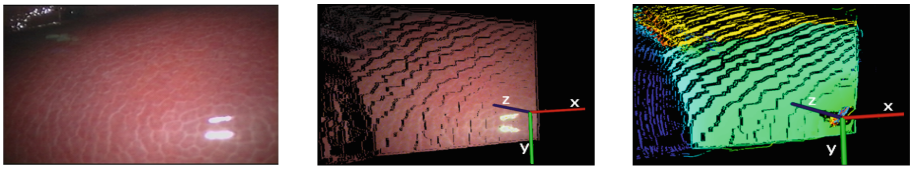


(a) (b) (c)

Fig. 3. Registration evaluation. Point to mesh error. (a) Before registration, (b) after rigid and non rigid CPD registration, (c) laparoscopic image

Table 2. Quantitative registration evaluation. Average point to mesh error in mm (standard deviation).

Dataset	ICP	CPD rigid	CPD non rigid
Rigid - Phantom Dataset 1	0.97 (0.88)	0.93 (0.93)	0.78 (0.98)
Rigid - Phantom Dataset 2	1.68 (1.61)	1.33 (1.48)	1.08 (1.50)
Non Rigid - Phantom Dataset 1	1.0 (0.74)	1.0 (0.8)	0.58 (0.72)
Non Rigid - Phantom Dataset 2	1.16 (1.0)	1.2 (1.2)	0.93 (1.24)



(a) (b) (c)

Fig. 4. An *in-vivo* reconstruction showing the original frames and the reconstructed point cloud.

The registration of the stereo surface reconstruction to the 3D image data is quantitatively evaluated on the phantom datasets. It was evaluated with respect the ICP which has been commonly used in the literature. Two experiments were carried out to evaluate the registration performance. Firstly, an experiment was conducted to evaluate performance with respect to errors in the initial alignment stage. Rigid offsets were applied to the 3D image using the same parameters defined above. The 32 datasets were evaluated with respect to the laparoscopic videos and results are provided in Table 2. The CPD non rigid results are initialized from the CPD rigid registration. Non rigid CPD out performs ICP and rigid CPD. This is increased performance is partly due to the non rigid registration which deforms the surface coherently to compensate for noise in the 3D image and the stereo reconstruction.

A second experiment was performed to evaluate the registration with respect to deformation. The surface reconstructed from each image in the video sequences was registered to 3D image corresponding to the first frame. This represents tissue deformation of up to 2 mm. Quantitatively evaluation is provided in Table 2. As expected the non rigid CPD performed best with respect to deformation. ICP and rigid CPD had similar performances. Figure 3 illustrates the point to mesh registration error as a heat map before registration and after rigid and non rigid CPD registration. It contains both rigid and non rigid transformations. Prior to registration the largest error occurs in the bottom right of the image. This error is significantly reduced after registration.

Qualitative *In-vivo* results are provided (Fig. 4) in the form of stereo reconstructions. These images demonstrate that the acquired surface visually corresponds to the target anatomy.

4 Conclusion

In this paper, a novel approach for registering 3D images to laparoscopic video is presented. The system incorporates prior information of patient specific organ surface into stereo reconstructions and uses robust non rigid registration to align the 3D images to stereo surface reconstructions. Its robustness to rigid offsets and non rigid tissue deformation is demonstrated on phantom data obtaining point to mesh registration errors of less than 1 mm. Future work will focus on the challenging issue of in vivo and clinical validation.

References

1. Su, L.-M., et al.: Augmented reality during robot-assisted laparoscopic partial nephrectomy: toward real-time 3D-CT to stereoscopic video registration. Urology **73**(4), 896–900 (2009)
2. Mirota, D.J., et al.: Vision-based navigation in image-guided interventions. Annu. Rev. Biomed. Eng. **13**, 297–319 (2011)
3. Maier-Hein, L., Schmidt, M., Franz, A.M., dos Santos, T.R., Seitel, A., Jähne, B., Fitzpatrick, J.M., Meinzer, H.P.: Accounting for anisotropic noise in fine registration of time-of-flight range data with high-resolution surface data. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 251–258. Springer, Heidelberg (2010)
4. Dumpuri, P., et al.: Model-updated image-guided liver surgery: preliminary results using intra-operative surface characterization. Prog. Biophys. Mol. Biol. **103**(2–3), 197–207 (2010)
5. Malti, A., Bartoli, A., Collins, T.: Template-based conformal shape-from-motion-and-shading for laparoscopy. In: Abolmaesumi, P., Joskowicz, L., Navab, N., Jannin, P. (eds.) IPCAI 2012. LNCS, vol. 7330, pp. 1–10. Springer, Heidelberg (2012)
6. Devernay, F., et al.: Towards endoscopic augmented reality for robotically assisted minimally invasive cardiac surgery. In: 2001 Proceedings of the International Workshop on Medical Imaging and Augmented Reality, pp. 16–20(2001)
7. Stoyanov, D., et al.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. MICCAI **13**(Pt 1), 275–282 (2010)
8. Lau, W.W., Ramey, N.A., Corso, J.J., Thakor, N.V., Hager, G.D.: Stereo-based endoscopic tracking of cardiac surface deformation. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3217, pp. 494–501. Springer, Heidelberg (2004)
9. Richa, R., Poignet, P., Liu, C.: Efficient 3D tracking for motion compensation in beating heart surgery. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 684–691. Springer, Heidelberg (2008)
10. dos Santos, T.R., et al.: Minimally deformed correspondences between surfaces for intra-operative registration. In: Proceedings of SPIE, vol. 8314, p. 83141C (2012)
11. Myronenko, A., et al.: Point set registration: coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 2262–2275 (2010)
12. Prince, S.: Computer Vision: Models Learning and Inference. Cambridge University Press, New York (2012)
13. Pratt, P., Stoyanov, D., Visentini-Scarzanella, M., Yang, G.-Z.: Dynamic guidance for robotic surgery using image-constrained biomechanical models. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 77–85. Springer, Heidelberg (2010)

A Novel Dual LevelSets Competition Model for Colon Region Segmentation

Huafeng Wang[1,4(✉)], Wenfeng Song[1], Lihong Li[3], Haixia Pan[1], Ming Ma[2], Weifeng Lv[5], Zhaohui Zhong[6], and Zhengrong Liang[2]

[1] School of Software, Beihang University of Beijing, Beijing 10083, China
wanghuafengbuaa@gmail.com
[2] Department of Radiology, Stony Brook University, Stony Brook, NY 11794, USA
[3] Department of Engineering Science and Physics, College of Staten Island, City University of New York, Staten Island, NY 10314, USA
[4] Northern China University of Technology, Beijing 10000, China
[5] School of Computer Science and Engineering, Beihang University of Beijing, Beijing 10083, China
[6] Department of Endoscopic Surgery, Peking University People's Hospital, Beijing 100034, China

Abstract. To segment the colon region is of much significance for colonic polyp's detection in Computed Tomographic Colonoscopy (CTC). However, not only the low contrast between CT attenuation values of the colon wall and the various surrounding tissues but also the pseudo enhancement effect by tag materials limit many traditional algorithms to achieve this task. Though few approaches suggested to depict colon walls by exploiting two steps: (1) find the inner colon wall; and (2) apply geodesic active contour based level set to extract outer boundary of colon wall, the failures happened when encounter the merging around houstral folds or adhesions of two very adjacent outer walls. Motivated by the observation that the interaction among 'forces' lead to a balance between the objects who caused those 'forces', we proposed a dual LevelSets competition model to simulate the mutual interference relationships among those compositions of the colon walls. Differ from the traditional LevelSet approach, the dual LevelSets competition model has a comprehensive cost function which take fully advantage of the essential characteristics of colon such as mixture, weak boundaries, volumetric, and so on. Compared with two already proved to be effective methods in literature: the graph cut and the geodesic active contour method, the proposed method has a much better performance to segment both the inner wall and the outer wall of colon. Both the comparison on if the method works well on weak boundaries of colon but also if it is capable of distinguishing the sticking boundaries of two very close walls is given. 200 CTC datasets are used to validate our proposed method. In conclusion, since the colon consists of various tissues, and they depend on and interact with each other, we could not consider the segmentation task in a static way, but a dynamic view works well.

Keywords: Dynamic · Levelset · Competition · Colon segmentation

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 117–128, 2016.
DOI: 10.1007/978-3-319-29965-5_12

1 Introduction

According to the recent statistics from American Cancer Society (ACS), colorectal cancer ranks the third most common occurrence of both cancer deaths and new cancer cases for both men and women in the United States. With the help of the computer assisted detection (CADe) and the computer assisted diagnosis (CADx), the colorectal cancer diagnosis process shall be facilitated. It is believed that to segment the whole colon region out be of much significance both for the CADe and for the CADx, where the segmented CT volume will help in determining potential polyps, muscular hypertrophy and diverticulitis of the colon [2]. And the accuracy of the segmentation also has effect on the sensitivity and specificity of the performance of CADe.

By measuring the segmented volume, not only the thickness but also the Shape Index, which are thought as much useful indicators for abnormal information on colon wall, can be highly reasonably calculated. As a kind of intuitive information display on the colon wall, the thickness is recommended as an effective way to show the abnormality based on the segmented volume.

However the thickness measurement depends on the segmentation result which we need to confirm if it is capable of describing the colon region reasonably and accurately to a full extent. In view of this, many researches bloomed in past decades [2, 3]. In general, many researches take three steps to get the whole colon region: (1) electronic colon cleansing (ECC) [4–6], which aims at removing the contrast agent; (2) determining the inner boundary of the colon region [7]; (3) separating the outer boundary from the colon region [8]. Implicitly, the most previous researches regard the acquired inner boundary and the outer boundary as the colon region without further distinguishing the voxels between them. Consequently, the merging phenomenon will happen as shown in Fig. 1, and it tends to lead to the any inaccuracy for thickness measure, thus the significance of the segmentation turns to less.

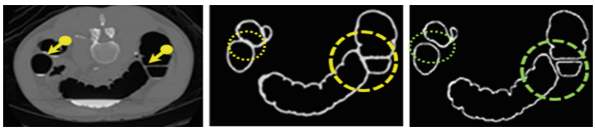


Fig. 1. An Illustration of merging result: the left is the raw slice, the middle is result given by previous approaches, and the right is the expected result; the yellow and green circles indicate the ROI (Color figure online).

The reason why this disadvantage happened on the traditional approaches can be attributed to the several existed challenges such as partial volume effect (PVE), low contrast, pseudo enhancement (PE) and so on [9]. Though all these challenges as shown in Fig. 2 have been thought over thoroughly before, the colon region has not been satisfying segmented yet because the merging on colon region still left unsolved.

Those existing segmentation methods can be grouped into two categories: exploiting static descriptors (SD) and using the dynamic descriptors (DD). SD methods

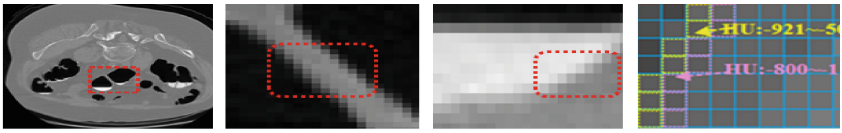


Fig. 2. The challenges on the colon region segmentation, red squares indicate the ROI (Color figure online)

suggested to fully use static characteristics in CT image such as pixel value, information contained in region [10, 11], gradient and so on. However, it has been proved that to simply use the SD method not be applicable for solving challenges above. On the contrary, DD approaches like LevelSet consider both the relationships among the SDs and the evolution discipline in a spatial and temporal context, so the above challenges were partially solved except the merging issue. In this paper, we will propose a new dynamic method which will estimate the colon regions by a novel competition model. Compared with the most popular methods such as GAC [12], and GraphCut [13], the newly proposed method shows much better performance on distinguishing the merging filed. For clearly understanding, here we define the colon region in the bending parts as BP, and the percentage for multiple tissue mixtures as 'mixrate'.

2 Related Work

Since the SD methods alone are not very robust to segment the reasonable colon region, here we will merely focus on those dynamic approaches. In literature, as one kind of most important dynamic methods, the LevelSet method and their derivatives, such as geodesic active contour (GAC) [3], DRLSE [14], decoupled active contour (DAC) [15] and so on, are highly recommended. They can be further classified into three main categories: the edge-based model [13], the region-based model [16] and the hybrid model [17, 18]. The geodesic active contour (GAC) model [3] for colon inner and outer boundaries as well as thickness measure (no thickness measure method was described in [2]) was proposed as one of the typical edge based models. In spite of its limited effect, the GAC model tends to enlarge the colon region thickness when near air-fluid boundaries. Recently, for the computational efficiency, Mishra et al. [14] proposed DAC for fast boundary detection, but as told by author that it won't work well on multiple tissues boundaries. It appears that the edge based methods put much emphasis on building a much faster LevelSet along the given edges, and the curve maybe cross the insufficient sharp boundaries [15]. While for the region based methods, such as Chan and Vese [16] model, they suggested to use a penalty term by measuring the distance between a piecewise constant approximation and the original image. However, the region based ones cannot work well when meet inhomogeneous objects due to ignoring the local features. As a complementary of respective strengths and weaknesses from both the edge based and the region based methods, the hybrid

models, such as Coupled Surfaces Propagation LevelSet [17] and couple LevelSet [18], evolve two embedded surfaces simultaneously driven by the image-derived information while maintaining the coupling. Though the previous hybrid method is capable of getting a volumetric region from bladder or cortex, they also suggest adapting this idea to new segmentation problem. Besides the LevelSet methods, the multiplicative intrinsic component optimization (MICO) [19], and MAP-EM [1] perform well in the task of segmentation of weak boundaries(with Low contrast, PVE and PE existed in image), but they are not good at keeping the reasonable contours of the objects since they tend to ignore the edge information.

Motivated by the hybrid method, we are trying to propose a competition LevelSet model by fully considering both the local features and the regional statistics, which is capable of describing the clear BP boundaries with help of a newly probability graph constraint term.

3 Method

At the beginning of the colon region segmentation, MAP-EM model [1] is introduced to complete the task of electronic colon cleansing (ECC). Since ECC has been discussed for years and tag material appears in CT image to vary from the background a lot, the new proposed method will focus on the colon region segmentation model based on the result outputted by MAP-EM ECC. As told above, our model is a combination of dynamic evolution and various order features, hence, the feature descriptors in each order we need to explain first.

3.1 Order Feature Descriptors in CT Image

According to the characteristics, such as pixel value (or voxel in 3D), gradient (first order) and curvature (second order) and so on, are usually applied to the process of segmentation. In terminology, I stands for raw data from CT image. An adequate volume image is defined as,

$$f(x, y, z; t) = G * I \tag{1}$$

where (x, y, z) is the Cartesian coordinates, and t is the scale and G is a Gaussian kernel; The gradient is expressed by

$$\nabla f = (f_x, f_y, f_z)^T \tag{2}$$

∇f represents the gradient.

$$g(I) = \frac{1}{1 + |\nabla f|^2} \tag{3}$$

g means the edge indicator;

$$k = \nabla(\frac{\nabla\phi}{|\nabla\phi|}) \tag{4}$$

where k defines curvature, and the surface tension is expressed as $k\delta_\varepsilon(\phi)\overrightarrow{n}$, where $\delta_\varepsilon(\phi)$ is the Dirac function, $\overrightarrow{n}$ is the normal vector and Φ is a LevelSet function; $\psi(\phi) = H_\varepsilon(\phi)$ acts as a Heasvide function to judge difference between pixels by the step function, and ε is the step.

Though the high order features are believed to be capable of describing geometrical information very well, it is believed that the statistical information held in raw image should help us a lot in identifying the distinguishing details among the neighborhood pixels. Thus, we define a zero order feature descriptor as bellow.

Since the colon region usually consists of various components, such as air, muscle, bone, Mucosa and so on, we need to 'know' what kind of components the current pixel probably belongs to. As shown in Fig. 3, the HU values appear to have overlaps among the possible components. So the pixel property is not possible determined in a simple threshold way. Enlightened by MAP-EM approach [1], a given pixel is regarded as a mixture of above mentioned types, such as air, mucosa, muscle and bone or tag material and so on. Hence we can simply evaluate the probability or the percentage of the known types in a given pixel. Thus we define the $tissue_i$ as the probability of a given pixel belongs to the i-th type, and the type labels we use for the colon region segmentation task are enumerated as {Air, Mucosa, Bone or Tag, Muscle}. Therefore, we have $\sum_{i=0} p(tissue_i) = 1$, where p stands for the probability.

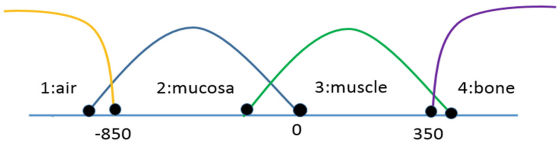


Fig. 3. The HU value range on abdomen CT image

Let T_i be the predicted result of a given pixel (x,y,z), i.e., which indicates if the pixel belongs to i-th tissue type. Then we have T_i to be defined as,

$$T_i(\mathrm{x,y,z,val}) = \begin{cases} 1, \textit{ if } \mathrm{tissue}_i \textit{equals to } \arg\max\limits_{tissue_i \in \Theta} p(\mathrm{val}|\mathrm{tissue}_i) \\ 0, \textit{ otherwise} \end{cases} \tag{5}$$

Where **val** is the HU value of the given pixel, **Θ is** the sample space for all the available tissue types. According to the Bayesian formula, $p(\mathrm{val}|tissue_i)$ can be calculated by,

$$p(\mathrm{val}|tissue_i) = \frac{\iiint_\Omega p(\mathrm{tissue}_i|\mathrm{val})\mathrm{p}(\mathrm{val})\mathrm{d}\Omega}{p(\mathrm{tissue}_i)} \tag{6}$$

Where Ω indicates the domain of the input raw data. The function p integrates all the points in the region of the domain. Please note, $p(\text{val}|tissue_i)$ is calculated based on the intuitively manually drawing regions in advance.

As shown in Fig. 4, the drawing region will be calculated automatically and we use the frequency percentage to express the $p(tissue_i|val)$ in line with the Large Number Law. Then $p(\text{tissue}_i)$ will be decided by,

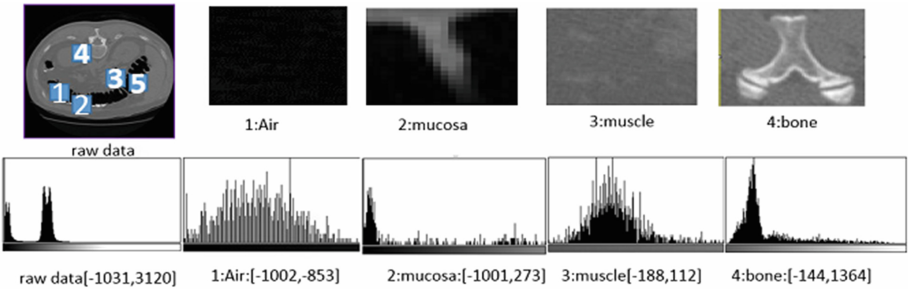


Fig. 4. The statistical HU value range of the given types

$$p(tissue_i) = \frac{\iiint_\Omega T_i(I(x,y,z))d\Omega}{\iiint_\Omega T(I(x,y,z))d\Omega} \tag{7}$$

Now, we can define the "mixrate" mentioned above of a given image I(x,y,z) as,

$$mixrate(I(x,y,z),t) = -\sum_{i=1}^{n} p(tissue_i) \cdot \log(p(tissue_i)) \tag{8}$$

where the negative sign stands for a bias to a homogeneous region, and mixrate (I(x,y, z),t) stands for the mixrate mapping for I(x,y,z) when t-th iteration finished.

3.2 Dynamic Forces Terms and the Competition Model (CM)

All the above discussed static order features descriptors are very meaningful for construct a dynamic model. For decades, LevelSet method is very popular to exploit an implicit surface to help depicting the boundary of a given object in a dynamic way. And it is usually expressed by an energy function with time as follows,

$$\phi_0(x,y,z,t=0) = \begin{cases} -i_0\ ,\ if\ (x,y,z) \in R_0 \\ i_0\ ,\ otherwise \end{cases} \tag{9}$$

Where Φ is a LevelSet function, i_0 is a positive constant, R_0 is the inner region of ϕ_0, t is time(or evolution time step) and (x,y,z) is the Cartesian coordinates. Φ will evolve to the boundary of the colon region constrained by a defined '**force**': $\boldsymbol{F}$,

$$\frac{\partial \phi}{\partial t} = F|\nabla \phi| \tag{10}$$

Since the forces defined in those previous methods tend to fail BP boundaries. Hence, a new force definition is necessary for achieving the expected result. Meantime, a cost function $L(x, y, z, t)$ is also required to help those forces converge.

In view of the previous defined LevelSet is not able to evolve into a much concave region by a single force, we need to consider more forces to work together. So we define a kinetic term,

$$\text{kinetic}_{g(I_i)} = \iint_{\Omega} g(I_i)\delta_\varepsilon(\phi)|\nabla \phi| ds \tag{11}$$

where $\boldsymbol{i}$ is the index. And let forces from different regions to compete as,

$$\varepsilon_1 = w_1\alpha_1 Kinetic_{g(I_1)}(\phi_1) + w_2\alpha_2 Kinetic_{g(I_2)}(\phi_2) \tag{12}$$

where w_i is the statistics weight coefficient of the term, α_i is a positive constant, the energy $Kinentic_g(\phi)$ computes the surface integral of the function **g** along the zero LevelSet. After parameterizing the zero LevelSet of ϕ as a surface, $\mathbf{I_1}$ is the input region to identify the boundaries, $\mathbf{I_2}$ is a guide image. In practice, it is a convex box of the object, and **it** has an opposite sign to $\mathbf{I_1}$. When LevelSet stops outside the sharp corner and the concave hull, $\mathbf{I_2}$ will help let the LevelSet continue until it reach the real interfaces because the double kinetic terms in the opposite sign will make the convex contour evolve into a concave one. However, if the LevelSet is evolving toward to a right and effective direction? Especially for the multiple regions, where the evolution directions are neither single nor consistent, the situation is getting much more difficult. Since energy functional is the integral of g on volume of Ω, a double anchors term ε_2 is given as,

$$\varepsilon_2 = w_1\beta_1 Anchor_{g(I_1)}(\phi_1) + w_2\beta_2 Anchor_{g(I_2)}(\phi_2) \tag{13}$$

Where Anchor is calculated by $\iint_{\Omega} gH_\varepsilon(-\phi)ds$, β_2 is a constant. Thus, the integration calculation among the neighbor regions is completed through changing the coefficient's sign of β_i, which will control the direction of the evolution. As an indirect effect, the changing of signs makes an increase or decrease on $|\nabla\phi|$, then change the direction of $|\nabla\phi|$. Moreover, the forces from the two terms is shift when the distance changes. When evolving close to the mixture boundaries, the second term takes a leading role and suppresses the fusion of two regions. In a word, it can be used to control the evolution of the step length and avoid excessive fusion.

Because the general size of colon CT data is up to 512 * 512 * 700 after interpolation, the fixed step evolution is apt to cause the LevelSet evolution with a very low efficiency. As told in [14], Potential energy will help speed up the evolution with an accuracy edge. And the traditional P term is computed by $\iint || \nabla\phi - 1||^2 ds$. So by the iteration, we have,

$$\frac{\partial P}{\partial t} = [\nabla^2\phi - div(\frac{\nabla\phi}{|\nabla\phi|})] = div((1 - \frac{1}{|\nabla\phi|})\nabla\phi) \tag{14}$$

By taking full advantage of the above terms, a cost function is defined as,

$$\begin{aligned} L = {} & W \cdot \mu Potential(\phi) + W \cdot A \cdot [Kinetic_{g(I_1)}(\phi), Kinetic_{g(I_2)}(\phi)]^T + \\ & W \cdot B \cdot [Anchor_{g(I_1)}(\phi), Anchor_{g(I_2)}(\phi)]^T \end{aligned} \tag{15}$$

Where W represents $[w_1, w_2]$, $w_i = \frac{1}{1+mixrate}$, **A** stands for $[\alpha_1, \alpha_2]$, and **B** means $[\beta_1, \beta_2]$. α_i is a positive constant with $\alpha_i \in R$. W represents the weight of the evolution step, where higher is the mixrate, lower is the evolution speed. When reaching the maximum, it will stop. Thus, the gradient decent flow will be,

$$\frac{\partial\phi}{\partial t} = \sum_{i=1}^{2} w_i\alpha_i\delta_\varepsilon(\phi_i) div\left(g(I)\frac{\nabla\phi_i}{|\nabla\phi_i|}\right) = \sum_{i=1}^{2} w_i\alpha_i\delta_\varepsilon(g|\nabla\phi|k + \nabla g \cdot \nabla\phi) \tag{16}$$

where k is the curvature and div is divergence operator. In conclusion, the above proposed model adequately considers not only the forces' direction but also their mutual interactions with two LevelSets (Φ_1, Φ_2), so, we call it as a dual LevelSet competition model.

4 Experiments

We selected a CTC database of 100 patients with 200 CT scans from both supine and prone positions from the Wisconsin hospital. All the selected datasets are in the DICOM formats, and the slice number ranged between 480 and 700.

4.1 Experiments for the Selection of the Parameters

Since the parameters are of much significance on the evaluation of the proposed algorithm, we set the several tests on relationships and the parameters. At the beginning, we select 5 datasets as the inputs, and apply a range of parameters to them. Then the results were grouped according to the parameter's value. As for the evaluation, 10 trained students are required to give each own scores on the results through a web system we built in advance. Finally the scores are collected automatically and averaged to show the performance based on each parameter. As shown in Fig. 5, most of the parameters have their respective best performance except α2.

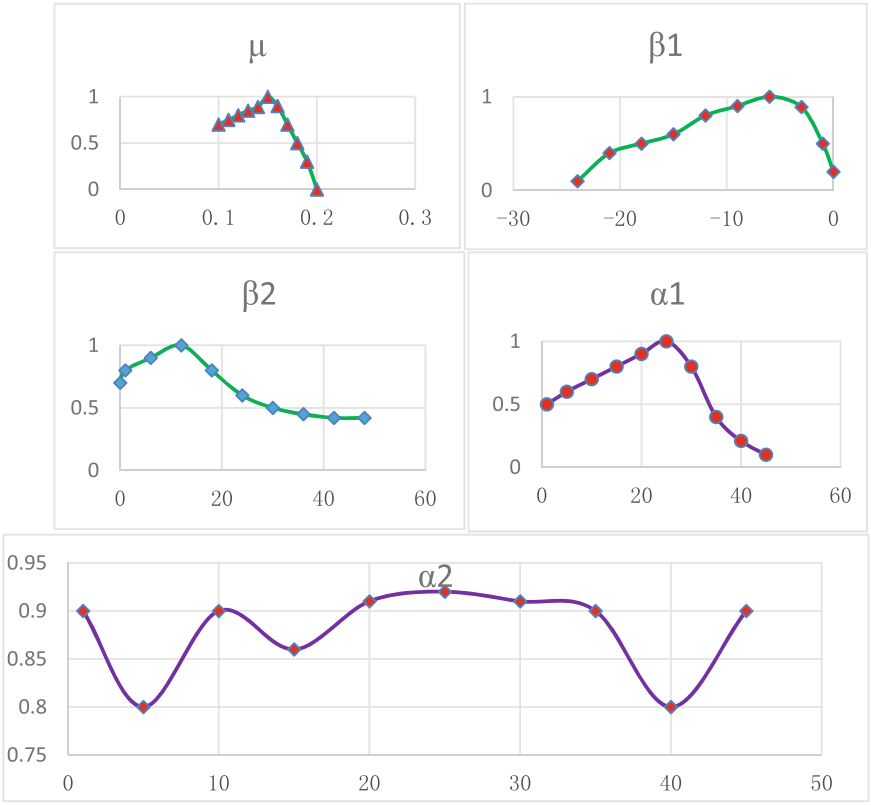


Fig. 5. Performance on different parameters respectively (x axis is the actual selected parameter value range and y axis is the normalized performance score)

According to the observation in the experiments, we also found that the higher the mixrate is, $|\beta_2|$ is larger. And when the mixrate is higher, the implicit surface evolves at low speed until a balance status achieved.

4.2 Comparison with the Other Methods

Based on the experimental results, we can intuitively compare the results given by CM with others such as GAC, GACK [13], GraphCut [14], MICO [19] and so on. As illustrated in Fig. 6, the proposed method CM shows much more gain than that of the rest because the results produced by GAC or GACK couldn't extend adequately into the concave wall, meantime the edges are rough and not keeping shapeness as the colon actually does.

Compared with GraphCut, the new approach performs better when dealing BPs on the colon wall. As shown in Fig. 7, CM keeps BPs very well, while the GraphCut misses some BPs' shape detail.

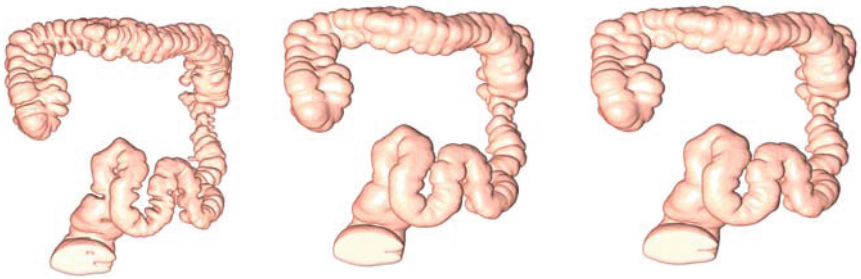


Fig. 6. The 3D rendering results for different methods.

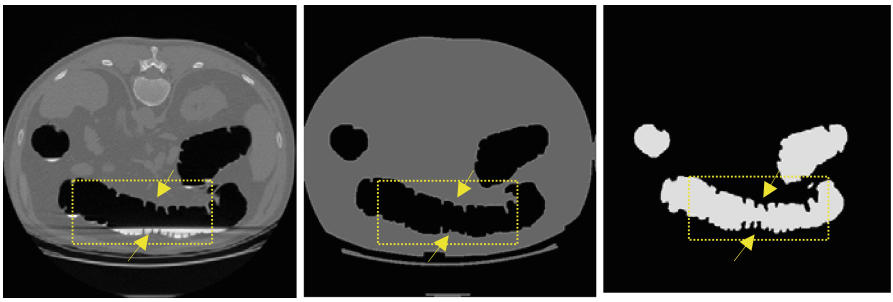


Fig. 7. The comparison between GraphCut and CM: left is the raw slice, middle is the corresponding segmented result given by GraphCut, right is produced by CM.

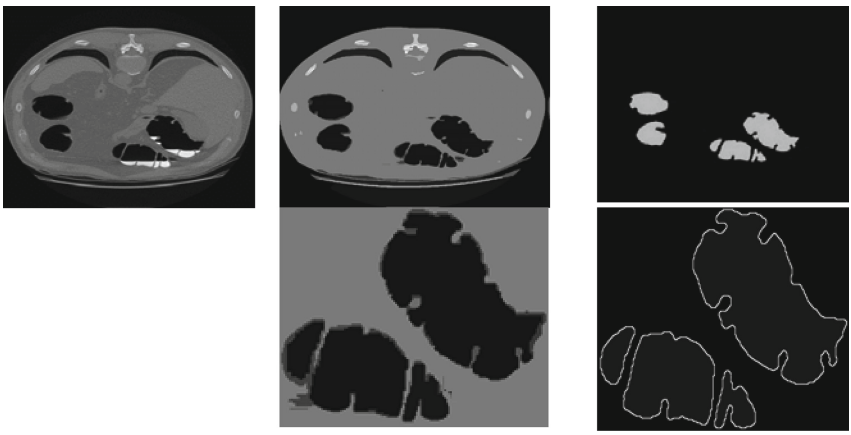


Fig. 8. The comparison between the statistics merging approach (SM) and CM: second row indicates the zoomed details, middle is the result given by SM and its' zoomed details, right is the result given by CM and its' zoomed details.

When we dig the segmentation details, a very interesting conclusion is also drawn by observation: the CM method is capable of both keeping shape of the colon wall and keeping the shape smoothly. As shown in Fig. 8, the left is raw slice, and the middle is the result given by statistics merging approach. Compared with the statistics merging approach, CM can preserve much smoothness of the shape.

5 Conclusion and Discussion

In this paper, we presented a new level set model, which has an intrinsic capability of segmenting the weak boundary and be able to depict BP regions. As shown in the experimental results, the model is feasible for colon region segmentation and illustrated a better performance than the previous methods. We regard our proposed model as a dynamic one, which not only takes fully the information of image features into consideration, but also combines the LevelSet method with the above mentioned descriptors in a mathematic way. In practice, we also observed that the terms we defined above are able to work properly as we expected. For example, double kinetic terms will smooth the result in the uniform way, while for the Anchor terms, they control segmented result in the opposite direction. With the energy function, the model shows many merits solving the weak boundaries segmentation problem.

References

1. Wang, S., et al.: An EM approach to MAP solution of segmenting tissue mixture percentages with application to CT-based virtual colonoscopy. Med. Phys. **35**(12), 5787–5798 (2008)
2. Van Uitert, R.L., Summers, R.M.: Colonic wall thickness using level sets for CT virtual colonoscopy visual assessment and polyp detection. In: Medical Imaging. International Society for Optics and Photonics (2007)
3. Chen, D., et al.: Accurate and fast 3D colon segmentation in CT colonography. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009. IEEE (2009)
4. Li, L., et al.: A new electronic colon cleansing method for virtual colonoscopy. In: Medical Imaging. International Society for Optics and Photonics (2007)
5. Zhang, H., et al.: An integrated electronic colon cleansing for CT colonoscopy via MAP-EM segmentation and scale-based scatter correction (2012)
6. Zhang, H., et al.: Integration of 3D scale-based pseudo-enhancement correction and partial volume image segmentation for improving electronic colon cleansing in CT colonograpy. J. X-ray Sci. Technol. **22**(2), 271–283 (2014)
7. Liang, Z., Wang, S.: An EM approach to MAP solution of segmenting tissue mixtures: a numerical analysis. IEEE Trans. Med. Imaging **28**(2), 297–310 (2009)
8. Nordin, N., et al.: Wall thickness measurement of colon based on ultrasound image segmentation. In: 1st WSEAS International Conference on Biomedicine and Health Engineering (BIHE 2012) (2012)
9. Zhang, H., et al.: Integration of 3D scale-based pseudo-enhancement correction and partial volume image segmentation for improving electronic colon cleansing in CT colonograpy. J. X-ray Sci. Technol. **22**(2), 271–283 (2014)

10. Mancas, M., Gosselin, B., Macq, B.: Segmentation using a region-growing thresholding. In: Electronic Imaging 2005. International Society for Optics and Photonics (2005)
11. Chang, K., et al.: Automatic colon segmentation using isolated-connected threshold. In: 2011 First International Conference on Robot, Vision and Signal Processing (RVSP), pp. 44–47. IEEE (2011)
12. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. **22**(1), 61–79 (1997)
13. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary region segmentation of objects in N-D images. In 2001 Proceedings of the Eighth IEEE International Conference on Computer Vision. ICCV 2001. IEEE, Vancouver (2001)
14. Chunming, L., et al.: Distance regularized level set evolution and its application to image segmentation. IEEE Trans. Image Process. **19**(12), 3243–3254 (2010)
15. Mishra, A.K., Fieguth, P.W., Clausi, D.A.: Decoupled active contour (DAC) for boundary detection. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 310–324 (2011)
16. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
17. Xiaolan, Z., et al.: Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation. IEEE Trans. Med. Imaging **18**(10), 927–937 (1999)
18. Duan, C., et al.: A coupled level set framework for bladder wall segmentation with application to MR cystography. IEEE Trans. Med. Imaging **29**, 903–915 (2010)
19. Li, C., Gore, J.C., Davatzikos, C.: Multiplicative intrinsic component optimization (MICO) for MRI bias field estimation and tissue segmentation. Magn. Reson. Imaging **32**(7), 913–923 (2014)

Enhancing Normal-Abnormal Classification Accuracy in Colonoscopy Videos via Temporal Consistency

Gustavo A. Puerto-Souza[1(✉)], Siyamalan Manivannan[2], María P. Trujillo[3], Jesus A. Hoyos[4], Emanuele Trucco[2], and Gian-Luca Mariottini[1]

[1] Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
gustavo.puerto@mavs.uta.edu
[2] CVIP, School of Computing, University of Dundee, Dundee, UK
[3] Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Cali, Colombia
[4] Hospital Universitario del Valle Evaristo García ESE, Cali, Colombia

Abstract. This paper proposes a novel hierarchical approach to improve the accuracy of the classification of normal-vs-abnormal frames in white-light colonoscopy videos. The existing approaches label each frame independently, without considering the temporal consistency between adjacent frames. Temporal consistency, however, can improve the classification accuracy in the presence of unclear/uncertain images. We propose to leverage temporal consistency between adjacent frames for colonoscopy video frame classification using a novel hierarchical classifier. Comparative experiments with five challenging full colonoscopy videos show that the proposed approach considerably improves the mean class normal/abnormal classification accuracy compared to the approaches where the frames are classified independently.

1 Introduction

Colorectal cancer is the second leading cause of cancer death in the world and the third most common cancer in the UK [1]. Although colonoscopy remains the gold standard for colorectal cancer screening, its miss rate for colorectal cancer has been reported to be as high as 6 % [2], posing the risk of developing colon cancer due to failure to detect treatable lesions in time. This motivates research into automated, repeatable systems detecting abnormalities (including polyps, cancer, ulcers, etc.) in colonoscopy videos, which could provide a second quantitative opinion and ultimately contribute to reduce the miss rate.

In this paper, we concentrate on classifying white-light colonoscopy images into 2 classes, normal and abnormal. Abnormal frames contain one or more lesions (e.g., polyps, adenomas); normal frames contain none and show a healthy colon wall. The majority of the work reported for colonoscopy image classification focuses mainly on designing or identifying appropriate features and classifiers.

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 129–139, 2016.
DOI: 10.1007/978-3-319-29965-5_13

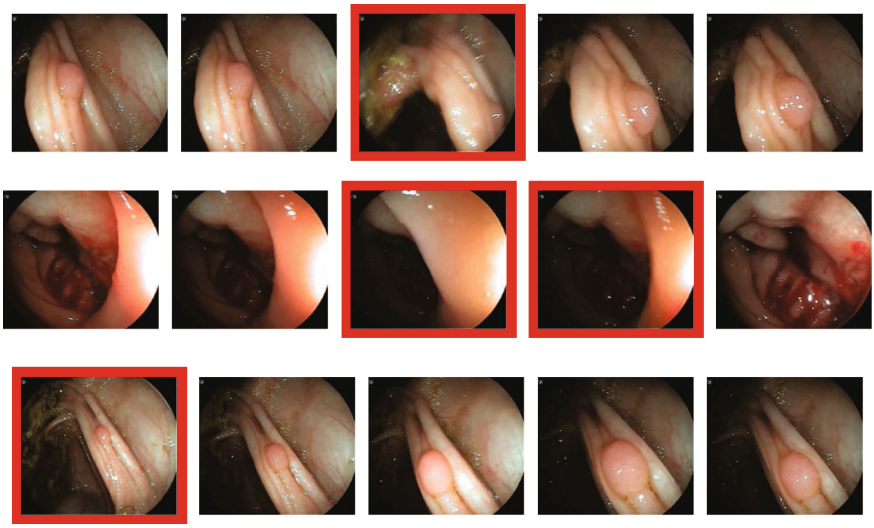


Fig. 1. Examples of three small video segments each contains 5 frames. The images which are difficult to classify due to (1) the lesion is not visible properly, (2) poor illumination, and (3) a very small lesion is highlighted in the 1st, 2nd and 3rd rows, respectively. These images, however, could be correctly classified as abnormal if the temporal information between adjacent frames were considered.

Texture, color, shape and their combinations, together with different classifiers, such as SVM and neural nets, have been explored for lesion detection and/or frame classification: texture features for normal/abnormal classification [3–5], lesion detection [6–8]; color histograms and related statistics for bleeding detection [9,10]; and shape-based features, such as edge orientation histograms for Crohn disease classification [11]. For a complete review of the aforementioned methods, we direct the reader to [12].

Up to our knowledge, the state-of-the-art colonoscopy video frame classification approaches assume frames independent of each other. In reality, if a lesion appears in a particular frame, previous and successive frames are very likely to include it, albeit from different viewpoints as the scope is moved. One expects, therefore, that temporal consistency should improve the accuracy of colonoscopy frame classification compared to single-frame schemes.

There are further reasons to expect that temporal consistency will improve the classification. First, some frames are genuinely ambiguous, and a single view will not be sufficient for reliable classification even for experts, whose decisions are based on multiple observations generated by moving the scope. Second, the colonic wall may not be clearly visible in specific frames due to poor illumination, blur due to fast camera movements, and surgical smoke. Third, the appearance of lesions (e.g., scale, orientation) varies in different frames. Fourth, frame-level representations for classification are often obtained by aggregating the statistics of the local features extracted from that frame (e.g. bag-of-visual-words).

Such representations may not capture small lesions sufficiently well, vis-á-vis the volume and appearance of background features (extracted from normal tissue).

Figure 1 shows three example video sequences, each containing a few frames which are difficult to classify. A system trained on individual frames independently is likely to classify these frames erroneously as normal. However, a classifier using temporal consistency information would classify these frames correctly as abnormal.

In this paper, we propose a three-level hierarchical classification approach which makes use of the temporal-context information across adjacent frames to classify any individual frame. In the first level, we assume the frames are independent to each other, hence we learn a classifier based on individual frame-level representations. The second level classifier is trained to leverage the temporal consistency information using the weighted similarities between frames in a temporal window and the classification outputs computed from the first level. We propose a max-margin approach to learn these weights based on the given training set. The third level applies a temporal filtering which refines the output from the second level by majority voting. We experimentally show that the proposed hierarchical approach outperforms the single-level classifier approaches such as SVM and random forests which were trained to classify frames independently. Note that our technique could be used to assess proficiency of gastroenterologist doctors either by analyzing colonoscopy videos both retrospectively or in real time depending on the parameters of the sliding window.

In the following, we first we explain the proposed hierarchical classification approach in detail, and then provide experimental evidence showing that the proposed approach performs better than any single level classification approach.

2 Methodology

In this section, we present an algorithm to classify normal-abnormal frames in colonoscopy videos. Our approach is based on a three-layer hierarchical classifier that leverages the strengths of SVM, in terms of accuracy and robustness, and the temporal consistency between adjacent frames based on a max-margin formulation.

In our proposed approach, we make use of the similarities between adjacent frames, in addition to the frame-level features. The similarities (e.g., number of image correspondences) between adjacent frames play an important role in this classification. Lets consider two consecutive frames I_i and I_j, if I_i has a high similarity with I_j it is most probable that both I_i and I_j are belonging to the same class.

Our approach is illustrated in Fig. 2. In the first level, frames are assumed to be independent to each other, and a SVM is trained to classify frames independently based on the frame-level features. In the second level, we make use of the temporal-context information between adjacent frames; which are measured by weighted similarity between a frame and its temporal neighbors, as well as the outputs obtained by the first level classifier. We propose an approach to learn

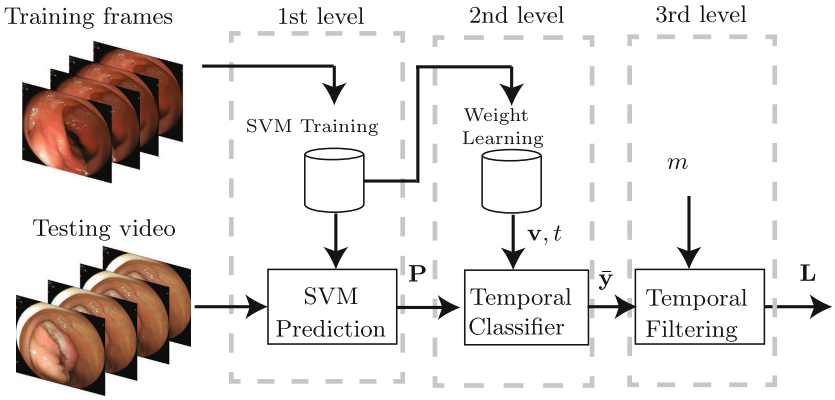


Fig. 2. The proposed hierarchical classifier. The first level outputs the confidence values based on classifying independent frames. The second level uses these confidence values in addition to the similarities between adjacent frames. The final level applies a majority voting on the second-level outputs to obtain the final labels of individual frames.

these weights by maximizing the margin between normal and abnormal classes. Finally, the resulting classification is passed to a third level that refines further the output from the second level by using a voting scheme over adjacent frames.

In the following, first we describe the first-level classifier and the Platt scaling which is used to convert the outputs of the first-level classifier to probability values. Then, the max-margin formulation of the second-level classifier is explained in detail. Lastly, the section concludes with the temporal filtering.

2.1 The First-Level Classifier

This classifier is trained on individual-frame representations to classify each test frame independently, i.e. without considering its temporal context.

Since the number of abnormal and the normal frames are highly unbalanced, we use a SVM with class balancing [13]. Learning the SVM weight vector $\mathbf{w}$ and the bias (b) for the first-level classifier $f(\mathbf{x})$ is achieved by the following formulation,

$$\arg\min_{\mathbf{w},b} \left\{ \|\mathbf{w}\|^2 + \lambda \left[C^+ \sum_{i \in A} h(\mathbf{w}^T \mathbf{x}_i + b, y_i) + C^- \sum_{j \in N} h(\mathbf{w}^T \mathbf{x}_j + b, y_i) \right] \right\} \quad (1)$$

where h is the hinge loss function $h(z, y) = \max(0; 1 - yz)$, with $\mathbf{x}_i$ and $y_i = \{-1, 1\}$ are the feature representation for I_i (the i^{th} frame) and its label, respectively. λ is a regularization parameter controlling the rate of misclassification, and C^+ and C^- are the class weighting parameters for the unbalanced abnormal (A) and the normal (N) classes, respectively. C^+ and C^- can be selected by setting $\frac{C^+}{C^-} = \frac{n^+}{n^-}$ [13], where n^+ and n^- are the total number of positive (abnormal) and the negative (normal) images in the training set.

Usually SVM outputs decision values represent how far the test feature is from the learned hyper-plane, which is defined by $(\mathbf{w}, b)$. The Platt calibration method [14] maps any SVM output $f(\mathbf{x}_i)$ with the range $[-\infty, \infty]$ to a posterior probability P with the range $[0, 1]$ by a sigmoid function, i.e.,

$$P(y = 1|f(\mathbf{x}_i)) = \frac{1}{1 + \exp(Af(\mathbf{x}_i) + B)} \tag{2}$$

where $P(\mathbf{x}_i)$ represents the probability of the ith image being positive. A and B are two parameters which has to be learned from the training set. As suggested by Platt [14], we use a three-fold cross validation on the training set to learn these parameters.

2.2 The Second-Level Classifier

This classifier aims to improve the classification accuracy of the first classifier by leveraging temporal consistency. The inputs are the probabilities obtained by the first-level classifier, as well as the similarities, in terms of image correspondences, between a frame and its neighbors.

Similarity Between Frames: We defined the similarity S_{ij} between two adjacent frames, I_i and I_j, as the number of image correspondences between them. In particular, we extract and match SIFT features because of their stability, distinctiveness, and repeatability, as well as their well known rotation and scale invariance, and robustness to affine distortions, illumination changes, and noise [15]. SIFT detects a sparse set of interest points (keypoints), in the image, obtained as the scale-space extrema of the difference of Gaussians operators. The extracted keypoints are matched according to the nearest neighbor distance ratio of their descriptors, discarding ambiguous matches with ratio greater than 0.8 [15].

The Temporal Classifier: The proposed temporal classifier assumes that the label of a particular frame I_i not only depends on the classification results of itself, but also on the weighted similarity between that frame and its neighbors as well as on the confidence values of its neighbors. From here and the following we will assume a centered sliding window since our approach targets for maximal performance over retrospective videos. However, our approach can achieve real time performance by using a queue-style sliding window.

Let $P_i = P(y_i = c)$ and $P_j = P(y_j = c)$ represents the probabilities obtained by the first-level classifier for the frames I_i and I_j. We define the label of the frame I_i based on the temporal classifier as follows,

$$\begin{aligned} d_i &= v_i P_i + \sum_{\substack{j=-n \\ j \neq 0}}^{n} v_j S'_{i,j} P_j \\ \bar{y}_i &= \begin{cases} 1 & \text{if} \quad d_i \geq t \\ -1 & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

where the set $\{v_j\}_{j=-n}^{n}$ are the weights applied to the current frame ($j = 0$) and its neighboring frames in the interval $[-n, n]$. Here t denotes the margin between classes, the size of the considered temporal window is represented by $2n+1$ (i.e., previous n and next n frames are considered around the frame I_i), and $\bar{y}_i$ is the predicted label for the frame I_i. S'_{ij} can be represented by

$$S'_{i,j} = 1 - \exp^{-\beta S_{i,j}} \tag{4}$$

where β is a decay parameter, empirically set to $\beta = 5$ in all the experiments reported in Sect. 3.

Lets define the vectors $\mathbf{v}$ and $\mathbf{u}$ be

$$\mathbf{v} = \begin{pmatrix} v_{i-n} \\ \vdots \\ v_{i-1} \\ v_i \\ v_{i+1} \\ \vdots \\ v_{i+n} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} P_{i-n}S_{i,i-n} \\ \vdots \\ P_{i-1}S_{i,i-1} \\ P_i \\ P_{i+1}S_{i,i+1} \\ \vdots \\ P_{i+n}S_{i,i+n} \end{pmatrix} \tag{5}$$

The classifier defined in Eq. (3) can be represented based on vector representations as follows,

$$\bar{y}_i = \begin{cases} 1 & \text{if} \quad \mathbf{v}^T\mathbf{u} - t \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{6}$$

where $\mathbf{v}$ and t define the temporal classifier, and can be easily learned in a similar manner to the max-margin approach given by Eq. (1).

2.3 Temporal Filtering

This final level refines further the results of the second-level by enforcing, within a sliding window, a temporal constraint based on the classes of the surrounding frames. As a result, the video has smoother transitions between abnormal and normal classes, i.e., the labels of video frames in segments containing lesions are consistently "abnormal", and do not contain noisy "normal" labels surviving the previous classifiers.

We use the second classifier prediction $\bar{y}_i$ to classify the frames, based on a majority-vote scheme over a sliding window. In particular, for each frame I_i, we gather the second-level classifier labels within a window with size $2m + 1$, centered on frame i. Each element within the window yields a vote for either abnormal or normal according to their class $\bar{y}_i$. The frame I_i is classified as the class with the larger number of votes. For example, the frame I_i is classified as abnormal if $C^A_{i,m} > C^N_{i,m}$, where $C^A_{i,m}$ and $C^N_{i,m}$ denote the number of votes for abnormal and normal classes within the window, respectively.

3 Experiments

The aim of these experiments is to compare different classifiers, with and without the hierarchical approach to incorporate temporal consistency, while keeping all the other factors unchanged, e.g. features for computing the frame representations.

In the following dataset, experimental settings and evaluation criteria are first explained. Then experimental validation and analysis of the results are presented.

3.1 Experimental Setup

We define abnormal frames as those that contain various lesions including polyps, cancer and bleeding. Our dataset consists of frames extracted from five colonoscopy videos (1 normal and 4 abnormal) from Hospital Universitario del Valle Evaristo Garcia ESE, Cali, Colombia. Each video has length of 8–15 min, image resolution of 640×480 and was recorded at 10 fps, leading to a total of 41518 extracted frames. For training and evaluation, the entire dataset was annotated at frame-level by an expert colonoscopist. In our two-label scheme and since lesion detection is the clinical target, large blurs and negligible frames were labeled as normal. The number of frames from different classes are given in Table 1; notice that the normal frames (N) constitute 77.5 % of the dataset while the 22.5 % of the frames are labeled as abnormal (A). All these frames were then rescaled by preserving their row to column aspect ratio to make their maximum size (row or column) equal to 300 pixels.

Table 1. The number of frames per video in each class (N-normal, A-abnormal)

Video	N	A	%A frames
1	5173	2944	36.3
2	3082	2555	45.3
3	8033	2056	20.4
4	5892	1823	23.6
5	9960	0	0
Total	32140	9378	22.6

Frame Representation: Each frame in the dataset was represented based on the Locality-constrained Linear Coding (LLC) [16] together with max-pooling on two types of local features: local color histograms and multi-resolution local patterns [17]. These features were extracted from patches of size 16×16 with an overlap of 12 pixels in the horizontal and vertical directions. Since the dimensionality of the local color histogram features are high (equal to 3 colors $\times$ 256 bins), we applied PCA to reduce its dimension to 400. Separate dictionaries of size 500 were learned for each feature type using k-means on a randomly sampled 200,000 features from the training set. Finally each frame was represented as a feature vector of size 1000, which is a concatenation of the frame representation obtained by each feature type.

Evaluation Criteria: The classification performance was evaluated based on leave-one-video-out experiments. Due to the highly unbalanced nature of the dataset, the average of the true positive rate (or sensitivity) and true negative rate (or specificity), namely the mean class accuracy (MCA), was used to evaluate the classification performance.

LibLinear [18] was used to train the SVM classifier. The regularization parameter of SVM is learned based on a three-fold cross validation applied on the training set. The `vlfeat` library [19] was used to create the dictionary and to extract the SIFT matches. The code from the authors of [16] was used for LLC.

3.2 Temporal Consistency for Classification

This section compares a single-layer SVM classifier, which is trained to classify frames independently, with the proposed hierarchical classifier which incorporates the temporal consistency.

Let SVM-TC and SVM-TF represent the second and the third level classifiers proposed in Sects. 2.2 and 2.3 respectively. Table 2 reports the MCA obtained by the single level (first row) and the proposed hierarchical (second and third rows) approaches for different videos.

Table 2. MCA per video with (2nd and 3rd rows) and without (1st row) the proposed hierarchical approach. SVM was used as the first-level classifier. The fourth row contains the percentage of improvement achieved by our approach (SVM-TF) with respect to the single-layer SVM.

Method	video 1	video 2	video 3	video 4	video 5
SVM	66.8	73.9	89.4	73.7	98.3
SVM-TC	**73.2**	84.7	90.5	74.6	99.1
SVM-TF	72.3	**84.9**	**91.5**	**75.9**	**99.4**
% improvement	**6.4**	**11.0**	**2.1**	**2.1**	**1.1**

As expected for all the videos adding temporal information considerably improve the MCA. The third level classifier gives modest improvements over the second level one, suggesting that the second level classifier already captures the temporal consistency information.

Figure 3 illustrates a qualitative comparison between the first level SVM and our approach. Note in Fig. 3(a–c) that the single-frame approach of SVM classifies erroneously few ambiguous frames, instead our approach, correctly classifies these frames by propagating the classification of SVM from more certain frames towards ambiguous ones. The example in Fig. 3(d) shows a challenging case when our approach obtains an incorrect classification, however this is mainly due to the classification obtained by the first level SVM classifier, which in this example is erroneous for the whole subsequence.

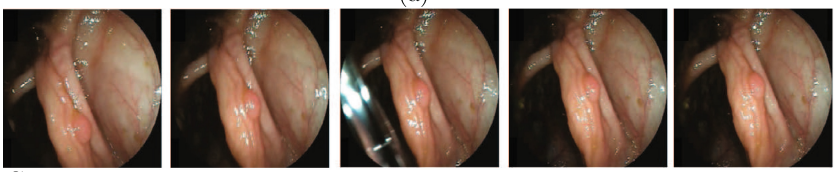
In this experiment the window sizes was empirically set to $n = 10$ for the second layer classifier and $m = 5$ for the third layer classifier respectively.

3.3 Generalization to Other Classifiers

The goal of this section is to show the applicability of our approach with respect to other first-level classifiers, i.e., by replacing the SVM classifier (used in Sect. 3.2) with a Random Forest (RF) classifier.

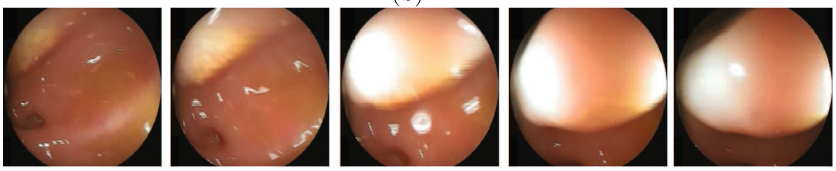
GT:	A	A	A	A	A
SVM:	A	A	N	A	A
Our:	A	A	A	A	A

(a)



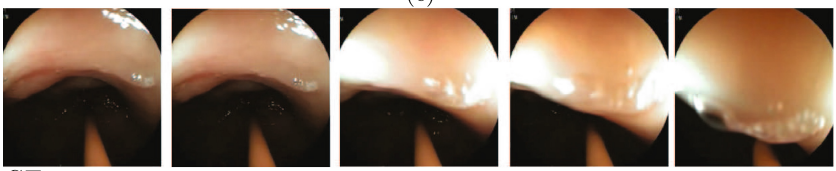
GT:	A	A	A	A	A
SVM:	A	A	N	A	A
Our:	A	A	A	A	A

(b)



GT:	N	N	N	N	N
SVM:	N	N	A	N	N
Our:	N	N	N	N	N

(c)



GT:	N	N	N	N	N
SVM:	A	A	A	A	A
Our:	A	A	A	A	A

(d)

N = Normal, A = Normal, SVM = SVM without temporal context, GT = Ground-Truth

Fig. 3. Qualitative example of the performance of the first level SVM and our proposed algorithm over four video subsequences (a–d) where our hierarchical classifier is able to correct the misclassified frames by enforcing temporal constraints.

Table 3 reports the MCA for RF with and without the temporal consistency. Adding the temporal consistency to the RF considerably improves the MCA for most of the videos. However, SVM without temporal information (Table 2) obtains better or very competitive results than RF without temporal information. When temporal consistency is added, SVM with temporal context performs better than RF with temporal context.

The number of trees in the RF classifier was set to 200 since we observed that increasing the number of trees leads to poor performance. This might happen because RF require very large training sets to perform optimally.

Table 3. MCA per video with (2nd and 3rd rows) and without (1st row) the proposed hierarchical approach. RF was used as the first-level classifier.

Method	video 1	video 2	video 3	video 4	video 5
RF	58.6	62.2	90.3	63.9	**100**
RF-TC	**59.6**	**68.0**	**91.6**	**67.2**	**100**
Improvement	**1.0**	**5.8**	**1.3**	**3.3**	0

4 Conclusions

We presented here a novel three-layer classifier to detect normal-abnormal frames in a colonoscopy video. Differently from other methods, our approach hierarchically combines the accuracy and robustness of SVM with the temporal consistency of two temporal classifiers. Experimental evaluation over five challenging colonoscopic videos shown improved classification accuracy, with two cases with significant improvements of 8.5 % and 14.9 %, when comparing against a SVM approach without any temporal information. Future work will be directed towards investigating other classification approaches as well as quantifying the impact of uninformative frames in the classification process.

References

1. Cancer research UK. http://info.cancerresearchuk.org/cancerstats
2. Bressler, B., Paszat, L.F., Chen, Z., Rothwell, D.M., Vinden, C., Rabeneck, L.: Rates of new or missed colorectal cancers after colonoscopy and their risk factors: a population-based analysis. Gastroenterology **132**(1), 96–102 (2007)
3. Lima, C., Barbosa, D., Ramos, A., Tavares, A., Montero, L., Carvalho, L.: Classification of endoscopic capsule images by using color wavelet features, higher order statistics and radial basis functions. In: IEEE EMBS (2008)
4. Manivannan, S., Wang, R., Trucco, E.: Extended gaussian-filtered local binary patterns for colonoscopy image classification. In: IEEE ICCV Workshops (2013)
5. Manivannan, S., Wang, R., Trucco, E., Hood, A.: Automatic normal-abnormal video frame classification for colonoscopy. In: IEEE ISBI (2013)

6. Engelhardt, S., Ameling, S., Paulus, D., Wirth, S.: Features for classification of polyps in colonoscopy. In: CEUR Workshop Proceedings (2010)
7. Karkanis, S.A., Iakovvidis, D.K., Maroulis, D.E., Karras, D.A., Tzivras, M.: Computer aided tumor detection in endoscopic video using color wavelet features. IEEE Trans. IT Biomed. **7**, 141–152 (2003)
8. Maroulis, D.E., Iakovidis, D.K., Karkanis, S.A., Karras, D.A.: Cold: a versatile detection system for colorectal lesions in endoscopy video-frames. Comput. Methods Programs Biomed. **70**, 151–166 (2003)
9. Cui, L., Hu, C., Zou, Y., Meng, M.Q.H.: Bleeding detection in wireless capsule endoscopy images by support vector classifier, IEEE International Conference on Information and Automation (2010)
10. Tjoa, M.P., Krishnan, S.: Feature extraction for the analysis of colon status from the endoscopic images. Biomed. Eng. Online **2**, 3–17 (2003)
11. Kumar, R., Zhao, Q., Seshamani, S., Mullin, G., Hanger, G., Dassopoulos, T.: Assessment of crohn's disease lesions in wireless capsule endoscopy images. Biomed. Eng. Online **59**, 355–362 (2012)
12. Liedlgruber, M., Uhl, A.: Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review. IEEE Rev. Biomed. Eng. **4**, 73–88 (2011)
13. Ben-Hur, A., Weston, J.: A user's guide to support vector machines. In: Data Mining Techniques for the Life Sciences. Methods in Molecular Biology, vol. 609, pp. 223–239. Humana Press (2010)
14. Lin, H.T., Lin, C.J., Weng, R.: A note on platt's probabilistic outputs for support vector machines. Mach. Learn. **68**(3), 267–276 (2007)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
16. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE CVPR (2010)
17. Manivannan, S., Li, W., Akbar, S., Wang, R., Zhang, J., McKenna, S.J.: HEp-2 cell classification using multi-resolution local patterns and ensemble SVMs. In: I3A 1st Workshop on Pattern Recognition Techniques for Indirect Immunoflurescence Images, ICPR (2014)
18. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
19. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/

3D Stable Spatio-Temporal Polyp Localization in Colonoscopy Videos

Debora Gil[1], F. Javier Sánchez[1], Gloria Fernández-Esparrach[2], and Jorge Bernal[1(✉)]

[1] Computer Vision Center and Computer Science Department, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
{debora,javier,jbernal}@cvc.uab.es
http://ivendis.cvc.uab.es/

[2] Endoscopy Unit, Gastroenterology Service, CIBERHED, IDIBAPS Hospital Clinic, Universidad de Barcelona, Barcelona, Spain
mgfernan@clinic.ub.es

Abstract. Computational intelligent systems could reduce polyp miss rate in colonoscopy for colon cancer diagnosis and, thus, increase the efficiency of the procedure. One of the main problems of existing polyp localization methods is a lack of spatio-temporal stability in their response. We propose to explore the response of a given polyp localization across temporal windows in order to select those image regions presenting the highest stable spatio-temporal response. Spatio-temporal stability is achieved by extracting 3D watershed regions on the temporal window. Stability in localization response is statistically determined by analysis of the variance of the output of the localization method inside each 3D region. We have explored the benefits of considering spatio-temporal stability in two different tasks: polyp localization and polyp detection. Experimental results indicate an average improvement of 21.5 % in polyp localization and 43.78 % in polyp detection.

Keywords: Colonoscopy · Polyp detection · Polyp localization · Region extraction · Watersheds

1 Introduction

1.1 Intelligent Systems for Colonoscopy

Colorectal cancer (CRC) is a serious health problem that affects the general population and is considered the fourth cause of cancer death worldwide with around 750.000 new cases diagnosed in 2012. Out of all found lesions, it is considered that at least two thirds of CRC develop through adenoma-carcinoma pathway [1]. Considering this, early screening with colonoscopy to search for CRC and its precursor lesion has become a generalized practice [2] and it is shown as crucial to patients' survival. Although colonoscopy has become the gold standard for colon screening, it still presents some drawbacks being polyp miss-rate-reported to be as high as 22 %- the most relevant affecting its effectiveness [3].

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 140–152, 2016.
DOI: 10.1007/978-3-319-29965-5_14

Several actions have been proposed to reduce polyp miss rate, such as optimal patient preparation and novel methodologies to carry out a complete examination of the mucosa. However, sometimes these new methodologies have impact in other quality metrics such as withdrawal time, as exposed in [4]. Regarding the technology itself, during the last years most of the developments in endoscopy have been focused on improving the quality of the images. This improvement in image quality has attracted the interest of computer scientists and has resulted in the creation of a new research field referred as intelligent systems for colonoscopy [5]. Among the different applications a given intelligent system can have, the one that has attracted higher research interest is the development of automatic polyp characterization methods.

Existing computational methods can be divided into those devoted to obtain an accurate localization of the polyp in the image -polyp localization- and those focus on providing as output an indicator of the presence or absence of polyps in the image -polyp detection-. The majority of these works rely on the extraction of shape, texture and color features to characterize polyps. The former includes methods which explore shape features of the different structures in the image to search for cues that discriminate polyps from other elements in the scene. Examples of methods belonging to this group can be found at [6–10]. Concerning texture-based approaches, we can find in the literature works that explore intensity patterns in the image to aid in polyp characterization, such as the works of [9,11]. Other approaches involve the use of state-of-the-art feature extraction methods such as local binary patterns [12] or MPEG-7 [13].

Although there is a great variety of methods, it is very difficult to compare them as they are commonly tested in private databases, hindering their actual performance in general cases of study that can appear in routinely procedures, therefore limiting their potential clinical deployment. In order to cope with this, efforts have been made to create and publish annotated databases of both still frames (CVC-ClinicDB database [7]) and videos (ASU-Mayo Clinic database [14]). Moreover, in order to gather researchers on the field, two different challenges on automatic polyp detection have been organized in 2015, at ISBI conference and as part of MICCAI Endoscopic Vision Challenge.

1.2 Motivation and Objectives of Research

After an analysis of the results of the different available methods, we have come to the conclusion that the majority of them present the following problem. Although they are able to locate/detect accurately the polyp in some frames, when this method is tested in a whole sequence performance scores decrease. We attribute this decrease in the performance to the lack of spatio-temporal stability in the response of the given methods, which can produce situations such as the one shown in Fig. 1. We can observe in this figure how a given polyp localization method (in this case, an implementation of Window-Median Depth of Valley Accumulation (WM-DOVA energy maps [7]) can provide a good localization output for isolated frames but, when analyzing its performance during a sequence

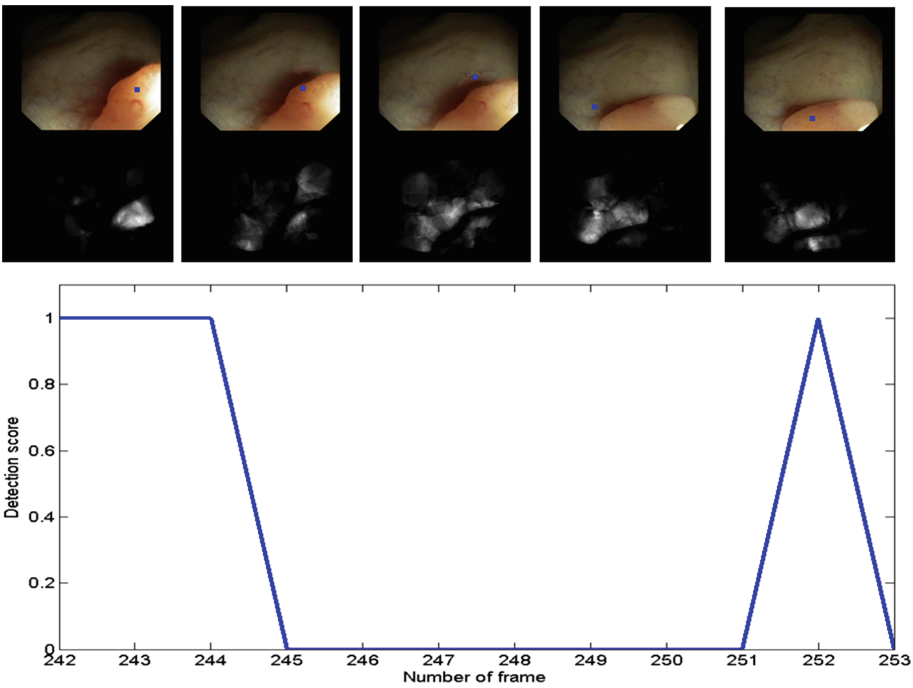


Fig. 1. Example of lack of spatio-temporal stability in the response of state-of-the-art polyp localization methods

of frames, we can observe how the polyp localization (in this case marked as a blue square) is not stable even for the two last images where the movement between them is small.

We propose in this paper a novel methodology to add spatial and temporal coherence to the response of a given polyp localization method. The use of temporal windows for increasing polyp detection capabilities has been explored in other works such as [15], although in this case the authors propose the use of Conditional Random Fields for adding spatio-temporal coherence to a texture-based polyp detection method. As WM-DOVA maps are the only ones tested on a public annotated frame-based database, we will take as base localization method these maps for this preliminary study, although our methodology could be used for any given energy-map based polyp localization method. Our methodology explores the consistency of the response by considering the displacement of the structures that appear in the image in a way such if a given polyp is localized in a region of the area with a high response, it is expected that a similar response will be given in a consecutive frame where the movement between frames is minimal. Moreover, we assess the potential of a given localization method as a polyp detection method by exploring if the response given by a polyp in a given frame loses stability when the polyp disappears from the scene. We validate

our methodology in terms of polyp localization, by comparing the performance of state-of-the-art method with and without applying spatio-temporal stability and, in terms of polyp detection, in a sequence with frames with presence and absence of the polyp.

The structure of the paper is as follows: we present our methodology in Sect. 2. Experimental results on both polyp localization and detection are discussed in Sect. 3 and we close this paper by exposing the main conclusions along with guidelines for future work in Sect. 4.

2 Methodology

The basis of our methodology is to improve the response of a given localization method in a given frame by incorporating information of neighboring frames in a temporal window centered at the frame (Fig. 2). Our method assumes that the response to the localization method keeps stable in such a window centered in a frame containing a polyp, in contrast to responses due to other structures (such as folds or specular highlights) which should be more spatially erratic. It is true that lumen region can also be considered as an stable structure that appears during consecutive frames and, in this case, we have used our methodology regarding non-informative region identification [16] to mitigate its impact in our approach. In order to explore such spatio-temporal response stability, we first need to obtain and track the different regions that appear in the given set of frames for a later classification of the regions in terms of polyp presence by performing a 3D statistical analysis of the output of the given localization methods for the extracted regions. By this, the output of the polyp localization in a given frame will depend on the output of polyp localization in a window of frames centered on it in a way such the output of a polyp localization method in a region of the image will rely on statistics over the output of the localization method for this specific region in a window of frames.

Before starting with the explanation of our 3D spatio-temporal stabilization of the output of polyp localization method, we will make a brief review of the localization method we will use as base, WM-DOVA energy maps.

Window Median Depth of Valley Accumulation (WM-DOVA) energy maps are based on a model of appearance for polyps which characterize polyp boundaries in terms of valley information [7]. This model is designed to foster those features characteristic of polyp boundaries (continuity, concavity, completeness and robustness to noisy structures) and it is specially designed to favor polyps from other structures on the image which also convey valley information such as folds, blood vessels or image artifacts such as specular highlights. The method is based on the accumulation on the output of a valley detector -in this case completed with information from morphological gradient to achieve a sense of the depth of the valleys- by using a ring of radial sectors. The final accumulation value for each pixel is calculated from the contribution of the different sectors centred on it but, in this case, the behaviour of a neighborhood of sectors is observed before calculating sectors' contribution to the final accumulation

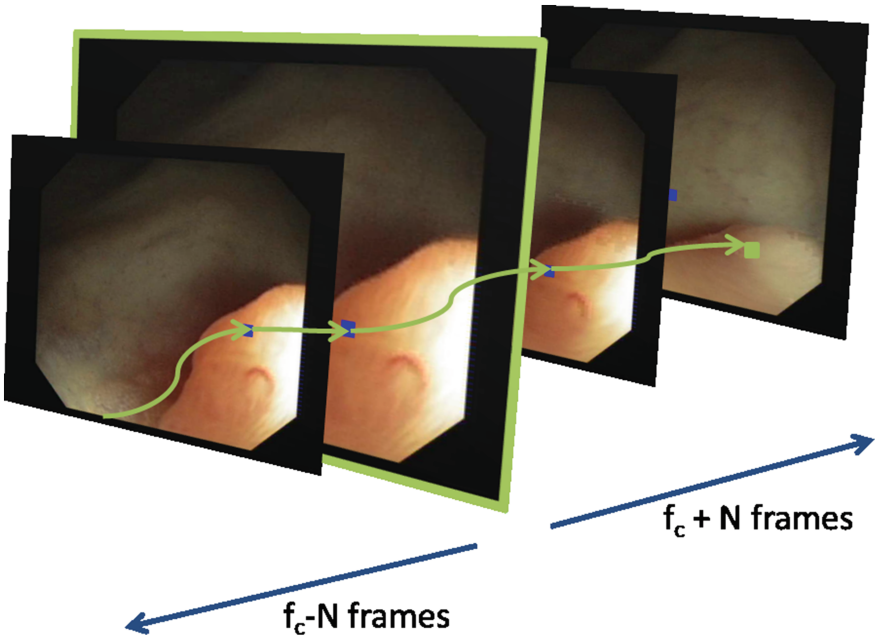


Fig. 2. Graphical scheme of the use of neighbor information to stabilize the output of a polyp localization method

value. More details on WM-DOVA energy maps creation can be found at [7]. WM-DOVA maps are proven to perform well for a wide range of images, appearing specially useful when having zenithal views of the polyps, regardless of their morphology and size.

2.1 Spatio-Temporal Region Extraction Using Watersheds

The first stage in our processing scheme aims at extracting a set of connected regions over a temporal window centered on each sequence frame. Each region represents an element which presence is kept in some consecutive frames from the temporal window. The question is to decide whether this element is a polyp or not and in order to take this decision we propose to perform a statistical study regarding characteristics of WM-DOVA maps during the temporal window where it appears. Considering this, region extraction should not be performed in a single-frame basis, but on a temporal window centered on the specific frame we are working with. In order to achieve this, we will perform watersheds in 3D over this window of greyscale images. In this context, the first two dimensions represent the image in 2D and the third dimension represents the time -understood as the temporal sequence of frames-.

3D watersheds extend the calculation of 2D watersheds to 3D volumes or sequences of frames and have already been applied in the context of medical

image segmentation [17,18]. The basic idea of watersheds consists of considering the given input image as a topographic surface. If we start to flood the regions starting by the regional minimums we will get to a point in which the water from one region invades a neighbor region. All the surface points at a given minimum constitute the catchment basin associated with that minimum. Watersheds are the zones which divide adjacent catchment basins. 3D extension aims to keep those regions which can be identified within this window of frames. In our particular case, we apply 3D watershed throughout all the frames belonging to the temporal window centered on the target frame. The methodology to calculate 3D watershed transformation for a given central frame f_c is:

1. Definition of a temporal window $w(f_c)$ of size r centered on f_c as $w(f_c) = \{f_i | i \in [f_c - r, f_c + r]\}$.
2. Calculation of the morphological gradient MG_i for each frame f_i contained in $w(f_c)$.
3. Calculation, for the central frame f_c, of the set of markers Mk_c as the local minima of MG_c.
4. Calculation of 3D watershed transform for all the frames in the temporal window, using the set of MG_i for the temporal window $w(f_c)$ as input image and Mk_c as markers .

Although we extend watershed calculation to the temporal window, the puncitual calculation for a given frame uses only as markers the local minima of the morphological gradient information calculated for the central frame. The catchment basin associated with those minima is extended over neighbor frames, following the movement of the image gradients. This behavior causes that each region should only represent one same element of the scene for a series of frames, which allows to perform a statistical study over the time. Unfortunately, this behavior is not common in the analysis of colonoscopy images where watershed

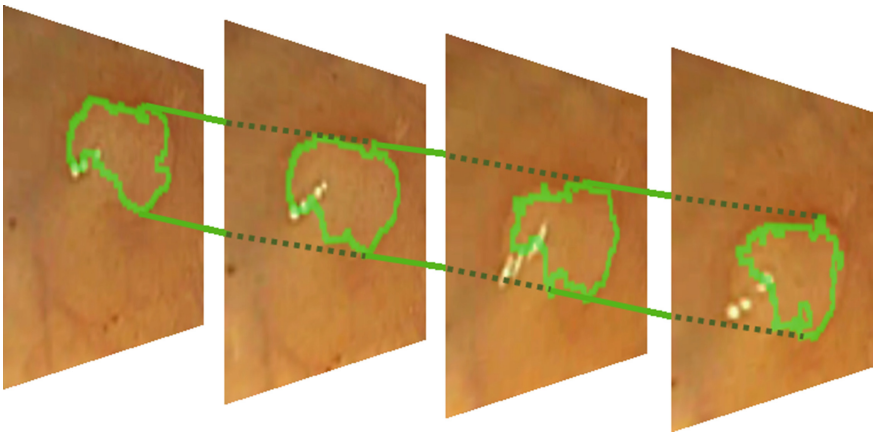


Fig. 3. Graphical explanation of the use of 3D watershed to track a polyp region.

fragments image elements in more than a region therefore reducing their statistical representativeness. Moreover, in some residual cases, a given region may cover more than one element of the scene.

For each frame, the output of this processing stage is a 3D representation of the stable regions inside the frame temporal window. By doing this analysis, we can easily observe which regions tend to be stable and which of them disappear, either because they are merged in a larger region or, following our hypothesis, due to the disappearance of the structure that originated them. Figure 3 illustrates the output of the 3D watershed in the case of a stable region corresponding to a polyp.

2.2 3D Region Statistical Analysis for Merging Polyp Region Information

Watershed stable regions provide an over-segmentation of the image in small regions that should be further selected and merged to provide a stable 3D localization of the polyp region. Under the assumption that polyp appearance keeps stable in temporal windows, those watershed regions inside a polyp should be significantly larger in frame size and DOVA values than regions outside polyps.

Small temporal regions are removed by a threshold, N_{Fr}, on the number of frames, N_{f_c}, contained in the watershed segmentation. To account for sudden scope motions, this threshold should be kept low. Regions with significant larger DOVA values are selected by an statistical analysis of the values obtained inside each watershed region. In order to detect significant differences we use Analysis of Variance (ANOVA) [19]. Given a grouping of a data set and a quantitative variable defined for each group, ANOVA is a statistical test that allows to decide if there are significant differences among the group's quantitative variable average with a given confidence α. The variability analysis is defined as soon as the ANOVA quantitative score and the different factors and methods are determined. In order to applied for polyp region selection, ANOVA groups and variable are defined as follows.

For each frame, f_c, the ANOVA groups are given by watershed labels of regions having more than N_{Fr} frames. For each such a region, the ANOVA variable is given by the median of DOVA values computed for each frame in the temporal window used to compute the 3D watershed. This gives a sampling of size N_{f_c}, being N_{f_c} the number of frames of the watershed region. ANOVA multicomparison is corrected using Tukey [20] to select those regions that have a median DOVA significantly higher.

Finally, the ANOVA selected regions are merged according to spatial connectivity to provide a single response per polyp.

3 Experimental Results

We validate our methodology by performing two separate experiments: the first one aims at assessing the impact of spatio-temporal stabilization of the response

of WM-DOVA maps in a sequence of frames, all of them containing a polyp. The second experiment is focused on exploring the potential of this stabilization method in polyp detection tasks when tested in a sequence with polyp and non-polyp frames.

In these experiments, we will note by DOVA the polyp localization given by WM-DOVA global maximum described in [7] and by DOVA3D, our spatio-temporal DOVA response.

3.1 Polyp Localization Results

In order to explore the benefits of DOVA3D, we have selected five different sequences with a polyp from those which compose CVC-ClinicDB [7]. We use as ground truth those frames from the original sequences that were included in the database. Our experiment consists of checking whether the performance of the localization method for these frames changes if we add spatio-temporal stabilization. To achieve this, we have been kindly granted with permission from the authors in order to analyze all the frames from the full sequences.

We define the following metrics for this experiment:

- Detection Rate (DR) defined as the ratio between the number of polyps in the sequence correctly located and the total number of polyps in the sequence:

$$DR = \frac{\#POk}{\#POk + \#PNOk}$$

  where POk represents a polyp correctly located and, conversely, $PNOk$, a polyp which was not located for a given image. In this case we label a polyp as correctly located whenever a polyp region is defined over the ground truth as an output of the statistical analysis.
- False Positive Rate (FPR) defined as the ratio between the total number of regions without polyp content (NPR) and the total number of final regions provided by our system, which also includes regions with polyp content, PR:

$$FPR = \frac{\#NPR}{\#PR + \#NPR}$$

We present DR and FPR results for both original WM-DOVA and spatio-temporal stable WM-DOVA for each sequence in Table 1. As we can observe from the Table, the spatio-temporal stabilization of WM-DOVA maps leads an general improvement of both DR and FPR for all the sequences. It is important to mention that there are some cases, such as sequence 4, where our methodology is able to improve DR in around 65 %, which indicates the potential of our approach to recover some mislocalizations by means of spatio-temporal coherence. Another important result is the reduction of FPR for all sequences. This is attributed to the statistical selection of the final regions that discard non-polyp information.

The benefits of our spatio-temporal analysis are assessed using a one-tailed t-test for paired data. For the DR score, we use a right-tailed test with

null hypothesis to $H_0 : \mu(DR_{DOVA3D} - DR_{DOVA}) < 0$, so that rejecting the test ($p_{val} < 0.05$) shows that DOVA3D has a significant larger detection rate. For the FPR score, we use a right-tailed test with null hypothesis to $H_0 : \mu(FPR_{DOVA3D} - FPR_{DOVA}) > 0$, so that rejecting the test ($p_{val} < 0.05$) shows that DOVA3D has a significant smaller false positive rate. We have also computed confidence intervals, CI, for the difference in means, $\mu(DR_{DOVA3D} - DR_{DOVA})$, $\mu(FPR_{DOVA3D} - FPR_{DOVA})$ to give the expected difference range. On one hand, the p-value for the DR test is $p_{val} = 3.9593e{-}005$, which clearly rejects the null hypothesis and, in fact, $CI = [11.3\,\%, 32.0\,\%]$, so that differences in average DR are at least 11 %. On the other hand, the p-value for the FPR test is $p_{val} = 0.0052$, which also rejects the null hypothesis and, in this case, $CI = [-23.9\,\%, -3.3\,\%]$, so that the reduction in average FPR is at least 3 %.

3.2 Polyp Detection Results

In order to illustrate the potential benefits of DOVA3D, we were provided by the authors of [7] with an additional sequence from an actual colonoscopy exploration. In this case we asked for a sequence in which the polyp is not present for all the frames, showing special interest in having a sequence in which the polyp is present, then it disappears for a set of frames and, finally, it appears again in the scene. We created a ground truth for all the frames in the sequence, which was validated by clinical personnel.

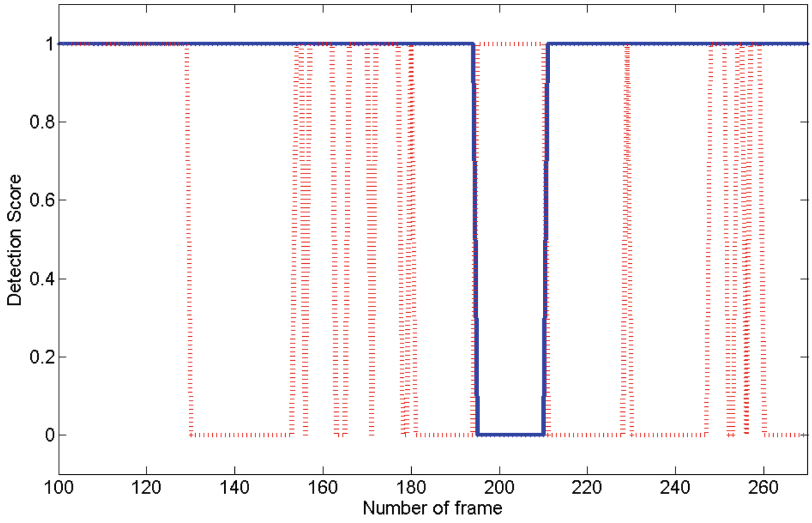
Regarding this second experiment, we propose to use FPR and a new metric, Detection Score:

$$DS = \frac{\#DOk}{\#DOk + \#DNOk}$$

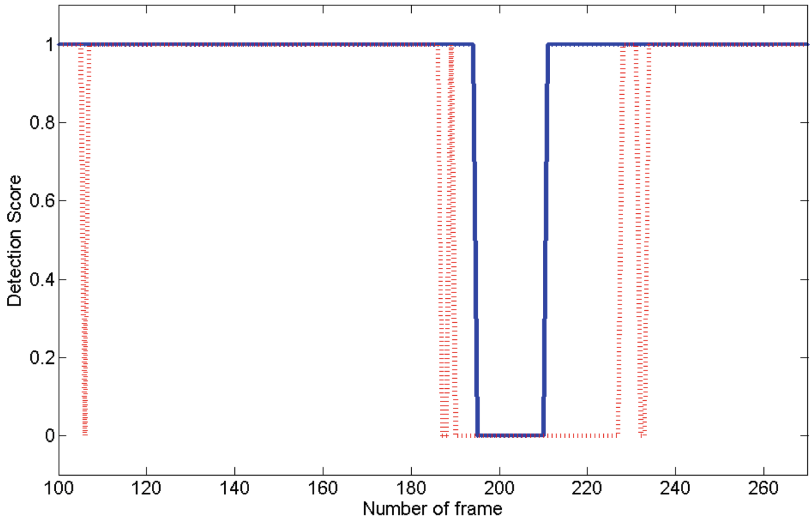
In this case we define a good detection DOk as the one whenever our method provides with an actual polyp location in a frame with a polyp or does not provide any kind of output for a frame without a polyp. Conversely, we define a bad detection $DNOk$ as our method providing a polyp location in a frame without polyp or not providing a polyp location in a frame with polyp.

Table 1. Comparison of DR and FPR results between original WM-DOVA and spatio-temporal stable WM-DOVA.

	Original WM-DOVA		Spatio-temporal Stable WM-DOVA	
Sequence	DR [%]	FPR [%]	DR [%]	FPR [%]
1	84.62	15.38	**91.67**	23.08
2	81.82	18.18	**90.91**	**15.91**
3	50.00	50.00	**66.67**	53.19
4	14.89	72.73	**80.49**	**58.05**
5	84.00	16.00	76.19	**10.29**



(a)



(b)

Fig. 4. Comparison of Detection Score between WM-DOVA maps (a) without and (b) with spatio-temporal stabilization. Blue line in the plots represents the presence (value 1) or absence or polyp in the image (value 0). Red line represents the performance of the method: good localization (value 1) or erroneous localization (value 0).

We present a graphical comparison of the performance of WM-DOVA maps with and without spatio-temporal stabilization in Fig. 4. By observing the plots for the two methods in the comparison, we can observe how spatio-temporal stabilization helps to improve polyp localization results in those frames with a polyp, as for the case of stable spatio-temporal WM-DOVA there is a general coincidence between the output of the method and the ground truth in both presence and absence of the polyp; this can be observed by having coincidence of blue (ground truth) and red (output of the method) lines for the majority of the frames. As can be seen, we can also observe how, in absence of a polyp -frames 190 to 210-, our methodology is able to correct the erroneous localization provided by WM-DOVA maps, which offer a candidate location for every frame analyzed. Overall, DS score improves from 38.71 % to a 82.58 %, which shows the potential of our approach to obtain good localization results using spatio-temporal coherence of WM-DOVA maps. Aside, we can also observe how the number of false alarms is also reduced, decreasing from 64.77 % to 19.87 %, indicating the potential of our approach to reduce the impact of noisy structures in overall localization results.

To close this section, we present a qualitative example of the benefits of adding spatio-temporal stability to the output of WM-DOVA in Fig. 5. We can observe that, for a same input image, original localization by means of global maximum of WM-DOVA provided a mislocalization outside the polyp (marked as a red square in the first image) whereas the stabilized response over a temporal window centered in this particular frame allows us to correctly localize the polyp (marked as a green square in the third image).

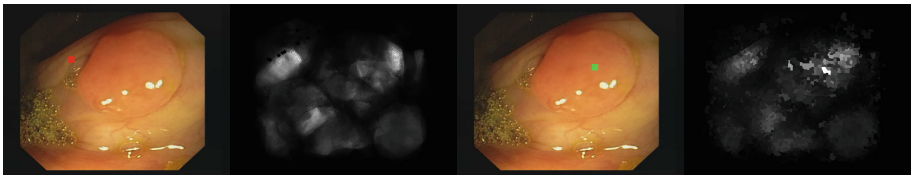


Fig. 5. Qualitative example of the benefits of adding spatio-temporal stabilization the output of WM-DOVA maps over a temporal window: (first image) original image; (second image) corresponding WM-DOVA map; (third image) original image; (fourth image) stabilized WM-DOVA map over 3D watershed regions. Correct localizations are marked as green squares over the original image, false positives as red squares.

4 Conclusions and Future Work

This paper addresses one of the main drawbacks of frame-based polyp localization algorithms, which is related to the lack of spatio-temporal stability in their output when applied to a sequence of frames. In order to cope with this we propose to incorporate information of a neighborhood of frames. Our methodology

is based on the observation of the response of the output of the method over a same region along a temporal window of frames. In order to extract stable regions over time we propose the use of 3D watersheds and then, in order to integrate the output of the localization method over time, we perform an statistical analysis over the output of the method along all the frames in which the region is present. Experimental results on polyp localization indicate the benefits of adding spatio-temporal stability which is observed by both an increase in Detection Rate (average improvement of 21.50 %) and by a strong decrease in the number of false alarms provided by the method (with an average decrease of 13.30 % in FPR). Moreover, we have also studied the potential of our methodology in polyp detection tasks: a preliminar study over a full annotated sequence with frames with both presence and absence of polyp shows an improvement over 43 % regarding detection score metric.

These preliminary results shows the potential of our methodology but also allows us to sketch future research lines. Although our methodology improves the 3D performance of the localization method, it is clear that an estimation of the movement between frames -using motion descriptors such as optical flow or particle filtering- could also add value to the system as we could complement the output of 3D watershed with this information in order to obtain a more accurate tracking of the regions over the defined temporal window of frames. Additionally, studies about setting the size of the temporal window should be undertaken, which could include definition of automatic systems to assess when the temporal window information should be restarted due to the apparition of a high number of consecutive frames with low quality (blurring, fecal content). Regarding region extraction, region merging strategies may be developed to reduce the number of regions to be tracked, which could ease to reduce the computational cost of the whole methodology, easing the statistical analysis. Finally, this preliminary study should be extended over more sequences in order to account the performance of the whole approach in a wide range of scenarios.

Acknowledgments. Work supported by the Spanish project TIN2012-33116, DPI2015-65286-R, and the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya 2014-SGR-1470. Debora Gil is a Serra Hunter fellow.

References

1. Kerr, J., Day, P., Broadstock, M., Weir, R., Bidwell, S.: Systematic review of the effectiveness of population screening for colorectal cancer (2007)
2. Quintero, E., Castells, A., Bujanda, L., Cubiella, J., Salas, D., Lanas, Á., Andreu, M., Carballo, F., Morillas, J.D., Hernández, C., et al.: Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. N. Engl. J. Med. **366**(8), 697–706 (2012)
3. van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. **101**(2), 343–350 (2006)

4. Barclay, R.L., Vicari, J.J., Doughty, A.S., Johanson, J.F., Greenlaw, R.L.: Colonoscopic withdrawal times and adenoma detection during screening colonoscopy. N. Engl. J. Med. **355**(24), 2533–2541 (2006)
5. Bernal, J., Vilariño, F., Sánchez, F.J.: Towards intelligent systems for colonoscopy. In: Miskovitz, P. (ed.) Colonoscopy, pp. 245–270. INTECH (2011). Doi:10.5772/19748. http://www.intechopen.com/books/colonoscopy/towards-intelligent-systems-for-colonoscopy. ISBN: 978-953-307-568-6
6. Iwahori, Y., Shinohara, T., Hattori, A., Woodham, R.J., Fukui, S., Bhuyan, M., Kasugai, K.: Automatic polyp detection in endoscope images using a hessian filter. In: Proceedings of MVA, pp. 21–24 (2013)
7. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015)
8. Hwang, S., Oh, J., Tavanapong, W., Wong, J., De Groen, P.C.: Polyp detection in colonoscopy video using elliptical shape feature. In: IEEE International Conference on Image Processing, ICIP 2007, vol. 2, pp. II-465. IEEE (2007)
9. Tajbakhsh, N., Chi, C., Gurudu, S.R., Liang, J.: Automatic polyp detection from learned boundaries. In: Proceedings of ISBI , pp. 97–100. IEEE (2014)
10. Wang, Y., Tavanapong, W., Wong, J., Oh, J., de Groen, P.: Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy. IEEE J. Biomed. Health Inform. **18**, 1379–1389 (2014)
11. Ameling, S., Wirth, S., Paulus, D., Lacey, G., Vilarino, F.: Texture-based polyp detection in colonoscopy. In: Meinzer, H.-P., Deserno, T.M., Handels, H., Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin 2009. Informatik aktuell, pp. 346–350. Springer, Heidelberg (2009)
12. Iakovidis, D.K., Maroulis, D.E., Karkanis, S.A., Brokos, A.: A comparative study of texture features for the discrimination of gastric polyps in endoscopic video. In: Proceedings of IEEE CBMS, pp. 575–580. IEEE (2005)
13. Coimbra, M.T., Cunha, J.P.S.: MPEG-7 visual descriptors-contributions forautomated feature extraction in capsule. IEEE Trans. Circuits Syst. Video Technol. **16**(5), 628–637 (2006)
14. Tajbakhsh, N., Liang, J., Gurudu, S.R.: Asu-mayo polyp detection database (2015)
15. Park, S.Y., Sargent, D., Spofford, I., Vosburgh, K.G., et al.: A colon video analysis framework for polyp detection. IEEE Trans. Biomed. Eng. **59**(5), 1408–1418 (2012)
16. Bernal, J., Gil, D., Sánchez, C., Sánchez, F.J.: Discarding non informative regions for efficient colonoscopy image analysis. In: Luo, X., Reich, T., Mirota, D., Soper, T. (eds.) CARE 2014. LNCS, vol. 8899, pp. 1–10. Springer, Heidelberg (2014)
17. Lin, G., Adiga, U., Olson, K., Guzowski, J.F., Barnes, C.A., Roysam, B.: A hybrid 3d watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. Cytometry Part A **56**(1), 23–36 (2003)
18. Kuhnigk, J.-M., Hahn, H., Hindennach, M., Dicken, V., Krass, S., Peitgen, H.-O.: Lung lobe segmentation by anatomy-guided 3d watershed transform. In: Medical Imaging, pp. 1482–1490. International Society for Optics and Photonics (2003)
19. Cohen, J.: Statistical Powe Analysis For the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale (1988)
20. Tukey, J.W.: Comparing individual means in the analysis of variance. Biometrics **5**, 99–114 (1949)

Uninformative Frame Detection in Colonoscopy Through Motion, Edge and Color Features

Mohammad Ali Armin[1,2(✉)], Girija Chetty[1], Fripp Jurgen[2], Hans De Visser[2], Cedric Dumas[2], Amir Fazlollahi[2], Florian Grimpen[3], and Olivier Salvado[2]

[1] Department of Computer Science, University of Canberra, Canberra, Australia
m.a.armin@gmail.com
[2] CSIRO Biomedical Informatics, The Australian e-Health Research Centre Herston, Brisbane, Australia
Olivier.Salvado@csiro.au
[3] Department of Gastroenterology and Hepatology, Royal Brisbane and Women's Hospital, Herston, Australia

Abstract. Colonoscopy is performed by using a long endoscope inserted in the colon of patients to inspect the internal mucosa. During the intervention, clinicians observe the colon under bright light to diagnose pathology and guide intervention. We are developing a computer aided system to facilitate navigation and diagnosis. One essential step is to estimate the camera pose relative to the colon from video frames. However, within every colonoscopy video is a large number of frames that provide no structural information (e.g. blurry or out of focus frames or those close to the colon wall). This hampers our camera pose estimation algorithm. To distinguish uninformative frames from informative ones, we investigated several features computed from each frame: corner and edge features matched with the previous frame, the percentage of edge pixels, and the mean and standard deviation of intensity in hue-saturation-value color space. A Random Forest classifier was used for classification. The method was validated on four colonoscopy videos that were manually classified. The resulting classification had a sensitivity of 75 % and specificity of 97 % for detecting uninformative frames. The proposed features not only compared favorably to existing techniques for detecting uninformative frames, but they also can be utilized for the camera navigation purpose.

Keywords: Optical colonoscopy · Uninformative frames · Colonoscopy quality · Feature · Random Forest

1 Introduction

Colorectal cancer is the second leading cause of cancer related death after lung cancer in Australia, however detection and removal of polyps in early stages can increase the chance of survival by up to 90 % [1]. Optical colonoscopy is the gold standard in

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, pp. 153–162, 2016.
DOI: 10.1007/978-3-319-29965-5_15

inspecting and removing polyps. Each year 500,000 colonoscopies are performed in Australia [1]. One of the main issues for clinicians is to estimate the position of the endoscope inside the colon, and software solutions to help with navigation would be desirable [2–5]. As a whole, we aim to provide a technology to estimate camera pose during colonoscopy. However, colonoscopy video streams contain many frames with no or little clinical information such as the result of colon cleansing, a dirty lens, or close inspection of colon wall. In Fig. 1 some examples of colonoscopy frames are illustrated. We categorized colonoscopy frames as informative or uninformative. The informative frames include clear shot of the lumen Fig. 1(a–b) or wall (Fig. 1(c–d)). Uninformative frames are a result of blurriness (blurred), colon cleansing with water jet (water), lens contact to the colon wall with a various illumination (indistinct) or indistinct with big bubbles or a bubbles' colony that reduce clinical information in a frame (Fig. 1(e–h)). The uninformative frames decrease the quality of colon inspection by clinicians and may hamper our camera motion estimation algorithm. Some studies showed that uninformative frames can compromise up to 30–40 % of the entire video stream [6, 7]. Therefore, it is important to detect uninformative frames and remove them.

In recent years, several studies have reported automated identification of uninformative frames from endoscopy videos [6, 8, 9]. Oh et al. [6, 10] developed a method which is based on analyzing the gray level co-occurrence matrix (GLCM) texture of the discrete Fourier transform images and edge detection. Following that, Arnold et al. [8] proposed a Bayesian classification method to analyze the norm of the detail coefficients of wavelet decomposition to classify colonoscopy frames. They reported 92.3 % accuracy only in detecting indistinct frames similar to Fig. 1(g). Color features have also been used in Wireless Capsule Endoscopy (WCE) to identify useful frames prior to diagnosis [9, 11].

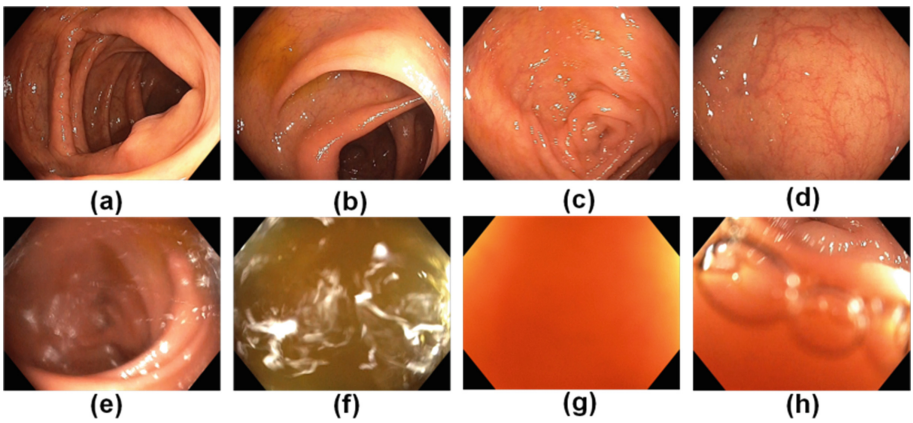


Fig. 1. First row is an example of informative frames, lumen view (a and b), wall view (c–d), and second row represent uninformative frames: blurred (e), water (f), indistinct (g), indistinct with bubble (h).

2 Method

The outline of the proposed method to classify colonoscopy frames is shown in Fig. 2. In this study, we investigate several features and use a Random Forest (RF) [12] classifier to detect uninformative frames. As the first step, all image frames were converted to the Hue-Saturation-Value (HSV) color space and smoothed using a Gaussian filter (By applying a Gaussian filter we aim at removing the noise, as well as moving mildly blurred frames to the blurred category). Subsequently, three shape-feature descriptors were investigated based on the following assumptions: (i) Consecutive uninformative frames results in a lower number of features detected by motion flow. For this, we computed the number of features detected by the Kanade Lucas Tomasi (KLT) tracker [13]. (ii) Uninformative frames such as Fig. 1(f–h) appear with a uniform color distribution. Therefore, to further emphasize on the color aspect, HSV color space was considered for computing the mean and standard deviation (STD) as features. (iii) Those uninformative frames which are blurred or mildly blurred have fewer sharp edges than a typical good quality colonoscopy image; for this we computed the percentage of edge pixels. The motivation of using these features is to utilize features currently computed for camera motion estimation to classify colonoscopy frames. This can also reduce the complexity of uninformative frame detection.

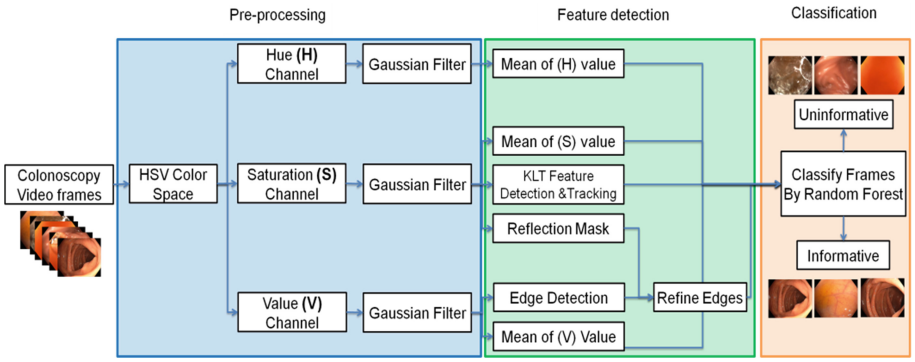


Fig. 2. The diagram of the proposed method for classification of colonoscopy frames.

2.1 Dataset

The data used for preparing this study were collected from four colonoscopy videos of different parts of the colon from different patients. Videos were captured by a 190HD Olympus colonoscope, with 50 frame/sec with a frame size of 1856 × 1044 pixels. A medical expert manually marked videos for uninformative frames. The details of our experimental videos are shown in Table 1.

Table 1. Dataset used in our experiment to detect uninformative frames

Dataset	Uninformative frames	Informative frames	Total frames	Informative and uninformative sequences
Patient 1	1205	1295	2500	2 × 30
Patient 2	112	1888	2000	2 × 14
Patient 3	201	1498	1699	2 × 11
Patient 4	702	2368	3070	2 × 40
Total	2220	7049	9269	190

2.2 Feature Detection

Number-of-features Descriptor Computed by KLT from Saturation Channel. The saturation color channel of HSV was used to extract and track features by the KLT method. This channel was used because our camera estimation parameters empirically obtained a better performance in feature detection. The KLT method detects corner like features with high contrast by measuring the minimum eigenvalue of each 2 × 2 gradient matrix in a frame. The displacement of selected features between consecutive frames was estimated by using an optimizer to minimize the difference between two feature windows for the image intensity. To address the large displacements, a pyramid based approach was used to track features.

Based on our assumption, frames with low numbers of features should have inadequate information to be used for camera motion estimation, and should be classified as uninformative frame. The number-of-features detected on a set of informative and uninformative frames are shown in Fig. 3.

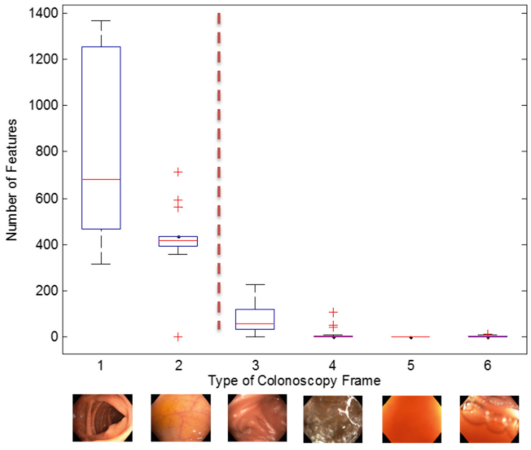


Fig. 3. Number-of-feature detected by the Kanade Lucas Tomasi (KLT) for several informative (1–2) and uninformative (3–6) frames using the Saturation color space.

Color Features from H-, S- and V-channel. Colonoscopy images are commonly presented in RGB color space. However, the HSV color space has shown a better ability in dichotomizing chromaticity (hue and saturation) from luminance [11]. Frames with no information such as the ones captured from a close inspection of the colon wall (Fig. 1(g–h)) or during colon cleansing have distinct signal from informative frames. Such distinction can be estimated by computing the STD and mean of the three HSV channels. The STD of hue, saturation and value for a set of informative and uninformative frames are presented in Fig. 4.

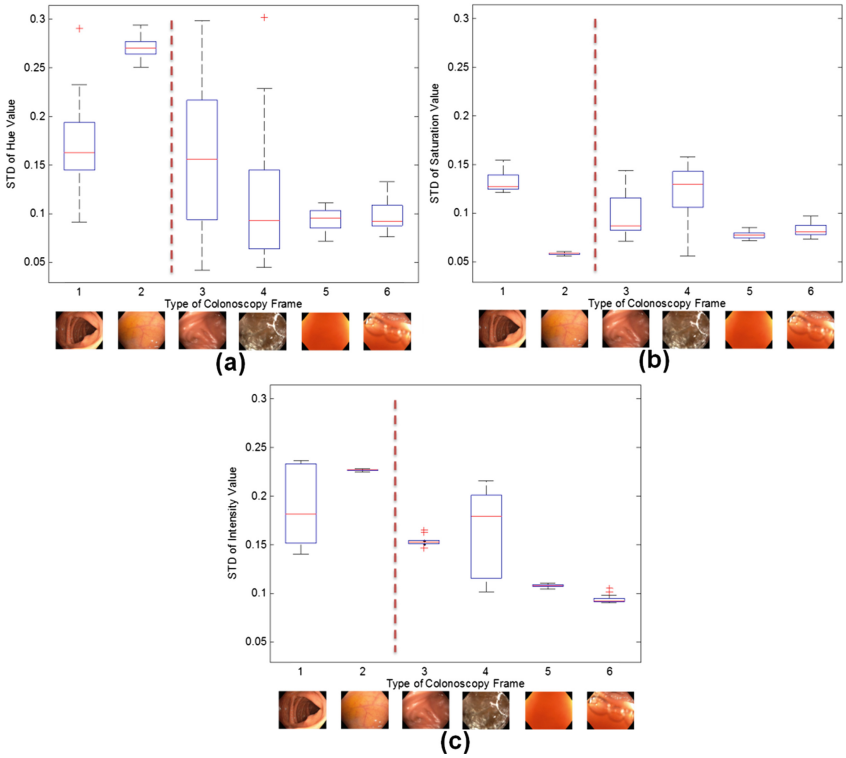


Fig. 4. The STD of Hue (a), Saturation (b), and Value (b) for several informative (1–2) and uninformative (3–6) frames.

Percentage-of-edge-pixel Feature Estimated from Value Channel. To detect uninformative frames, we analyzed the percentage of the edge pixels as the number of the edge pixels to all pixels in a frame. The edges were detected by using the Canny edge detector [14] from the Value channel. The percentage of isolated pixels introduced by Oh et al. [6] was also estimated for comparison.

Based on our experiments, frames with a higher percentage of edge pixels were informative whereas uninformative frames (including blurred, mild blurred and indistinct) had a lower percentage. Reflections can increase the number of edges,

especially when there are bubbles or a water jet for cleaning the colon. The reflection effect was removed by generating a mask using an automatic Otsu thresholding from a frame in the Saturation channel. The percentage of edge pixels on a set of informative and uninformative frames is shown in Fig. 5.

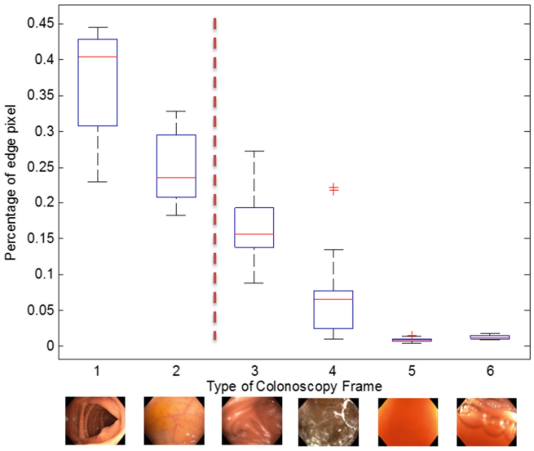


Fig. 5. Percentages-of-edge-pixel to all pixels for different types of colonoscopy frames (1and 2 represent informative and 3 to 6 represent uninformative frames) using the Value channel and Canny edge detector.

2.3 Random Forest Classification

To classify frames into informative and uninformative classes, a binary Random Forest (RF) classifier [12] was used. On all available frames, feature metrics, including number of motion features, mean and STD of each color channel, and percentage-of-edge-pixel were calculated. In a colonoscopy video, consecutive frames may provide similar information which reduces the efficiency of the RF classifier if selected together. Therefore, prior to classification, all the informative and uninformative sequences were divided into half. We used the first half for training and second half for testing. The parameters used for RF training were: 100 trees, sample selection without replacement, and a node size of maximum 2.

2.4 Experimental Evaluation

To evaluate the performance of the proposed detection technique, sensitivity, precision, specificity and accuracy were considered. To compare the effectiveness of the proposed feature descriptors with similar studies, the gray level co-occurrence matrix (GLCM) and percentage-of-isolated-pixel (IPR) [6] were also included.

3 Results

Two representative examples of the KLT, edge and color features computed on informative and uninformative frames are illustrated in Figs. 6 and 7. A high number of motion vectors and edge pixels were identified for informative frames which demonstrate the potentials of the proposed features.

The performance of the above mentioned features in detecting uninformative frames using RF classifier is shown in Table 2. The collective performance of the proposed features, with accuracy of 94 % and specificity of 97 %, compares favorably to GLCM + IPR features, with accuracy of 92 % and specificity of 96 %.The calculation

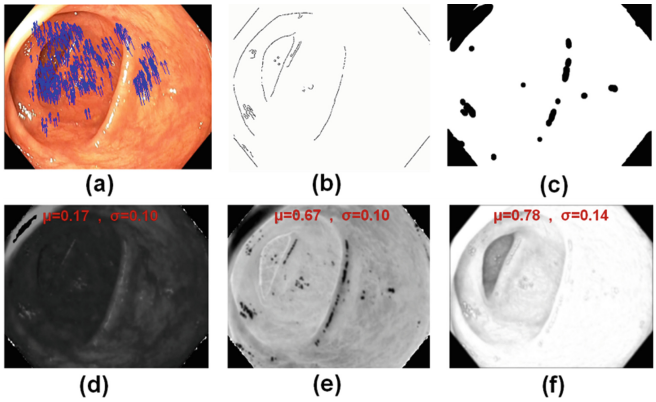


Fig. 6. The proposed motion (a) and edge features (b) computed on a representative informative frame along with the reflection mask (c) and three HSV channels (d–f). The mean (μ) and STD (σ) features are also shown on each color space.

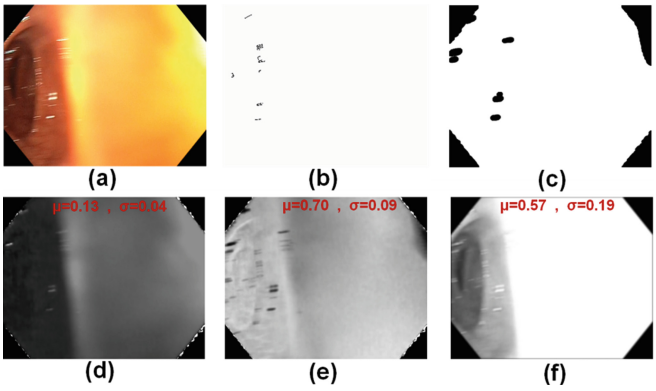


Fig. 7. The proposed motion (a) and edge features (b) computed on a representative uninformative frame along with the reflection mask (c) and three HSV channels (d–f). The mean (μ) and STD (σ) features are also shown on each color space.

Table 2. The performance of different feature descriptors on identifying uninformative frames

Feature	Classification results			
	Precision	Sensitivity	Specificity	Accuracy
KLT	0.83	0.72	0.97	0.93
STD color	0.62	0.36	0.96	0.85
Mean color	0.76	0.55	0.96	0.90
Percentage of edge pixels	0.75	0.51	0.97	0.89
All proposed features	**0.86**	**0.75**	**0.97**	**0.94**
GLCM	0.76	0.64	0.96	0.91
IPR	0.49	0.32	0.90	0.75
GLCM + IPR	**0.79**	**0.67**	**0.96**	**0.92**

time on average for KLT features computation was 0.16 s/frame whereas 0.072 s/frame was spent for GLCM feature calculation by using a standard PC, MATLAB, and non-optimized scripts.

4 Discussion

This paper proposes a method based on the KLT motion, color and edge features for detecting uninformative frames as the initial stage of our main pipeline for camera motion estimation algorithm. The proposed features were evaluated using a binary Random Forest classifier and obtained 86 % precision, 75 % sensitivity, 97 % specificity, and 94 % accuracy.

In the present work, the KLT motion features were proposed as a metric for identifying uninformative frames. To increase the number of these features in each frame, the HSV color space was found more suitable. More importantly, motion features were already available for estimating camera motion in our algorithm which will reduce the computational complexity. Besides, there are some frames, e.g. wall view, with fewer textures compared to lumen which will result in less motion features. To identify these frames, color information in HSV color space was calculated.

Adding more features such as GLCM, IPR or wavelet as used in literature might improve our method in classifying more complicated colonoscopy frames. For instance, these features might be useful when color features show a partial overlapping between a subset of uninformative and informative frames. However, the aim of this study was to investigate the feasibility of using KLT features which were concurrently computed for camera pose estimation. Furthermore, this approach can be used in endoscopy videos such as bronchoscopy and wireless capsule endoscopy to remove uninformative frames during camera motion estimation.

The main limitation of the current study is the small dataset size, increasing the number of frames might slightly change the reported performance. In future work, we aim to validate our method on a bigger dataset acquired from different colonoscopes with different field of view and resolution. A diverse colonoscopy video datasets from different patients will allow us to validate the proposed features with other training

approaches for RF classifier such as one-video-leave-out approach and clustering based methods such as K-mean clustering. Furthermore, other feature descriptors such as scale invariant feature transform (SIFT) or speeded up robust features (SURF) will be investigated.

5 Conclusion

This study demonstrated that KLT motion, color and edge features can together provide effective detection of uninformative colonoscopy frames. The proposed method can be performed simultaneously with camera pose estimation. This would reduce the computational burden and necessity to compute other complex features for uninformative frame detection.

References

1. Australian Institute of Health and Welfare. http://www.aihw.gov.au/
2. Liu, J., Subramanian, K.R., Yoo, T.S.: A robust method to track colonoscopy videos with non-informative images. Int. J. Comput. Assist. Radiol. Surg. **8**, 575–592 (2013)
3. Puerto-Souza, G.A., Staranowicz, A.N., Bell, C.S., Valdastri, P., Mariottini, G.-L.: A comparative study of ego-motion estimation algorithms for teleoperated robotic endoscopes. In: Luo, X., Reich, T., Mirota, D., Soper, T. (eds.) CARE 2014. LNCS, vol. 8899, pp. 64–76. Springer, Heidelberg (2014)
4. Mori, K., Deguchi, D., Sugiyama, J., Suenaga, Y., Toriwaki, J., Maurer, C.R., Takabatake, H., Natori, H.: Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. A preliminary version of this paper was presented at the Medical Image Computing and Computer-Assisted Intervention (MICCAI) Conference, Utrecht, The Netherlands (Mori et al. 2001). Med. Image Anal. **6**, 321–336 (2002)
5. Rai, L., Helferty, J.P., Higgins, W.E.: Combined video tracking and image-video registration for continuous bronchoscopic guidance. Int. J. Comput. Assist. Radiol. Surg. **3**, 315–329 (2008)
6. Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., de Groen, P.C.: Informative frame classification for endoscopy video. Med. Image Anal. **11**, 110–127 (2007)
7. Oh, J., Hwang, S., Cao, Y., Tavanapong, W., Liu, D., Wong, J., de Groen, P.C.: Measuring objective quality of colonoscopy. IEEE Trans. Biomed. Eng. **56**, 2190–2196 (2009)
8. Arnold, M., Ghosh, A., Lacey, G., Patchett, S., Mulcahy, H.: Indistinct frame detection in colonoscopy videos. In: 13th International Machine Vision and Image Processing Conference (IMVIP), pp. 47–52 (2009)
9. Mackiewicz, M., Berens, J., Fisher, M.: Wireless capsule endoscopy color video segmentation. IEEE Trans. Med. Imaging **27**, 1769–1781 (2008)
10. Oh, J., Hwang, S., Tavanapong, W., de Groen, P.C., Wong, J.: Blurry-frame detection and shot segmentation in colonoscopy videos. In: Proceedings of SPIE, pp. 531–542 (2003)
11. Bashar, M.K., Mori, K., Suenaga, Y., Kitasaka, T., Mekada, Y.: Detecting informative frames from wireless capsule endoscopic video using color and texture features. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 603–610. Springer, Heidelberg (2008)

12. Random Forest (Regression, Classification and Clustering) implementation for MATLAB. https://code.google.com/p/randomforest-matlab/
13. Shi, J., Carlo, T.: Good features to track. In: CVPR, pp.593–600 (1994)
14. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-8**, 679–698 (1986)

Erratum to: Computer-Assisted and Robotic Endoscopy

Xiongbiao Luo[1(✉)], Tobias Reichl[2], Austin Reiter[3], and Gian-Luca Mariottini[4]

[1] Xiamen University, Fujian, China
xbluo@xmu.edu.cn
[2] KUKA Robotics, Augsburg, Germany
[3] Johns Hopkins University, Baltimore, MD, USA
[4] University of Texas at Arlington, Arlington, TX, USA

Erratum to:
X. Luo et al. (Eds.)
Computer-Assisted and Robotic Endoscopy
DOI: 10.1007/978-3-319-29965-5

The original version of the cover was revised. The logo was inserted by mistake.

The updated original online version for this Book can be found at 10.1007/978-3-319-29965-5

X. Luo et al. (Eds.): CARE 2015, LNCS 9515, p. E1, 2016.
DOI: 10.1007/978-3-319-29965-5_16

Author Index