3D Surface Reconstruction using Dense Optical Flow combined to Feature Matching: Application to Endoscopy

Tan-Binh PHAN[1], Dinh-Hoan TRINH[1], Dominique LAMARQUE[2], Didier WOLF[1], Christian DAUL[1] *

[1]Université de Lorraine, CNRS, CRAN, 2 avenue de la Forêt de Haye, 54518 Vandœuvre-lès-Nancy, France.

[2] Hôpital Ambroise Paré, 9 Avenue Charles de Gaulle, 92100 Boulogne-Billancourt, France.

tan-binh.phan@univ-lorraine.fr, dinh-hoan.trinh@univ-lorraine.fr
lamarquedominique@gmail.com, didier.wolf@univ-lorraine.fr,
christian.daul@univ-lorraine.fr

Résumé – Dans les algorithmes de structures à partir du mouvement (SfM), la performance de la reconstruction des surfaces dépend fortement de la qualité de la détermination des points homologues entre images. Les méthodes SfM de référence sont souvent inopérantes pour les scènes avec peu de structures et textures faiblement contrastées car elles reposent uniquement sur l'appariement de caractéristiques. Cette contribution présente une solution associant un flot optique dense à la mise en correspondance de caractéristiques. La précision et la robustesse de la reconstruction ont été validées via des résultats obtenus pour un fantôme avec des dimensions connues et avec des données patient en cystoscopie et en gastroscopie, respectivement. Plus généralement, cette approche a un fort potentiel pour toute scènes peu constrastée, médicales ou non.

Abstract – In structure from motion (SfM) algorithms, the surface reconstruction performance strongly depends on the quality of the determination of homologous points between images. Classical feature matching-based methods as integrated in the state-of the-art SfM-algorithms are often inoperative for scenes including weak structures and textures (e.g., as those in medical endoscopic videos). This contribution introduces an effective solution based on the combination of dense optical flow and feature matching. The accuracy and robustness of the proposed method were validated using results obtained for a phantom with known dimensions and with patient data, respectively. Apart from the high performance obtained for cystoscopy and gastroscopy, the proposed solution has a high potential in other medical and non-medical scenes.

1 Introduction

Endoscopy plays a key role in lesion diagnosis, patient follow-up and minimally invasive surgery. However, the lack of extended and textured 3D surfaces is an obstacle to an easy visual interpretation of the scene, whereas the very limited 2D field of view (FoV) does not allow for a diagnosis made from lesions seen entirely.

First attempts to reconstruct 3D FoV extented endoscopic surfaces were based on structured light approaches [1]. However, these approaches led to hardware changes considered as being too significant by endoscope manufacturers. Solutions using only 2D images were proposed by some authors to tackle the 3D reconstruction problem. In the particular case of cystoscopy [2], structure from motion (SfM) methods were used to reconstruct the internal bladder wall surface. Other methods combined SfM with shape from shading approaches [3] to obtain surfaces from endoscopic data.

The SfM-based methods make the assumption that the scene is rigid, and that point correspondences can be established by detecting and matching feature points. Feature points are located with sub-pixel accuracy, while their feature descriptors can be invariant to scale, rotation and intensity changes. Homologous image points are classically obtained by matching feature points using their descriptor vectors, and by rejecting outliers with a RANSAC method [4] taking a homography as transformation model between image pairs. However, numerous endoscopic scenes often consist of surfaces with small (tissue) deformations and the images, affected by strong illumination changes, only include weak structures or textures. As shown for instance in Fig. 1 for gastroscopy, feature based matching approaches lead to too few correspondences when the epithelial wall of the stomach includes poor structure and texture information.

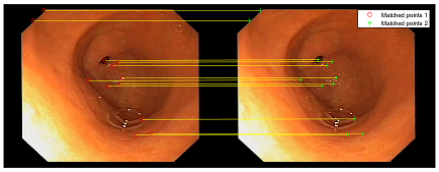

FIGURE 1 – Too few SIFT matches [5] making SfM inapplicable in gastroscopy.

In the case of scenes affected by strong illumination changes and with few information as in Fig. 1, variational optical flow (OF) using illumination-invariant descriptors [6,7] can favourably replace feature based methods. OF methods have an advantage to provide dense point-to-point correspondences between two overlapping images. The authors in [8] integrated an OF step in their SfM approach.

*This work was partially funded by the Agence Nationale de la Recherche (EMMIE project, ANR-15-CE17-0015).

This contribution combines dense OF (DOF) and feature matching to take advantage of both methods : on the one hand, OF is able to give a dense correspondance even in complex scene conditions and, on the other hand, the accuracy of feature points is exploited whenever possible.

This paper is organized as follows. Section 2 gives an overview of a novel 3D reconstruction pipeline in which the SfM algorithm is only one step. The algorithm for the determination of groups of homologous points in the SfM step is the major contribution of this paper and is detailed in Section 3. Section 4 successively illustrates the accuracy and the robustness of the proposed method using 3D surfaces with known dimensions and (endoscopic) patient data, respectively. Finally, a conclusion is given in Section 5.

2 3D reconstruction pipeline

The proposed surface reconstruction algorithm consists of four steps.

Preprocessing : In this step, a set of frames is selected from the input video. Although in this preliminar work the images are manually selected, they could also be automatically chosen according the amount of segmented specular reflection [9] and the measurement of motion blur. The images are also undistorted using the algorithm in [10]. The output of this step is set $S = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N\}$ of N temporally ordered images with a size of $H \times W$ pixels.

Structure from Motion : This step provides a sparse 3D point cloud close to the surface to be reconstructed, as well as the camera poses (i.e., their position and orientation) of the images in set S. These results can only be obtained using groups of homologous 2D points (homologous 2D points are those issuing from the projection of a same 3D point on the images of different viewpoints). Obtaining groups with numerous and accurate points is crucial in SfM. These point groups, together with a classical triangulation algorithm, are used for an initial estimation of the 3D point positions and camera poses which are refined with a bundle adjustment technique [11]. This step ends with a dense point cloud computation algorithm [12] which performs a completion of the surface initially represented by the sparse SfM point cloud.

Mesh generation : The algorithm described in [13] is used to build a meshed surface of triangular facets using the dense point cloud provided by previous step.

Surface texturing : The algorithm described in [14] is used to superimpose the image textures onto the meshed surface.

Next section details the proposed joint integration of DOF and feature matching into the SfM step.

3 Determination of point groups

A point group is defined from at least three images $\mathbf{I}_i, \mathbf{I}_j, \mathbf{I}_k$ taken all from different viewpoints and with $1 \leq i, j, k \leq N$. If $\left(\mathbf{p}_i^{a^1}, \mathbf{p}_j^{a^2}\right)$ and $\left(\mathbf{p}_j^{a^2}, \mathbf{p}_k^{a^3}\right)$ are homologous point pairs in images pairs $(\mathbf{I}_i, \mathbf{I}_j)$ and $(\mathbf{I}_j, \mathbf{I}_k)$ respectively, then $\mathbf{p}_i^{a^1}, \mathbf{p}_j^{a^2}, \mathbf{p}_k^{a^3}$ belong to a group a with a minimal length of three points.

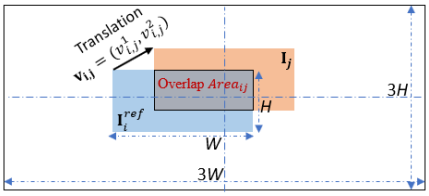

FIGURE 2 – Rectangular overlap of $\mathbf{I}_j$ with $\mathbf{I}_i^{ref}$ and vector $\mathbf{v}_{i,j}$ between their centres.

The main idea of proposed method is to search scene regions seen in as numerous images as possible and to use the DOF and/or feature matching methods to determine the homologous points between image pairs. SfM is accurate when the 3D points are reconstructed from numerous viewpoints. In scene regions with a large number of image overlaps, the point groups have the highest probability to be large. Since in these common scene regions the corresponding images have not to be registered (homologous points have only to be determined), a simple rectangle can be used to delineate the common parts of image pairs geometrically linked by a translation vector (vector $\mathbf{v}_{i,i+1}$ in Fig. 2). Next sections detail the proposed three step algorithm.

Step 1 : *Determination of image translations*. Two matching methods are used to find the translation vectors $\mathbf{v}_{i,i+1}$ between two consecutive images $\mathbf{I}_i$, $\mathbf{I}_{i+1}$ of a sequence. Since feature-based methods have the highest accuracy, it is first checked if the SIFT algorithm can be used to find the translation between $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$. When this attempt with SIFT fails, a DOF method [6] is used to determine $\mathbf{v}_{i,i+1}$ in a robust way.

Let K_i ($i \in [1, \ldots, N]$) be the set of $|K_i|$ feature points detected in $\mathbf{I}_i$ by the SIFT algorithm [5]. The feature points of sets K_i and K_{i+1} are matched using their descriptor vectors, and by rejecting outliers with the RANSAC method [4] taking a homography as transformation model between images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$. Set $M^{i,i+1}$ corresponds to the set of $\left|M^{i,i+1}\right|$ point pairs which were successfully matched.

The feature-based matching is considered as valid under two conditions : (*i*) the number of detected features must be above a threshold α for images $\mathbf{I}_i$, $\mathbf{I}_{i+1}$ (i.e., $|K_i|$ and $|K_{i+1}| > \alpha$) and (*ii*) the number of matches $\left|M^{i,i+1}\right|$ must be larger than threshold β. If these two conditions are fulfilled, the components $(v^1_{i,i+1}, v^2_{i,i+1})$ of vector $\mathbf{v}_{i,i+1}$ take the value of the translation parameters located in the last column of the homography matrix taken as model in RANSAC. The DOF from $\mathbf{I}_i$ to $\mathbf{I}_{i+1}$ is computed when at least one of the two previous conditions is not fulfilled. This vector field between consecutive images (denoted by $\mathbf{F}_{i,i+1}$) is computed with a robust variational method [6] developed for scenes with few textures and affected by strong illumination changes. The central vector of flow field $\mathbf{F}_{i,i+1}$ of $\mathbf{I}_i$ is taken as translation $\mathbf{v}_{i,i+1}$.

The translation vectors between two non-consecutive images $\mathbf{I}_i$ and $\mathbf{I}_j$ (with $j > i+1$) are defined by the sum of the vectors between the consecutive images from i to j :

$$\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j}) = \sum_{t=i}^{j-1} \mathbf{v}_{t,t+1}(v^1_{t,t+1}, v^2_{t,t+1}). \qquad (1)$$

Step 2 : *Determination of reference images favouring large point groups.* As sketched in Fig. 2, images $\mathbf{I}_i$ and $\mathbf{I}_j$ are called τ-overlapped if and only if :

$$\begin{cases} Area_{i,j} = (W - |v^1_{i,j}|)(H - |v^2_{i,j}|) \geq \tau \\ -W < v^1_{i,j} < W \\ -H < v^2_{i,j} < H, \end{cases} \quad (2)$$

where $W \times H$ is the image size, $Area_{i,j}$ is the overlap area and $\tau > 0$ is a threshold parameter.

Reference images $\mathbf{I}_i^{ref}$ share common scene regions with numerous other images. Hence, images $\mathbf{I}_i^{ref}$ are images $\mathbf{I}_i$ that fulfill two conditions : a reference image must be τ-overlapped with as much as possible of other images and two reference images cannot be τ-overlapped (i.e. they do not satisfy equation (2)). For each image $\mathbf{I}_i$ ($i = 1, 2, \ldots, N$), S_i is the set of $|S_i|$ images which are τ-overlapped with $\mathbf{I}_i$. Let Ω^{ref} ($\Omega^{ref} \subset S$) be the set of reference images $\mathbf{I}_i^{ref}$ to be selected. The determination of set Ω^{ref} is detailed in Algorithm 1.

Algorithm 1 Determination of Reference Images

Input: set S of N consecutive images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$, area threshold τ, and vectors $\mathbf{v}_{1,2}, \mathbf{v}_{2,3}, \ldots, \mathbf{v}_{N-1,N}$.
Initiation : $\Omega^{ref} = \emptyset$, $G = \{S_1, S_2, \ldots, S_N\}$.
While $G \neq \emptyset$
– $\Omega^{ref} \leftarrow \Omega^{ref} \cup \mathbf{I}_k$, where k satisfies $|S_k| \geq |S_i|$, for all $S_{i \neq k} \in G$.
– For all images $\mathbf{I}_j \in S_k$, removing corresponding set S_j from G :

$$G \leftarrow G \setminus \bigcup_{j:\mathbf{I}_j \in S_k} S_j. \quad (3)$$

End
Output: Set Ω^{ref} of the reference images $\mathbf{I}_i^{ref}$.

Step 3 : *Point group determination.* Point groups are computed for each reference image $\mathbf{I}_i^{ref}$ by determining the homologous points for all pairs $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$, with $\mathbf{I}_j \in S_i$. According to the SIFT algorithm efficiency defined in step 1, one among three methods is used to optimize the accuracy and robustness of the homologous point determination between $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$:

- If enough SIFT points are detected in both images ($|K_i|$ and $|K_j| > \alpha$) and successfully matched ($\left|M^{i,j}\right| > \beta$), then the homologous points are computed with SIFT and RANSAC.

- If enough SIFT points are detected in the reference $\mathbf{I}_i^{ref}$, but not enough SIFT points were found in $\mathbf{I}_j$ ($|K_j| \leq \alpha$) or the matching failed ($\left|M^{i,j}\right| \leq \beta$), when for each point $\mathbf{p}_a^{i,ref} \in K_i$, the point $\mathbf{p}_a^j \in \mathbf{I}_j$ defined by $\mathbf{p}_a^j = \mathbf{p}_a^{i,ref} + \mathbf{F}_{i,j}\left(\mathbf{p}_a^{i,ref}\right)$, is the homomologous of $\mathbf{p}_a^{i,ref}$ if it is preserved by specular reflections and occlusions in $\mathbf{I}_j$.

- If not enough SIFT points can be found in $\mathbf{I}_i^{ref}$, the homologous point search is completely based on the flow field $\mathbf{F}_{i,j}$ from $\mathbf{I}_i^{ref}$ to $\mathbf{I}_j$. A grid $\mathbf{C}_i^{ref}$ of 2D points in $\mathbf{I}_i^{ref}$ is created, $h \times h$ being the square cell grid size :

$$\mathbf{C}_i^{ref} = \{\mathbf{p}_{xy}^{i,ref}(xh, yh) \mid x, y \in \mathbb{N}, x \leq \frac{W}{h}, y \leq \frac{H}{h}\}. \quad (4)$$

Each $\mathbf{p}_{xy}^j \in \mathbf{I}_j$, defined by $\mathbf{p}_{xy}^j = \mathbf{p}_{xy}^{i,ref} + \mathbf{F}_{i,j}\left(\mathbf{p}_{xy}^{i,ref}\right)$, is a homologous point of $\mathbf{p}_{xy}^{i,ref}$ in $\mathbf{I}_i^{ref}$.

4 Experimental Results

Experimental results are given for both phantom and real endoscopic data. For all results, the grid size in equation (4) is $h \times h = 10 \times 10$, while the SIFT point detection and matching thresholds are $\alpha = 100$ and $\beta = 50$, respectively. τ in equation (2) is set to $\frac{2}{3}WH$. The results of the proposed method are compared with those of the COLMAP software [15] which is a state-of-the-art solution for multi-view 3D reconstruction based on SfM. COLMAP uses SIFT features to find homologous point groups.

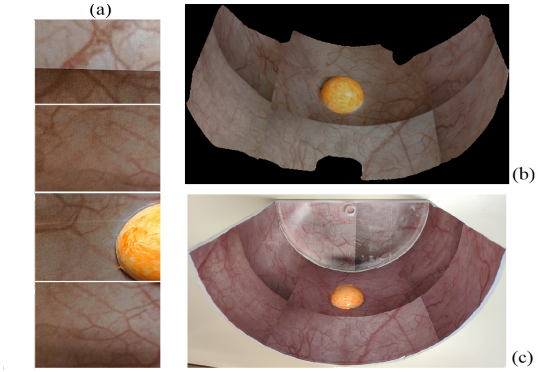

FIGURE 3 – Phantom tests. (a) Four small FoV images. (b) Reconstructed surface under the viewpoint of the snapshot in (c). (c) Snapshot of the phantom (top view).

4.1 Objective evaluation based on ground truths

An objective evaluation is impossible on endoscopic data since for patients no ground truth is available. The phantom in Fig. 3 consists of a cylinder with known diameter ($D = 191.8$ mm) and that carries an orange sphere those diameter ($d = 40.1$ mm). Cystoscopic images were printed on a paper sheet that was glued onto the cylinder surface. A camera and an objective with a 12 mm focal length were used to acquire a sequence of 293 images (with a size of 780×580 pixels) of the phantom. Four of these small FoV images acquired from different viewpoints are shown in Fig. 3(a). As in medical endoscopy, where the acquisitions are done close to the tissue, each image only visualise a small internal object region.

For both SfM methods, the dense point cloud is used, together with a fitting technique, to obtain the cylinder and sphere equations. The maximum allowable distance from a 3D inlier point to the cylinder and to the sphere is set to 1 mm. An objective evaluation of the dense 3D point clouds accuracy is possible by comparing the diameter ratio D/d of the reconstructed surfaces with the ground truth $D/d = 4.78$. This ratio is constant even if the two SfM algorithms reconstruct surfaces at an unknown scale. The diameter ratios obtained with the COLMAP software and with the proposed method are 4.72 (98.76%) and 4.87 (98.29%), respectively (a percentage corresponds to the absolute value of the difference between the ground truth and the computed ratio divided by the ground

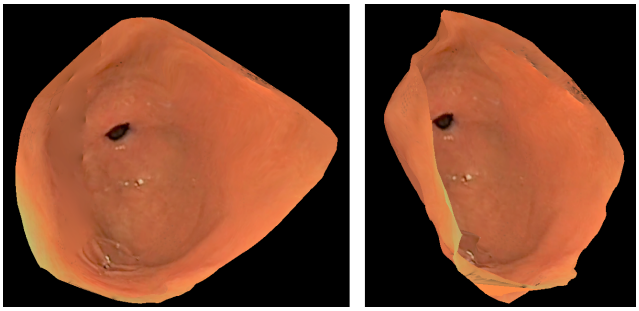

FIGURE 4 – Internal stomach surface under two viewpoints.

truth ratio). These results highlight the accuracy of the proposed method since its performances are the same to those of COLMAP which has a high precision in presence of contrasted textures. Both reconstruction methods are really close to the ground truth.

4.2 Subjective evaluation on patient data

The surface in Fig. 4 was reconstructed using 39 images from a gastroscopic video of the stomach. The shape of the pyloric antrum region is very realistic and was recovered mainly due to OF matches. COLMAP failed completely in the reconstruction of this surface since only few SIFT points can be matched in these images (see Fig 1).

A cystoscopic video-sequence of 2468 images was used to reconstruct a large part of the internal bladder wall surface (see two images of the sequence in Fig. 5(a)). The 3D surface in Fig. 5(b) was constructed with homologous points given by DOF fields for images with few textures (top image in Fig. 5(a)) and by SIFT matches (bottom image in Fig. 5(a)). It allows for a second diagnosis (after the endoscopy) by zooming on regions of interest (polyp in Fig. 5(c)) of the archived map.

5 Conclusion

This paper gives an overview on a robust SfM-based pipeline. The main contribution lies in the integration of a joint DOF and feature matching in the SfM step for generating large 2D point groups using Algorithm 1, even in complex scenes. Although results were only shown in endoscopy, the proposed solution can be used for scenes with few textures and strong illumination changes.

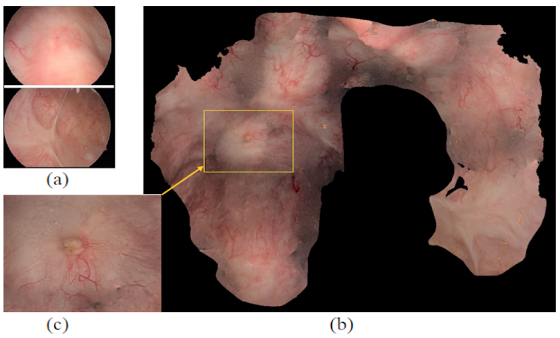

FIGURE 5 – Bladder reconstruction. (a) Two small FoV images. (b) Extended internal surface. (c) Zoom on a polyp.

Références

[1] A. Ben-Hamadou, C. Daul, and C. Soussen, "Construction of extended 3D field of views of the internal bladder wall surface : A proof of concept," *3D Research*, vol. 7, no. 3, pp. 1–23, 2016.

[2] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. Ellerbee Bowden, "3D reconstruction of cystoscopy videos for comprehensive bladder records," *Biomedical Optics Express*, vol. 8, no. 4, pp. 2106–2123, 2017.

[3] Q. Zhao, T. Price, S. Pizer, M. Niethammer, R. Alterovitz, and J. Rosenman, "The endoscopogram : A 3D model reconstructed from endoscopic video frames," in *MICCAI*, 2016, pp. 239–447.

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

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] D.-H. Trinh, W. Blondel, and C. Daul, "A general form of illumination-invariant descriptors in variational optical flow estimation," in *IEEE ICIP*, 2017, pp. 2533–2537.

[7] D.-H. Trinh and C. Daul, "On illumination-invariant variational optical flow for weakly textured scenes," *Computer Vision and Image Understanding*, vol. 179, pp. 1–18, 2019.

[8] T. Schneevoigt, C. Schroers, and J. Weickert, "A dense pipeline for 3D reconstruction from image sequences," in *GCPR*, 2014, pp. 629–640.

[9] D.-H. Trinh, C. Daul, W. Blondel, and D. Lamarque, "Mosaicing of images with few textures and strong illumination changes : Application to gastroscopic scenes," in *IEEE ICIP*, Athens, Greece, 2018, pp. 1263–1267.

[10] R. Miranda-Luna, W. Blondel, C. Daul, Y. Hernandez-Mier, R. Posada, and D. Wolf, "A simplified method of endoscopic image distortion correction based on grey level registration," in *IEEE ICIP*, 2004, pp. 3383–3386.

[11] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment - A modern synthesis," in *IWVA*, 1999, pp. 298–372.

[12] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch : a randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 24 :1–24 :11, 2009.

[13] M. M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *ESGP*, 2006, pp. 61–70.

[14] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color ! large-scale texturing of 3D reconstructions," in *ECCV*, 2014, pp. 836–850.

[15] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE CVPR*, 2016, pp. 4104–4113.