3D Cylinder Pose Estimation by Maximization of Binary Masks Similarity: A simulation Study for Multispectral Endoscopy Image Registration

O. Zenteno, S. Treuillet and Y. Lucas

PRISME, Univ. d'Orléans, F-45072 Orléans, France

Keywords: Multispectral Endoscopy, Image Registration, Optical Biopsy.

Abstract: In this paper we address the problem of simultaneous pose estimation for multi-modal registration of images captured using a fiberscope (multispectral) inserted through the instrument channel of a commercial endoscope (RGB). We developed a virtual frame using the homography-derived extrinsics parameters using a chessboard pattern to estimate the initial pose of both cameras and simulate two types of fiberscope movements (i.e, insertion and precession). The fiberscope pose is calculated by the maximization of similarity measures between the 2D projection of the simulated fiberscope and the fiberscope tip segmentation from the endoscopic images. We used the virtual frame to generate sets of synthetic fiberscope data at two different poses and compared them after the maximization of similarity. The performance was assessed by measuring the reprojection error of the control points for each pair of images and the pose absolute error in a sequential movement mimicking scenario. The mean reprojection error was 0.38 ± 0.5 pixels and absolute error in the tracking scenario was 0.05 ± 0.07 mm.

1 INTRODUCTION

Gastrointestinal complications are usually produced by a bacterial pathogen called Helicobacter pylori (Hp). About 50% of the world's population is infected with Hp but most individuals remain asymptomatic until developing clinical disease. The primary clinical manifestations of the infection are chronic inflammation which produce cellular alterations of the gastric mucosa (degeneration and infiltration) that can lead to peptic ulcers and malignous complication.

Current early detection capabilities are primarily based on gastro-endoscopic exploration under sedation or anesthesia during which the clinician may perform a biopsy for further histopathological examination if needed. Some endoscopic systems propose alternative spectral tools for helping gastric screening by the use of optical biopsies. Typical examples are Fuji Intelligent Chromo Endoscopy (FICE), proposed by Fuji and Narrow Band Imaging (NBI), proposed by Olympus (Song et al., 2008). These techniques have shown the benefits of using multiple wavelengths to improve the visibility of blood vessels and other important features. However, they are limited in the number of wavelengths processed. We believe that using a larger number of bands in the visible and near infrared (400-1000 nm) could improve characterization of reflectance properties of the gastric mucosa varying between healthy tissue and inflammatory or malignant lesions.

For this purpose we developed a multispectral-augmented endoscopic prototype illustrated on Figure 1. It is based on a Olympus (Tokyo, Japan) EVIS EXERA III endoscopic system and a fiberscope (IT-Concepts microflex m2.5-2500) is inserted into the operators canal and connected to a multispectral camera. This allows simultaneous acquisition of white light (WL) and a multispectral video (i.e., 41 spectral bands in the range of 470 to 975 nm). This prototype offers a familiar protocol for the clinician: he first introduces the endoscope on the patient, then the fiberscope into the operator's canal and performs the simultaneous multi-modal exploration as with a conventional endoscope. The multispectral probe works as a localized optical biopsy for medical exploration with a much smaller field of view.

There are several registration issues to overlay both modalities (i.e., WL and multispectral): the two images have different points of view, different resolutions, different focal lengths and distortions. In addition, a conventional off-line static calibration using a chessboard pattern can be used to estimate intrinsic

Zenteno, O., Treuillet, S. and Lucas, Y.
3D Cylinder Pose Estimation by Maximization of Binary Masks Similarity: A simulation Study for Multispectral Endoscopy Image Registration.
DOI: 10.5220/0007400808570864
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 857-864
ISBN: 978-989-758-354-4

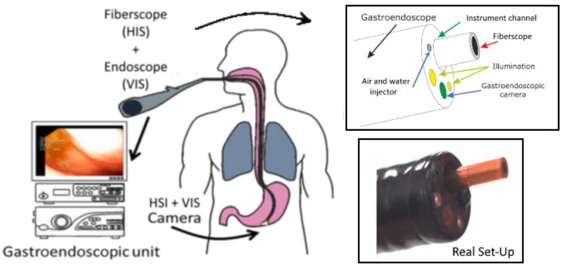

Figure 1: Augmented multispectral prototype.

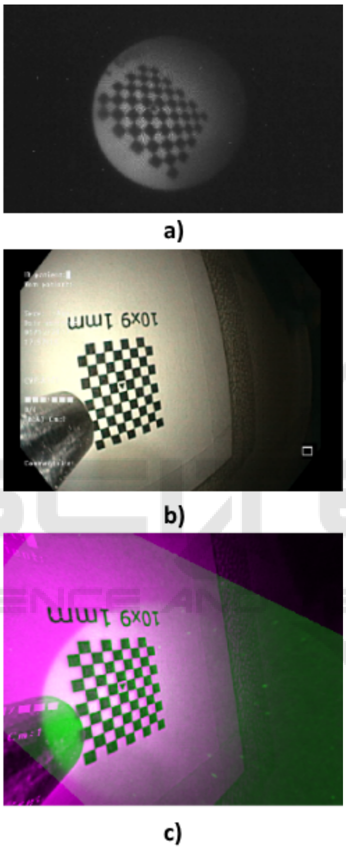

Figure 2: a) Fiberscopic Image b) Endoscopic Image c) Multimodal-enhanced image.

parameters of each cameras (focal length and distortions) but the relation of both cameras cannot be considered rigidly fixed due to the fiberscope being removed between patients for sterilization. Therefore, a variation in the initial fiberscope insertion and a slight bending of the fiberscope tip (precession) is always present.

To perform the multimodal image registration during exploration we used the fiberscope's tip, which is always visible in the endoscopic image. A first approach was proposed in (Zenteno et al., 2018) based on an off-line training with a chessboard pattern of an adaptive affine transform. The transformation compensates the zooming and decentering effect produced by the insertion/retraction movement of the fiberscope during in-vivo exploration. Although this is really useful for insertion and retraction, it does not take into account more complex movements which can be induced by fiberscope manipulation like precession or axis displacement.

This paper presents a new approach based on a 3D cylinder model of the fiberscope to achieve a more robust tracking of the tip's pose and improve the image registration accuracy. The remainder of this document is organized as follow: Section 2 makes a review of related works, Section 3 describes the method, Section 4 the results obtained and Sections 5 concludes the manuscript.

2 RELATED WORKS

The present multimodal image registration problem is similar to the pose estimation of a tubular instrument which is a classic issue of visual servoing for laparoscopy and has been presented before in the literature. The application of artificial landmarks is a common practice as in (Kim et al., 2003) or (Tonet et al., 2007). However, in the case of surgical instruments with direct contact to human tissue, particular medical requirements such as the biocompatibility and the sterilisability of the artificial markers have to be met. (Doignon et al., 2008) presents several 3-D pose estimation algorithms and visual servoing-based tracking of tubular instruments with monocular vision systems such as endoscopes and CT scanners. Another approaches using the video information provided by the endoscopic camera have been proposed in (Cabras et al., 2017) and (Reilink et al., 2013). The first relies on colored markers attached onto the bending section. The image of the instrument is segmented using a graphbased method and the corners of the markers are extracted by detecting the color transitions along Bezier curves fitted on edge points. The latter uses the positions of three markers in the endoscopic image or three feature points to update the state of a kinematic model of the endoscopic instrument. However, these existing solutions does not use multimodal images or have been applied to ad-hoc laboratory setup which cannot be directly used for real surgical systems. In this paper, we propose a landmark-free approach to dynamically estimate the pose changes between the two cameras using only a binary segmentation of the fiberscope tip in the endoscopic images, for a robust real time image registration.

3 METHODOLOGY

3.1 Dual Camera Calibration

A set of chessboard pattern images is used to estimate the intrinsic parameters and the distortion of both cameras (endoscope and fibroscope). Calibration from real images of a chess-board pattern gives a realistic initial pose for the two cameras using the homography-derived extrinsics parameters which provide the translation vector T and rotation matrix R relative to the world coordinates. The reference frames attached to each camera take the optical center C for origin with its optical axis Z and the image plane XY designated as C_z and C_{xy} respectively.

Figure 4 depicts the position of the three coordinate reference frames used for the virtual modelization:

- The world reference frame (W) which origin is at pattern's left-top corner. W_{xy} correspond to the pattern horizontal and vertical dimension and W_z is oriented on the acquisition system opposite direction.
- The Fiberscopic reference frame (FC) which its origin is at the end of the fibercope. FC_{xy} and FC_z correspond to the camera plane and its camera optical axis respectively.
- The Endoscopic reference frame (EC) which its origin is at the end of the endoscope. EC_{xy} and EC_z correspond to the camera plane and its camera optical axis respectively.

3.2 Fiberscope Model

The fiberscope is represented following its real geometrical properties as a straight cylinder with a fixed diameter of 2.5mm. The cylinder's pose is defined by two points: Z_1 and Z_2, which are the two extremities of its axis. The point Z_1 is fixed behind the initial FC_z to be used as a pivot point. The point Z_2 is at the end of the tip and it moves with the center of the camera.

We modeled two different movements: insertion and precession (Fig 3). To model insertion the location of Z_2 is translated along FC_z according to the desired depth and to model precession the position of Z_2 moves in FC_{xy} by two values (p_x and p_y). All coordinate points are then transformed into FC.

The final extremity of the fiberscope's tubular model (estimated camera location) is given by:

$$Z_2 = R_F \cdot [p_x, p_y, depth]^T + L_F \tag{1}$$

$$L_F = -T_F \cdot {R_F}^T \tag{2}$$

where L_F is the initial pose given by the initial calibration (i.e, extrinsics parameters of fiberscopic cameras).

Therefore the fiberscopic camera pose will be defined by the vector $[p_x, p_y, depth]^T$ regarding FC.

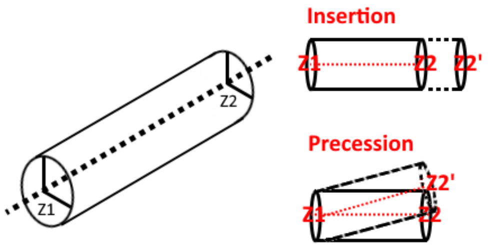

Figure 3: Simulated fiberscope movements: insertion and precession.

3.2.1 Image Projection

To simulate the observation of the fiberscope in the endoscopic image the 3D cylindrical model is projected in 2D by using the projection matrix of the endoscopic camera defined as follow

$$m_E = P \cdot MP = K_E \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} [R_E | T_E] \tag{3}$$

where
m_E = Endoscopic 2D projection
P = Projection matrix
M = 3D point coordinates
K_E = Endoscope calibration matrix

Figure 4 depicts the simulation interface where the insertion depth, angles of precession of the fiberscope and rotation of the camera can be modified interactively. The lower part presents: c) the cylinder model and virtual pattern projection superposed to the original endoscopic image, d) the original fiberscopic image and e) the binary mask provided by the projection of the world points in the 2D endoscopic plane.

3.3 Pose Estimation by Binary Mask Similarity Maximization

The relative pose estimation between the two heterogeneous cameras (endoscopic/fiberscopic) is expressed as a maximization problem by fitting the projection of the 3D cylinder into the fiberscope's segmented tip in the endoscopic image. To do this we maximize a similarity measure between two binary masks A and B.

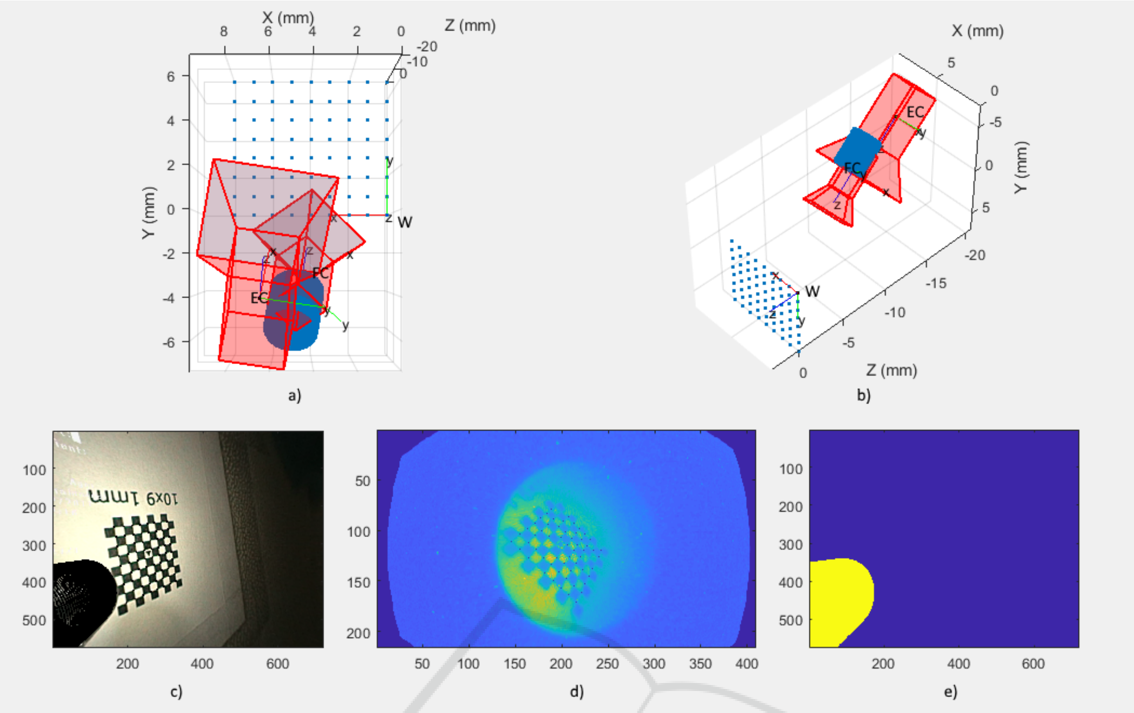

Figure 4: MATLAB simulator 3D view.

$$X = \arg\min[1 - S(A,B)] \tag{4}$$

where X is the vector $[p_x, p_y, \text{depth}]^T$ representing the pose of the cylinder's extreme point.

3.4 Similarity Index

To compare the two binary masks we tested three commonly used indexes (Csurka et al., 2013). All similarity index S comply the condition $0 \leq S \leq 1$

3.4.1 Jaccard

The Jaccard index J, also known as intersection over union is defined as the size of the intersection divided by the size of the union of the sample sets A and B. :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{5}$$

3.4.2 Dice

Dice or Srensen's measure it is defined as twice the number of elements common to both sets divided by the sum of the number of elements in each set.

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{6}$$

where $|A|$ and $|B|$ are the cardinalities of the two sets (i.e. the number of elements in each set).

3.4.3 BF-score

The BF score measures how close the predicted boundary of an object matches the ground truth boundary. It is defined as the harmonic mean of the precision and recall values. Precision is the fraction of detections that are true positives rather than false positives. Recall is the fraction of true positives that are detected rather than missed.

$$BF(A,B) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

4 SIMULATION RESULTS AND DISCUSSION

4.1 Objective Function

The behavior of the similarity indexes are observed varying the three parameters of the fiberscope pose (i.e., $[p_x, p_y, \text{depth}]^T$) as shown in Figure 5.

Althought the BF-score behaves as a strongly convex function while evaluating the two precession parameters, it lacks of singular minimum values when analyzing the insertion parameter. Jaccard and Dice behave similarly in the three cases, describing a singular min value in the exact fit case. However, Jac-

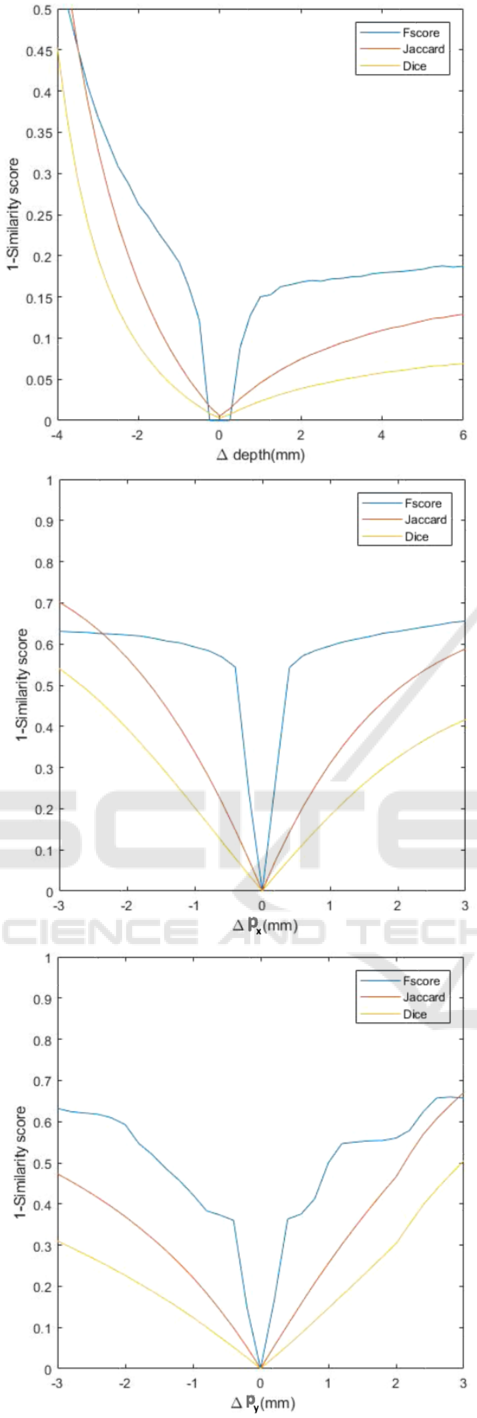

Figure 5: Behavior of objective functions.

card's function presents itself as a better option due to the higher value of its local tangents which may lead to a faster convergence.

4.2 Convergence Validation

Twelve scenarios including highly ill-posed initialization were simulated. An example of four representative cases is presented in Figure 6. The first column depicts the superposition of the initial and the target segmentation. In the same manner, the second column depicts the superposition of the final and the target segmentation. In both cases yellow and green represent the intersecting and individual pixels to the two segmentations. The third column describes the evolution of the three parameters estimates trough iterations, the dotted lines represent the expected values and the crosses the estimated values on each iteration. Finally, the fourth column compare the corners location of the virtual grids for estimated pose vs ground truth pose.

Detailed results including the error statistics for the projected points and a comparison between the initial and final similarity measure for each case is presented in Table 1. In addition a summary of statistics is presented in Table 1. The median error for the sample set was 0.38 ± 0.5 pixels. We expected the initial Jaccard similarity measure to be a determinant for convergence. However, even when its value is low (e.g., cases 5,6,7,9) the minimization can be performed effectively. In contrast, we observed failed convergences were more associated to inaccurate final estimates (i.e.,Table 1). This may be related to the fact that large variations in the optimized parameters lead to small variation in the projected images due to non-linearity included in the perspetive projection

4.3 Tracking Scenario

Figure 7 depicts the evolution of the estimated parameters trough a simulation of combined movements of the tip during an exploration. From initialization, the pose of the fiberscope is estimated frame by frame by using the final estimate of the previous frame as initial values for the current one. The trajectory was determined by a combination of the parameters of insertion and precession in an aleatory manner. The overall mean absolute error in this measurements was of 0.05 ± 0.07 mm. The number of iterations needed for the initial fit was 75 and the number of iterations needed for the following cases was 35 ± 5. Indeed, the initial fit requires around the double of iterations than the other points of the curve

We observe that in all cases the convergence was achieved satisfactorily with high similarity measure values and all differential errors being around zero. This was expected due to the relative similarity between successive frame. In addition, the convergence

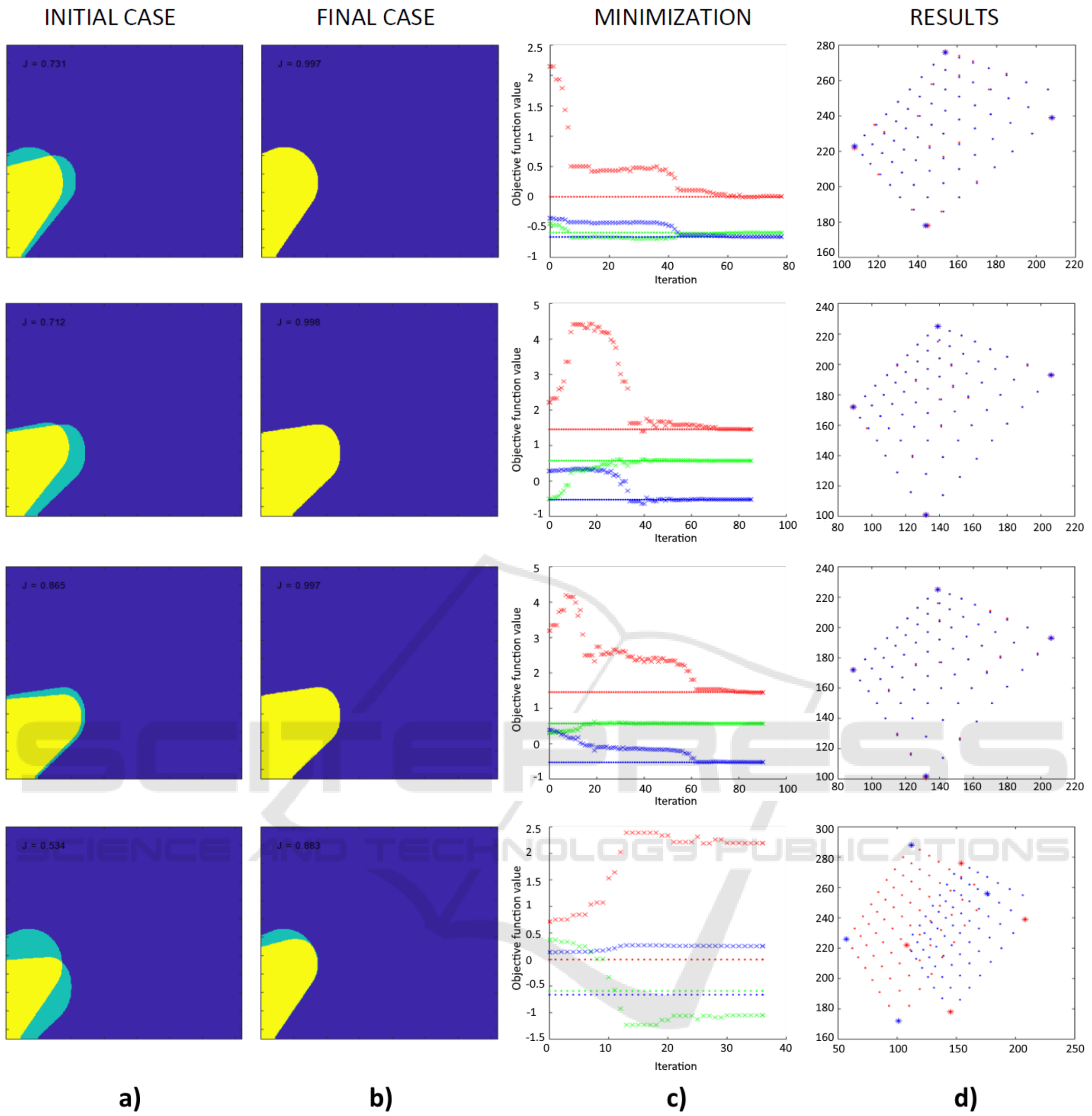

Figure 6: a) 2D projection of simulated sets before maximization, b) 2D projection of simulated sets after maximization, c) Detail of the three objective function parameters over iteration time, d) reprojection results.

speed is also slower. So the minimization problem could be critical for initialization in the first frame only. However the approach presented in (Zenteno et al., 2018) can provide a precise first initialization.

5 CONCLUSIONS

This paper has presented a method for simulating and compensate two sources of movement encountered during multi-spectral endoscopic acquisition for multimodal registration (i.e, the insertion and precession motion of a fiberscope inserted in the instrument channel of an endoscope). The technique relies on applying an homographic transformation between modalities by using a virtual reference pattern projected in both frames as control points. The results showed that the method can track the camera insertion and precession motion. Although the pipeline is still currently executed off-line, this paper demonstrates the potential of image-based tracking of a fiberscope.

Table 1: Similarity measure comparison and projection error statistics.

	Target pose (mm)			Initial pose (mm)			Estimated pose (mm)			Jaccard		$E_{Reprojection}$	
No	depth	p_x	p_y	depth	p_x	p_y	depth	p_x	p_y	Initial	Final	Mean	STD
1	3.17	0.00	0.00	1.62	0.00	0.00	3.177	-0.00	0.00	0.93	0.99	0.25	0.50
2	3.17	0.00	0.00	4.20	0.00	0.00	3.18	0.00	0.01	0.97	0.99	0.00	0.00
3	3.17	0.00	0.00	3.55	-0.21	-0.31	3.16	0.00	-0.00	0.87	0.99	0.00	0.00
4	3.17	0.00	0.00	3.55	0.64	0.29	4.29	-0.08	0.37	0.77	0.96	14.53	7.55
5	0.00	-0.59	-0.66	2.15	-0.44	-0.35	0.00	-0.59	-0.66	0.71	0.99	0.00	0.00
6	1.45	0.57	-0.52	2.21	-0.50	0.28	1.45	0.57	-0.52	0.86	0.99	0.25	0.50
7	1.45	0.57	-0.52	3.19	0.30	0.39	1.44	0.56	-0.52	0.53	0.88	43.87	6.10
8	0.00	-0.59	-0.66	0.71	0.37	0.13	2.19	-1.05	0.25	0.73	0.99	0.50	0.58
9	1.85	-0.59	0.66	2.21	0.36	-0.31	1.86	-0.58	0.66	0.68	0.99	0.25	0.50
10	1.85	-0.59	0.66	2.99	0.36	0.77	4.64	-0.99	1.74	0.66	0.92	39.52	17.68
11	2.34	0.00	0.00	2.41	-0.21	-0.16	2.22	0.03	-0.04	0.89	0.99	1.97	0.39
12	1.37	0.00	0.00	6.48	0.00	0.00	1.38	0.00	0.00	0.85	0.99	0.75	0.50

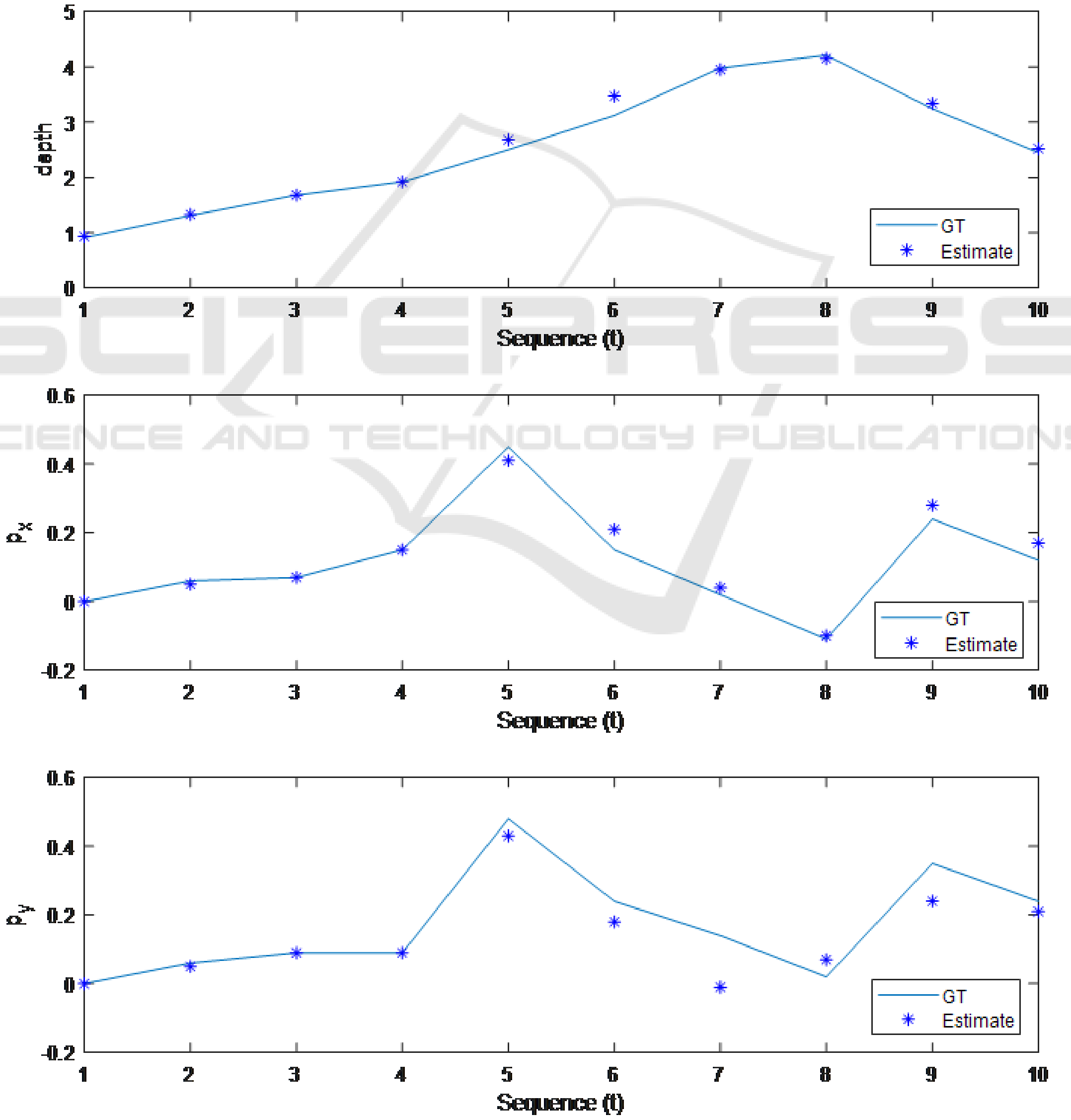

Figure 7: Comparison of estimated values versus ground truth along a 10 step sequential movement mimicking scenario.

ACKNOWLEDGEMENTS

This work was supported by the EMMIE (Endoscopie MultiModale pour les lésions Inflammatoires de l'Estomac) project funded by the ANR-15-CE17-0015 grant.

REFERENCES

Cabras, P., Nageotte, F., Zanne, P., and Doignon, C. (2017). An adaptive and fully automatic method for estimating the 3d position of bendable instruments using endoscopic images. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 13(4):e1812.

Csurka, G., Larlus, D., Perronnin, F., and Meylan, F. (2013). What is a good evaluation measure for semantic segmentation?. In *BMVC*, volume 27, page 2013. Citeseer.

Doignon, C., Nageotte, F., Maurin, B., and Krupa, A. (2008). Pose estimation and feature tracking for robot assisted surgery with medical imaging. In *Unifying perspectives in computational and robot vision*, pages 79–101. Springer.

Kim, M., Lee, J., and Huh, J. (2003). Real time visual servoing for laparoscopic surgery. In *International Symposium on Artificial Life and Robotics*, pages 0–0. ISALR.

Reilink, R., Stramigioli, S., and Misra, S. (2013). 3d position estimation of flexible instruments: marker-less and marker-based methods. *International journal of computer assisted radiology and surgery*, 8(3):407–417.

Song, L. M. W. K., Adler, D. G., Conway, J. D., Diehl, D. L., Farraye, F. A., Kantsevoy, S. V., Kwon, R., Mamula, P., Rodriguez, B., Shah, R. J., et al. (2008). Narrow band imaging and multiband imaging. *Gastrointestinal endoscopy*, 67(4):581–589.

Tonet, O., Thoranaghatte, R. U., Megali, G., and Dario, P. (2007). Tracking endoscopic instruments without a localizer: a shape-analysis-based approach. *Computer Aided Surgery*, 12(1):35–42.

Zenteno, O., Krebs, A., Treuillet, S., Lucas, Y., Benezeth, Y., and Marzani, F. (2018). Dual-channel geometric registration of a multispectral-augmented endoscopic prototype. In *VISIGRAPP (4: VISAPP)*, pages 75–82.