# Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces

Tan-Binh Phan, Dinh-Hoan Trinh, Didier Wolf, Christian Daul

▶ **To cite this version:**

Tan-Binh Phan, Dinh-Hoan Trinh, Didier Wolf, Christian Daul. Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. Pattern Recognition, 2020, 105, pp.107391. 10.1016/j.patcog.2020.107391 . hal-02666537

**HAL Id: hal-02666537**
**https://hal.science/hal-02666537**

Submitted on 20 May 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0031320320301941
Manuscript_f4ae083ce737bc46a192ccb64c127198

# Optical Flow-based Structure-from-Motion for the Reconstruction of Epithelial Surfaces

Tan-Binh Phan[a], Dinh-Hoan Trinh[a], Didier Wolf[a], Christian Daul[a,*]

[a] *CRAN, UMR 7039, Université de Lorraine and CNRS, 2 avenue de la Forêt de Haye, 54518 Vandœuvre-Lès-Nancy Cedex, France.*

**Abstract**

This paper details a novel optical flow-based structure from motion (SfM) approach for the reconstruction of surfaces with few textures using video sequences acquired under strong illumination changes. An original image search and grouping strategy allows to reconstruct each 3D scene point using a large set of 2D homologous points extracted from a reference image and its superimposed images acquired from different viewpoints. A variational optical flow scheme with a descriptor-based data term leads to a robust, accurate and dense homologous point determination between the image pairs. Thus, contrary to classical SfM usable for textured scenes, the proposed dense point cloud reconstruction algorithm requires neither a feature point tracking method nor any multi-view stereo technique. The performance of the proposed SfM approach is assessed on phantoms with known ground truth and on very complex patient data of various medical examinations and image modalities.

*Keywords:* 3D image mosaicing, Structure-from-Motion (SfM), Dense optical flow, Endoscopy, Dermatology.

## 1. Introduction

Multiview 3D techniques aim to reconstruct scenes with an extended field of view (FoV) using sequences of 2D images with limited FoV. The intrinsic cam-

*Corresponding author.
*Email address:* christian.daul@univ-lorraine.fr (Christian Daul)



© 2020 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

era parameters are usually obtained either through a offline calibration or are directly estimated with the images used to reconstruct the scene [1]. Multiview 3D techniques recover a scene in several steps. The acquired sequences are first preprocessed to correct image distortions or remove images with poor quality (e.g., blurred data). The 3D scene structure is reconstructed in the second step, referred to as structure-from-motion (SfM). According to the image contents, this step is among the most challenging in the whole process. In this SfM step, 3D geometrical structures are obtained using triangulation techniques applied on groups of homologous points seen (preferably) in numerous 2D images [2], the 3D point positions being refined by a bundle adjustment [3]. The performance of the determination of homologous points (matching) is a key issue in SfM. Almost all SfM methods in the literature determine homologous points using feature detection and matching algorithms (as SIFT [4] or SURF [5]). The SfM step delivers sparse 3D point clouds since feature based methods detect a limited number of points in images of most of the scenes. Multi-view Stereo (MVS) techniques represent a classical step used to increase the density of the 3D point clouds. Patched-MVS [6], CMPMVS [7], and MVS [8] are state-of-the-art MVS methods. In the next step, a mesh generation algorithm (as the Poisson surface algorithm in [9]) uses the dense 3D point cloud to approximate surfaces with triangular facets. These meshes are usually refined to obtain the final surface [10]. Finally, the superimposition of the 2D image textures onto the meshed surface leads to a visually coherent scene rendering [11].

Feature-based SfM methods were used to recover the surface of objects of a few centimeters of diameter up to one kilometer across (see [12]). SfM-based 3D reconstruction was also used to reconstruct large monuments [13], or even complete city districts [14] with high accuracy. However, there is a class of medical scenes for which feature based-SfM approaches are an optimal solution.

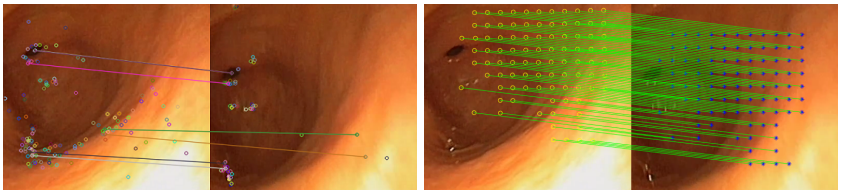
### *1.1. Medical context*

The epithelium (tissue that covers the external human body surface or that lines the internal wall of all hollow organs) is visualized by cameras in various

medical examinations. In dermatology, in gastroenterology and in cystoscopy the epithelial surfaces (respectively corresponding to the skin, the inner stomach wall or the inner bladder wall) are scanned by a camera to search for lesions or to assess their evolution. All these medical applications have a common point: the images are acquired close to the tissue to ensure high image resolution.

Due to these acquisition conditions, the FoV of the images is very limited. Small FoVs do not facilitate the diagnosis since, on the one hand, cancerous lesions on the skin or in the bladder have to be completely seen and, on the other hand, an urologist or a gastroenterologist cannot mentally visualize the endoscope position in hollow organs. Extending the FoV using mosaicing algorithms favours a simultaneous visualization of complete lesions and of anatomical landmarks helping endoscopists to localize the instrument into the organ. In the last two decades, 2D image mosaicing algorithms were proposed in endoscopy [15, 16]. 2D mosaics increase the FoV, but have two major drawbacks. On the one hand, the 3D organs are projected on a 2D plane defined by the image taken as a reference for the mosaicing. When moving away from this reference image in the mosaic plane, the projection distortions become strong and result both in a loss of image resolution and in an incorrect organ representation at the borders of the mosaics which remain of limited size. On the other hand, 2D mosaics are not in accordance with the 3D mental organ representation of endoscopists or dermatologists. Obtaining extended 3D FoV mosaics using SfM techniques can be of high interest in dermatology and endoscopy.

However, in medical examinations both the acquisition conditions and the scene characteristics are significantly different from those of the applications for which SfM has been proven efficient. First, the reconstruction of 3D points is more accurate when homologous points can be acquired from very different viewpoints. In classical SfM applications (e.g., manufactured part or monument surface construction), the acquisition conditions are controlled in the sense that scene parts can effectively be acquired from very different viewpoints. In dermatology, and more particularly in endoscopy, the camera trajectory is quite difficult to control. Obtaining images of the same organ part from very differ-



(a) SIFT combined with RANSAC (b) Dense optical flow

Figure 1: Comparison of two methods used to determine homologous points in two gastroscopic images. (a) Few homologous points obtained using SIFT and RANSAC. In the lower left image corners some false features points are due to specular refections. (b) Numerous homologous points found with the optical flow method proposed in this contribution.

ent viewpoints is a difficult task. Secondly, images of natural scenes or manufactured parts usually include image primitives (corners, line segments, etc.), contrasted textures and/or a great variation in terms of colours. On the contrary, the color variations are very small in dermatology, while in gastroscopy most images are with very few and weakly contrasted textures and structures.

As shown in Fig. 1(a) for two pyloric antrum images, only few homologous points were found when associating the SIFT algorithm [4] to the RANSAC outlier rejection method [17]. Besides the lack of textures, homologous point determination is also impeded by the strong illumination changes between two acquisitions and inhomogeneous lighting due to viewpoint changes. Specular reflections also favor false point correspondences. Such few and partially wrong matches are not appropriate for a 3D reconstruction using SfM approaches.

### *1.2. Scientific context*

As shown in Fig. 1(b), the optical flow (OF) approach described in this work is usable for scenes with few textures and structures. Although dense optical flow (DOF) provides numerous homologous points between two images, DOF matching techniques have been rarely used in SfM up to now.

Let us consider the following situation to understand the reason for this. Suppose that $\mathbf{I}_i$ and $\mathbf{I}_j$ (with $j \neq i \pm 1$) are two non (temporally) consecutive video-sequence images that share a common scene part. If the images are well

structured/textured, feature detectors and descriptor (e.g., SIFT) can effectively determine (both in quantity and quality) the homologous points between $\mathbf{I}_i$ and $\mathbf{I}_j$. The advantage of the feature matching methods is that the points detected by detector (keypoints) and their descriptors are often invariant to geometric and photometric changes. Thus, point-tracks determined by feature matching often ensure a high accuracy (the key-points can be localized with a subpixel accuracy). In contrary, if an OF-based tracking method is used to find homologous points between $\mathbf{I}_i$ and $\mathbf{I}_j$, flow fields $\mathbf{F}_{k,k+1}$ (with $k = i, i+1, \ldots, j-1$) have to be computed for consecutive image pairs $(\mathbf{I}_k, \mathbf{I}_{k+1})$ from $\mathbf{I}_i$ to $\mathbf{I}_j$. With a starting point $A_i$ in $\mathbf{I}_i$, the tracked sequence of points $(A_i, A_{i+1}, \ldots, A_j)$ is determined, with $A_k = A_{k-1} + \mathbf{F}_{k-1,k}(A_{k-1})$ and $k = i+1, \ldots, j$, and $A_j$ is defined to be the homologous point of $A_i$. Two issues are related to this way to track homologous points: *(i)* even if a very accurate OF method providing a dense flow field between images is used, it is impossible to reach the subpixel accuracy of feature matching methods, and *(ii)* although the errors affecting the OF vectors linking points in consecutive images are weak, these errors accumulate themselves along the sequence and may become quickly large when the length of the point track increases. Therefore, $A_i$ and $A_j$ are often wrong homologous points when the temporal distance $|j - i|$ is large. This lack of accuracy explains why DOF is rarely used in SfM approaches.

The proposed SfM approach is based on the fact that in the scenes where feature detectors are unusable, DOF may be the only option for point correspondance establishment. The global aim of this paper is to show that a DOF-approach can lead to an efficient surface reconstruction solution for scenes for which feature-based methods cannot be used. The described solution is based on two contributions. The paper shows first how a dense point correspondence can be established even in complexe scenes with few textures and strong illumination changes as in Fig. 1(b). Then, one proposes an image grouping strategy that leads to numerous and large homologous point sets enabling a robust surface reconstruction. Compared to the point tracking in consecutive images of a sequence, the proposed image grouping strategy avoids accumulated errors

leading to inaccurate or false correspondences. Moreover, unlike feature-based SfM methods, the proposed DOF-based SfM method directly provides dense 3D point clouds and makes the implementation of a MVS method unnecessary.

### *1.3. Paper organization*

Section 2 presents previous contributions relating to the reconstruction of medical scenes. Section 3 focuses on the two main contributions enabling SfM methods to be robust, namely the OF method for finding numerous homologous points between two images and the strategy for finding numerous 2D homologous image points for each 3D scene point to be reconstructed. Results are first given in Section 4 to compare the performance of a state-of-the-art SfM method based on feature detection (COLMAP [18]) with that of the proposed DOF-based SfM approach. Epithelial surface construction examples are then given for three medical examinations (gastroscopy, cystoscopy and dermatology) to show the large scene variability which can be handled by the proposed SfM scheme. A conclusion and perspectives are presented in Section 5.

## 2. Related work

A straightforward solution to tackle the issue relating to the lack of feature points would be to use active stereo-vision systems projecting light patterns on the surfaces to be reconstructed [19]. An active vision method was developed to show the feasibility of 3D bladder mosaicing [20]. However, such a solution lead to too significant hardware changes for endoscope manufacturers who prefer passive vision solutions keeping the instruments unchanged. Moreover, active vision solutions are usually application dedicated.

Beside active vision systems, passive vision techniques based on shape from shading (SfS), SfM and simultaneous localization and mapping (SLAM) were proposed. A SfS technique was successfully used in endoscopy to reconstruct Lambertian surfaces of bone structures [21]. However, SfS techniques alone are not appropriate for cystoscopic or gastroscopic scenes in which the illumination

conditions drastically change with the viewpoint. Several works [22, 23] have associated SfS with SfM methods in order to simultaneously exploit shading and feature information for the reconstruction of surfaces from endoscopic images. All these works show the potential of SfM, SfM associated with SfS, and SLAM (SLAM [24] can be seen as a particular case of SfM) methods in endoscopy when homologous points can be robustly determined between images.

SfM methods were tested in the specific case of cystoscopy. In [25], the authors replaced the cystoscope by a non-standard system acquiring image sequences using an ultrathin fiber whose trajectory is controlled by a robotic steering system. The spiral shaped camera trajectory ensures numerous image overlaps which favors robust SfM. Surface reconstruction tests were successfully conducted on pig bladders. Although no test on human data was performed, the results obtained in [25] show the potential of SfM in cystoscopy. This potential was confirmed in [26] on clinical data. However, the method in [26] is based on the assumption that a significant amount of homologous points can be extracted and matched using SIFT features for almost all images. This assumption is not always true. On the one hand, there is no warranty to obtain contrasted textures since strong blur affect often images due to the difficulty of controlling the cystoscope speed and the distance between the instrument's distal tip and the inner epithelial surface. On the other hand, large image regions may be without textures due to surgical intervention for lesion removal for instance.

At the best of our knowledge, no solution was proposed in the medical field to build surfaces using textureless images as in gastroscopy.

## 3. Dense Optical Flow for SfM

This section begins by briefly describing a robust illumination-invariant OF method that delivers accurate correspondences even with weakly structured and textured images. Then, section 3.2 details the image grouping strategy which maximizes the sizes of the homologous point sets by uniquely computing the DOF between image pairs (i.e., without tracking homologous points along a

sequence of more than two images). This image grouping strategy is intergraded in the incremental SfM pipeline given in [18] to generate 3D points of scene[1].

### *3.1. Optical flow estimation*

Although numerous state-of-the-art OF methods have been proposed, OF estimation remains challenging in medical scenes. Learning-based OF methods (e.g., FlowNet [27]) are difficult to apply to endoscopic images due to the lack of ground-truth data for training, while feature matching-based methods [28, 29] are often inoperative for weakly structured/textured scenes under strong illumination changes. As shown in [30], the variational OF approach is the more appropriate for complex scenes as in endoscopy. Numerous studies on variational OF have been proposed, e.g. [31, 32, 33].

The variational model for determining the flow field from source image $\mathbf{I}_s$ to target image $\mathbf{I}_t$ is defined as

$$\min_{\mathbf{u}} \left[ E_{reg}\left(\mathbf{u}\right) + \lambda E_{data}\left(\mathbf{I}_s, \mathbf{I}_t, \mathbf{u}\right) \right], \tag{1}$$

where $\mathbf{u}\left(u_x, u_y\right)$ denotes the flow field, $E_{reg}$ is a regularization term that assumes smoothness of solution $\mathbf{u}$, $E_{data}$ is a data-term that measures the similarity of pixels in $\mathbf{I}_s$ and $\mathbf{I}_t$, and $\lambda > 0$ is a parameter controlling the relative importance between the two terms.

In endoscopy, images are often affected by uncontrolled illumination variations and specular reflection (SR). Therefore, the data and regularization terms have to be appropriately designed. To this end, we follow the variational OF model given in [34] where SR pixels and the saturated pixels surrounding SR regions are excluded from the OF estimation, while the illumination variations are controlled using an illumination-invariant descriptor in the data-term.

More precisely, SR regions in $\mathbf{I}_s$ and $\mathbf{I}_t$ are first segmented using the method

[1] An overview video of the proposed algorithm and the MATLAB code for homologous point grouping can be downloaded from `https://github.com/CRAN-BioSiS-Imaging/PR2020`

described in [34]. A binary mask $M_{SR}$ is then computed as follows:

$$M_{SR} = (R_{\mathbf{I}_s} \oplus se) \cup (R_{\mathbf{I}_t} \oplus se)\,, \tag{2}$$

where $R_{\mathbf{I}}$ denotes a binary image in which $R_{\mathbf{I}}(i,j) = 1$ when $(i,j)$ corresponds to coordinates of a SR pixel in image $\mathbf{I}$, and $\oplus$ is the morphological dilation operator associated with a square structuring element $se$. Values at 1 in binary mask $M_{SR}$ correspond to pixels located either inside SR regions in $\mathbf{I}_s$ or $\mathbf{I}_t$ (pixels at 1 before dilation) or close to reflections (pixels at 1 after dilation). After determining SR pixels and their neighbors, the data- and regularization terms in (1) are defined by:

$$E_{data} = \sum_{\mathbf{x}\in\Omega} \theta_{\mathbf{x}} \left\| \mathbf{D}\left(P_{\mathbf{I}_s}(\mathbf{x})\right) - \mathbf{D}\left(P_{\mathbf{I}_t}(\mathbf{x}+\mathbf{u}_{\mathbf{x}})\right) \right\|_2^2, \tag{3}$$

$$E_{reg} = \sum_{\mathbf{x}\in\Omega} \sum_{\mathbf{x}'\in\mathcal{N}_{\mathbf{x}}} \theta_{\mathbf{x}}\theta_{\mathbf{x}'}\omega_{\mathbf{x}}^{\mathbf{x}'} \left\| \mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}'} \right\|_1, \tag{4}$$

where $\Omega$ stands for the image domain and $\mathbf{u}_{\mathbf{x}}$ is the displacement vector from pixel $\mathbf{x}$ in source $\mathbf{I}_s$. $L_1$-regularisation is used in (4) because it is known to better preserve discontinuities compared to $L_2$-regularisation [35]. Parameter $\theta_{\mathbf{x}}$ equals 0 for $M_{SR}(\mathbf{x}) = 1$, and $\theta_{\mathbf{x}} = 1$, otherwise. This ensures that saturated pixels and their close neighbors are not involved in the OF determination. Symbol $P_{\mathbf{I}}(\mathbf{x})$ denotes a small patch[2] centered on pixel $\mathbf{x}$ in image $\mathbf{I}$, and $\mathbf{D}(P_{\mathbf{I}}(\mathbf{x}))$ is a descriptor vector computed with the colours of the pixels in $P_{\mathbf{I}}(\mathbf{x})$. In (4),

[2]The size of the descriptor patches relates to the illumination variation model detailed in [30]. To sum up, illumination changes between two small homologous rectangular regions of images $\mathbf{I}_s$ and $\mathbf{I}_t$ are modelled by an affine relationship between the colors. Both the multiplicative and the additive coefficients of the affine relationship are constant for all pixels of two homologous regions. Complex illumination changes can be modelled by choosing a size of $3 \times 3$ pixels for these regions (the illumination differences can by locally very strong since the values of the coefficients can vary for each small homologous region pairs of $\mathbf{I}_s$ and $\mathbf{I}_t$). The descriptor patches have the same size as the small regions in this illumination change model ($3 \times 3$ pixels) and, as shown in this section, the values of the components of the descriptor vectors have to be independent of the values of the coefficients of the affine relationship between the colors.

$\mathcal{N}_{\mathbf{x}}$ is the set of neighbor pixels $\mathbf{x}'$ in a rectangular region centered on $\mathbf{x}$, and $\omega_{\mathbf{x}}^{\mathbf{x}'}$ is a weighting function which is used to define the mutual support between the pixels at positions $\mathbf{x}$ and $\mathbf{x}'$. The support-weight is computed based both on the color-similarities of pixels, and on their spatial distances:

$$\omega_{\mathbf{x}}^{\mathbf{x}'} = \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|_2^2}{\gamma_1} + \frac{-\|\mathbf{c}_{\mathbf{I}_s}(\mathbf{x}) - \mathbf{c}_{\mathbf{I}_s}(\mathbf{x}')\|_2^2}{\gamma_2}\right). \tag{5}$$

Vector $\mathbf{c}_{\mathbf{I}_s}(\mathbf{x}) = [L(\mathbf{x}), a(\mathbf{x}), b(\mathbf{x})]$ encodes the color of image $\mathbf{I}_s$ at pixel $\mathbf{x}$ in the CIELab space [36], while $\gamma_1$ and $\gamma_2$ are parameters controlling the importance of the colour similarity and the spatial distance.

The epithelial images as in Fig. 1 often include few contrasted textures and structures and are affected by strong illumination changes due, for instance, to viewpoint changes between two image acquisitions. Descriptor vector $\mathbf{D}\left(P_{\mathbf{I}}\left(\mathbf{x}\right)\right)$ of the data-term in (3) has to capture weak intensity variations, while being invariant to illumination changes between $\mathbf{I}_s$ and $\mathbf{I}_t$.

The new descriptor perceives local intensity changes in patches $P_{\mathbf{I}}\left(\mathbf{x}\right)$ having a size of $3 \times 3$ pixels and centered on $\mathbf{x}$. Twelve convolution kernels ($K_1$, $K_2$,...,$K_{12}$, see Fig. 2) are used to capture intensity variations in patches $P_{\mathbf{I}}^g$ ($P_{\mathbf{I}}^g$ are grey-level patches computed with the original $RGB$ patches $P_{\mathbf{I}}$). While kernels $K_d$ with $d = 1, 2, \ldots, 8$ allow to encode gradient components approximating line segments with different orientations, kernels $K_d$ with $d = 9, 10, 11, 12$ are rather similar to corner detectors where the vertex of the detected corners are oriented in the direction of positive x-axis values ($d = 10$), of positive y-axis values ($d = 9$), of negative x-axis values ($d = 12$), or of negative y-axis values ($d = 11$). The twelve component vector $\mathbf{D}(P_{\mathbf{I}})$, is defined as follows.

$$\mathbf{D}(P_{\mathbf{I}}) = \frac{\mathcal{V}(P_{\mathbf{I}})}{\|\mathcal{V}(P_{\mathbf{I}})\|_2}, \tag{6}$$

with $\mathcal{V}(P_{\mathbf{I}}) = [K_1 \otimes P_{\mathbf{I}}^g, K_2 \otimes P_{\mathbf{I}}^g, \ldots, K_{12} \otimes P_{\mathbf{I}}^g]^T \in \mathbb{R}^{12}$, where $\otimes$ denotes the convolution operator. In patches $P_{\mathbf{I}}^g$, the central pixel (whose grey-level value is multiplied by 3) can be seen as the origin of a star shaped structure from which grey-level variations are computed along 12 directions. These grey-level variations encode the shape and sharpness of the local texture or intensity vari-

$$K_1 = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad K_2 = \begin{bmatrix} 0 & -1 & -1 \\ 0 & 3 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad K_3 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 3 & -1 \\ 0 & 0 & -1 \end{bmatrix} \quad K_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & -1 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ -1 & -1 & -1 \end{bmatrix} \quad K_6 = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & -1 & 0 \end{bmatrix} \quad K_7 = \begin{bmatrix} -1 & 0 & 0 \\ -1 & 3 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad K_8 = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$K_9 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad K_{10} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad K_{11} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad K_{12} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

Figure 2: Convolution kernels used to compute the components of the new illumination-invariant descriptor (vector $\mathbf{D}$ in (3)).

ations. With or without illumination changes between images, two descriptors vectors $\mathbf{D}\left(P_{\mathbf{I_s}}\left(\mathbf{x}\right)\right)$ and $\mathbf{D}\left(P_{\mathbf{I}_t}\left(\mathbf{x}+\mathbf{u_x}\right)\right)$ should have the same component values. In [30] it was shown that a $\mathbf{D}$ is invariant to illumination changes when:

$$\mathbf{D}(P_{\mathbf{I}}) = \mathbf{D}(a_{\mathbf{x}}P_{\mathbf{I}} + b_{\mathbf{x}}), \forall a_{\mathbf{x}} \in \mathbb{R}_{>0}, \forall b_{\mathbf{x}} \in \mathbb{R}. \tag{7}$$

As seen in Fig. 2, the sum of the coefficients is null in each convolution kernel $K_d$. It follows that the effect of additive term $b_{\mathbf{x}}$ is compensated since

$$K_d \otimes (a_{\mathbf{x}}P^g_{\mathbf{I}} + b_{\mathbf{x}}) = a_{\mathbf{x}}(K_d \otimes P^g_{\mathbf{I}}), \forall d = 1, 2, \ldots, 12. \tag{8}$$

This leads to $\mathcal{V}(a_{\mathbf{x}}P_{\mathbf{I}} + b_{\mathbf{x}}) = a_{\mathbf{x}}\mathcal{V}(P_{\mathbf{I}})$ and $\|\mathcal{V}(a_{\mathbf{x}}P_{\mathbf{I}} + b_{\mathbf{x}})\|_2 = a_{\mathbf{x}}\|\mathcal{V}(P_{\mathbf{I}})\|_2$. Therefore, the effect of multiplicative term $a_{\mathbf{x}}$ is also compensated:

$$\frac{\mathcal{V}(a_{\mathbf{x}}P_{\mathbf{I}} + b_{\mathbf{x}})}{\|\mathcal{V}(a_{\mathbf{x}}P_{\mathbf{I}} + b_{\mathbf{x}})\|_2} = \frac{a_{\mathbf{x}}\mathcal{V}(P_{\mathbf{I}})}{a_{\mathbf{x}}\|\mathcal{V}(P_{\mathbf{I}})\|_2} = \frac{\mathcal{V}(P_{\mathbf{I}})}{\|\mathcal{V}(P_{\mathbf{I}})\|_2} \tag{9}$$

$$\Leftrightarrow \mathbf{D}(a_{\mathbf{x}}P_{\mathbf{I}} + b_{\mathbf{x}}) = \mathbf{D}(P_{\mathbf{I}}), \forall a_{\mathbf{x}} \in \mathbb{R}_{>0}, \forall b_{\mathbf{x}} \in \mathbb{R}. \tag{10}$$

Thus, vector $\mathbf{D}$, as defined in (6), is an illumination-invariant descriptor.

In this work, the optimization problem defined by (1), (3) and (4) is solved using the projection-proximal point algorithm [37, 38]. Moreover, the well-known coarse-to-fine multiscale warping strategy is also used to deal with large displacements. The size of structural element $se$ in (2) was empirically set to $7 \times 7$, whereas the size of neighborhood $\mathcal{N}_x$ in (4) was $5 \times 5$. The experimental approach described in [30] was adopted to search the optimal values of parameters $\lambda$ in (1),$\gamma_1$ and $\gamma_2$ in (5), as well as the value of the pyramid scale $Py_s$ in the coarse-to-fine strategy. The values of these parameters are constant and given in Section 4.

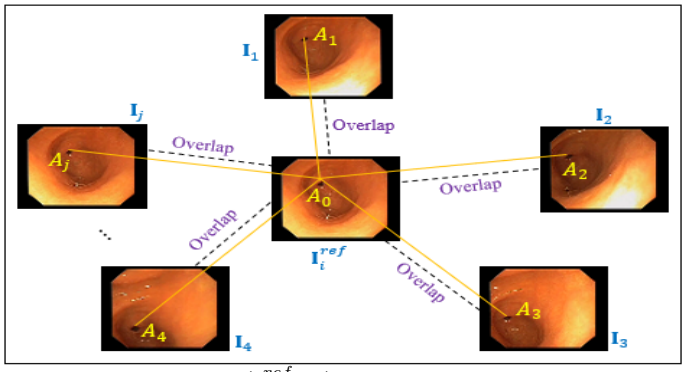



Figure 3: HP-group example. Pairs $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$ consist both of consecutive and non-consecutive images in video-sequence $S$.

When comparing Figs. 1(a) and 1(b), it appears that the DOF-approach can robustly determine homologous point-pairs.

### *3.2. Homologous point set determination for SfM*

Suppose that the input of the SfM algorithm is a video-sequence $S = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N\}$ of $N$ temporally numbered images having a size of $H \times W$ pixels. Let $Z = \{1, 2, \ldots, N\}$ be the index set of $S$. The proposed matching method is based on the fact that if $\{A_k\}_{k \in Z_i \subset Z}$ is a group of 2D homologous points on images in sequence $S$, then $A_k \in \bigcap_{j \in Z_i} \mathbf{I}_j, \forall k \in Z_i$. Therefore, to find homologous point groups, our idea is to first determine reference images, referred to as $\mathbf{I}_i^{ref}$, which have an overlap with a maximum of other images. The set of images overlapping $\mathbf{I}_i^{ref}$ is denoted by $S_i = \{\mathbf{I}_k\}_{k \in Z_i} \subset S$. Then, if $A_0^{ref}$ is a point in $\mathbf{I}_i^{ref}$ and $A_k$ $(k \in Z_i)$ are corresponding points of $A_0^{ref}$ in images $\mathbf{I}_k$ with $k \in Z_i$ and $i \neq k$, then the set $\{A_0^{ref}, A_k\}_{k \in Z_i}$ is defined as a group of homologous points or, in abbreviated form, as a *HP-group* (see Fig. 3).

The next three sub-sections successively present the method for determining overlapping image pairs in sequence $S$, the algorithm which determines the reference images $\mathbf{I}_i^{ref}$ as well as their corresponding sets $S_i$, and the solution for generating HP-groups based on the DOF technique detailed in section 3.1.

*3.2.1. Overlap estimation*

**Definition 1.** *Two images* $\mathbf{I}_i$ *and* $\mathbf{I}_j$ *are called* $\tau$*-overlapped when their common area* $\mathbf{I}_i \cap \mathbf{I}_j$ *is greater than a given threshold* $\tau$ *in pixels.*

When the acquisition distance is small (e.g., as in gastroscopy or cystoscopy where the endoscope distal tip is close to the tissue), the FoV is limited and the displacement field between consecutive images mainly consists of translation vectors. For this reason, simple translations can be used to represent the displacement between common scene parts approximated by rectangular subregions in the images.

The translation vector between two images $\mathbf{I}_i$ and $\mathbf{I}_j$ in $S$ is denoted by $\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j})$, where $v^1_{i,j}$ and $v^2_{i,j}$ are the vector components. Vector $\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j})$ is determined with the DOF fields $\mathbf{F}_{t,t+1}$ between the consecutive images $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ of sequence $\mathbf{I}_i, \mathbf{I}_{i+1}, \ldots, \mathbf{I}_{j-1}, \mathbf{I}_j$. The motion vector at the central pixel $(W/2, H/2)$ of image $\mathbf{I}_t$ to image $\mathbf{I}_{t+1}$ is denoted by $\mathbf{c}_{t,t+1}(c^1_{t,t+1}, c^2_{t,t+1})$ with:

$$\mathbf{c}_{t,t+1}(c^1_{t,t+1}, c^2_{t,t+1}) = \mathbf{F}_{t,t+1}\left(\frac{W}{2}, \frac{H}{2}\right). \tag{11}$$

If two images $\mathbf{I}_i$ and $\mathbf{I}_j$ are consecutive (i.e., $|i-j| = 1$), then the translation vector between images pair $(\mathbf{I}_i, \mathbf{I}_j)$ is defined by:

$$\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j}) = \mathbf{c}_{i_0,i_0+1}(c^1_{i_0,i_0+1}, c^2_{i_0,i_0+1}), \tag{12}$$

with image index $i_0 = \min(i, j)$ making (12) valid for two cases: $j = i - 1$ and $j = i + 1$. For two non-consecutive images $\mathbf{I}_i$ and $\mathbf{I}_j$ (i.e., $|i-j| > 1$), two image indexes are considered : $i_0 = \min(i, j)$ and $j_0 = \max(i, j)$. In this case, the translation between $\mathbf{I}_i$ and $\mathbf{I}_j$ is defined (both for $i > j$ and $i < j$) by the sum of the translation vectors between consecutive images from $\mathbf{I}_{i_0}$ to $\mathbf{I}_{j_0}$:

$$\mathbf{v}_{i,j}(v^1_{i,j}, v^2_{i,j}) = \sum_{t=i_0}^{j_0-1} \mathbf{c}_{t,t+1}(c^1_{t,t+1}, c^2_{t,t+1}). \tag{13}$$

Two images $\mathbf{I}_i$ and $\mathbf{I}_j$ with translation vector $\mathbf{v}_{i,j}$ are $\tau$-overlapped when the

following condition is fulfilled:

$$\begin{cases} -W < v_{i,j}^1 < W \\ -H < v_{i,j}^2 < H \\ Area_{i,j} = (W - |v_{i,j}^1|)(H - |v_{i,j}^2|) \geq \tau, \end{cases} \tag{14}$$

where $W$ and $H$ are the width and height of the images, $Area_{i,j}$ is the overlap area in pixels, and $\tau$ ($0 < \tau \leq WH$) is the threshold parameter. The two first equations in (14) ensure that $\mathbf{I}_i \cap \mathbf{I}_j \neq \emptyset$, whereas the third equation defines the area of the overlap region $\mathbf{I}_i \cap \mathbf{I}_j$.

*3.2.2. Reference images*

The proposed algorithm determines the reference images $\mathbf{I}_i^{ref}$, which favor groups consisting of numerous homologous points. From the practical point of view, such $\mathbf{I}_i^{ref}$ images are those including scene parts which are geometrically surrounded by other scene parts seen in numerous other images acquired from different viewpoints and having common areas with $\mathbf{I}_i^{ref}$. References $\mathbf{I}_i^{ref}$ are images in $S$ that simultaneously fulfill two conditions: *(i)* a reference image must be $\tau$-overlapped with as much as possible of other images, and *(ii)* two reference images cannot be $\tau$-overlapped. The first condition ensures that HP-groups involve numerous images, whereas the second condition favours the distribution of the 3D points over the complete surface. The proposed algorithm (see Algorithm 1) for the determination of reference images consists of two parts.

**Part 1: Determination of the $\tau$-overlapped image sets.** A set $S_i$ (with $i = 1, 2, \ldots, N$) of $\tau$-overlapped images consists of all images $\mathbf{I}_j$ of $S$ which are $\tau$-overlapped with $\mathbf{I}_i$, and of $\mathbf{I}_i$ itself. For all image pairs $(\mathbf{I}_i, \mathbf{I}_j)$ with $j \neq i$, translation vector $\mathbf{v}_{i,j}$ is computed differently depending on whether $\mathbf{I}_i$ and $\mathbf{I}_j$ are consecutive images or not. When $|i - j| = 1$ (consecutive images), translation $\mathbf{v}_{i,j}$ is computed with (12). If $|i - j| > 1$ (non-consecutive images), vector $\mathbf{v}_{i,j}$ is obtained with (13). Set $S_i$ is updated with image $\mathbf{I}_j$ only when the $\tau$-overlap condition given in (14) is fulfilled for image pair $(\mathbf{I}_i, \mathbf{I}_j)$. This algorithm part leads to set $G$ gathering all sets $S_i$: $G = \{S_1, S_2, \ldots, S_N\}$.

**Algorithm 1** Reference image determination

**Input:** Set $S$ of $N$ consecutive images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$, area threshold $\tau$, and central flow field vectors $\mathbf{c}_{1,2}$, $\mathbf{c}_{2,3}$, …, $\mathbf{c}_{N-1,N}$ given by (11).

**Initialization:** $\Omega^{ref} = \emptyset$.

---

/* ***Part 1****: Determination of sets $S_i$: */*

**for** $i = 1$ to $N$ **do**

$S_i = \{\mathbf{I}_i\}$

**for** $j = 1$ to $N$ **do**

**if** $(|i - j| = 1)$ **then**

Compute $\mathbf{v}_{i,j}$ using (12).

**end if**

**if** $(|i - j| > 1)$ **then**

Compute $\mathbf{v}_{i,j}$ using (13).

**end if**

**if** $(j \neq i)$ and ($\mathbf{I}_j$ $\tau$-overlaps with $\mathbf{I}_i$) **then**

$S_i \leftarrow S_i \cup \mathbf{I}_j$

**end if**

**end for**

**end for**

$G = \{S_1, S_2, \ldots, S_N\}$.

---

/* ***Part 2****: Reference image determination: */*

**while** $G \neq \emptyset$ **do**

- $\Omega^{ref} \leftarrow \Omega^{ref} \cup \mathbf{I}_i$, where $i$ satisfies $|S_i| \geq |S_k|$, for all $S_{k \neq i} \in G$.
- For all images $\mathbf{I}_j \in S_i$, removing corresponding set $S_j$ from $G$:

$$G \leftarrow G \setminus \bigcup_{j:\mathbf{I}_j \in S_i} S_j. \tag{15}$$

**end while**

**Output:** Set $\Omega^{ref}$ of images $\mathbf{I}_i^{ref}$ and their groups $S_i$.

**Part 2: Reference image determination.** Let $\mathbf{I}_i^{ref}$ be a reference image and let $\Omega^{ref}$ ($\Omega^{ref} \subset S$) be the set of reference images maximizing the number $|S_i|$ of $\tau$-overlapped images of set $S_i$ associated with $\mathbf{I}_i^{ref}$. At each iteration of part 2, the algorithm searches for set $S_i$ in $G$ with the highest image number $|S_i|$. Image $\mathbf{I}_i^{ref}$ is added to $\Omega^{ref} \subset S$ and becomes a reference image. Before the next iteration, all image sets $S_j$ corresponding to an image $\mathbf{I}_j \in S_i$ are removed from set $G$. The iterative process ends when set $G$ is empty. After the last iteration, all reference images are gathered in set $\Omega^{ref}$ and image group $S_i$ is known for each $\mathbf{I}_i^{ref}$.

*3.2.3. Accurate point correspondences for a robust SfM step*

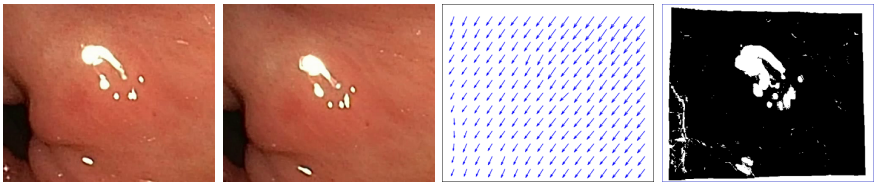
After obtaining the set of reference images $\Omega^{ref} = \{\mathbf{I}_i^{ref}\}_{i \in \hat{Z} \subset Z}$ (where $\hat{Z}$ denotes the index set of the reference images) and the sets $S_i = \{\mathbf{I}_j\}_{j \in Z_i \subset Z}$ with $i \in \hat{Z}$, HP-groups can be easily established based on the DOF fields between images $\mathbf{I}_i^{ref}$ and their $\tau$-overlapped images belonging to the sets $S_i$. Suppose the DOF fields $\mathbf{F}_{i,j}$ between $\mathbf{I}_i^{ref}$ and images $\mathbf{I}_j$ in $S_i$ as determined. For every reference image $\mathbf{I}_i^{ref}$ in $\Omega^{ref}$, one first considers the set $\Xi_i^{ref}$ of regularly distributed 2D points $A_{xy}^{i,ref}$ on $\mathbf{I}_i^{ref}$ given by

$$\Xi_i^{ref} = \left\{ A_{xy}^{i,ref}(xh, yh) \mid x, y \in \mathbb{N}, x \leq \frac{W}{h}, y \leq \frac{H}{h} \right\}, \tag{16}$$

where parameter $h$ represents the distance between neighbor points on a grid visible in Fig. 1(b). Then, from each point $A_{xy}^{i,ref} \in \Xi_i^{ref}$ in $\mathbf{I}_i^{ref}$ which is not indicated as a specular reflection pixel (see mask $M_{SR}$ defined in (2)), one computes its corresponding points in images $\mathbf{I}_j \in S_i$ using the DOF fields $\mathbf{F}_{i,j}$:

$$A_{xy}^{j} = A_{xy}^{i,ref} + \mathbf{F}_{i,j}(A_{xy}^{i,ref}), \forall j \in Z_i. \tag{17}$$

In the proposed SfM pipeline, not only pixels in SR regions (mask $M_{SR}$), but also those in occluded regions are excluded from the homologous point determination. The term "occluded" refers classically to scene parts visible only in one image. For non-occluded pixels, the forward OF from the first image should be the opposite of the backward OF at the corresponding pixels in the second



(a) $\mathbf{I}_i^{ref}$ (b) $\mathbf{I}_j$ (c) OF field (d) $M_{i,j}$

Figure 4: Valid homologous points between image pair $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$. (a) Source image $\mathbf{I}_i^{ref}$. (b) Target image $\mathbf{I}_j$. (c) OF field from $\mathbf{I}_i^{ref}$ to $\mathbf{I}_j$. (d) Mask $M_{i,j}$ indicating specular reflections, occlusions, and inaccurate correspondences, i.e. white pixels indicate "invalid" of too inaccurate homologous points in (c).

image. Moreover, consider two images $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$, every pixels $A_{xy}^{i,ref}$ in $\mathbf{I}_s$ violating at least one of the three constraints:

$$\begin{cases} A_{xy}^{j} = A_{xy}^{i,ref} + \mathbf{F}_{s,t}\left(A_{xy}^{i,ref}\right) \\ A'^{i,ref}_{xy} = A_{xy}^{j} + \mathbf{F}_{t,s}\left(A_{xy}^{j}\right) \\ \|A_{xy}^{i,ref} - A'^{i,ref}_{xy}\|_2 \leq \epsilon \end{cases} \tag{18}$$

is marked as having an inaccurate flow field vector, where a weak $\epsilon$ threshold parameter value ensures an accurate pixel correspondence. Both occluded pixels and pixels with too inaccurate OF vectors are encoded in binary image $M_{inac}$, where $M_{inac}(\mathbf{x}) = 1$ refers either to an occluded pixel (also detected with (18)) or a pixel without a very accurate OF vector (the latter pixels are not necessarily associated with a wrong OF vector, but they simply not lead to correspondences with a high accuracy). Binary mask $M_{i,j}$ defined as $M_{i,j} = M_{SR} \cup M_{inac}$ is used to mark pixels which will be excluded from the homologous point determination of two images $\mathbf{I}_i^{ref}$ and $\mathbf{I}_j$. An example of mask $M_{i,j}$ can bee seen in Fig. 4(d). The flow field obtained using the proposed variational OF method for the textureless image pair $(\mathbf{I}_i^{ref}, \mathbf{I}_j)$ in Figs. 4(a)-4(b) is illustrated in Fig. 4(c). Only the OF vectors corresponding to black pixels which verify $M_{i,j} = 0$ in Fig. 4(d) are used to determine the homologous point sets.

Finally, a HP-group is defined by a point $A_{xy}^{i,ref}$ in $\mathbf{I}_i^{ref}$ and all its homologous points in images $\mathbf{I}_j \in S_i$. It is noticeable that with the proposed method

| | Optical flow | | | | Point grouping | | |
|---|---|---|---|---|---|---|---|
| **Parameter** | $\lambda$ | $Py_s$ | $\gamma_1$ | $\gamma_2$ | $\tau$ | $h$ | $\epsilon$ |
| **Adjusted value** | 9 | 0.7 | 3 | 5 | $\frac{2}{3}WH$ | 10 | 0.1 |

Table 1: Constant parameter values used for experiments with the proposed DOF-based SfM method.

numerous HP-groups consisting each of a large amount of accurately matched points can be established. The number of HP-groups depends on the values of parameter $h$ and of the overlap parameter $\tau$. Moreover, HP-groups are robustly and accurately determined because no optical flow errors accumulate.

## 4. Results and discussion

This section successively quantifies the accuracy of the DOF-based SfM scheme (subsection 4.1), demonstrates the differences between feature and OF approaches when textures are missing (subsection 4.2), and highlight the robustness of the proposed method which can deal with very different scenes and acquisition conditions (subsection 4.3). The code (computation of homologous point groups usable as input by any SfM approach) and all data used in this section are given as supplementary material.

The proposed homologous point grouping method detailed in section 3.2 has three important parameters: overlap threshold $\tau$ in (14), cell size $h$ in (16), and error threshold $\epsilon$ in (18). The optimal values of these parameters are determined using a grid search method performed on the phantom data-sets. The quality criteria given in Section 4.1 were used to find the best values of triplet $(\tau, h, \epsilon)$. The optimal values of all parameters (including the OF parameters in Section 3.1) are summarized in Table 1. These parameter values were held constant for all experiments with the proposed SfM method.

(a) (b)

Figure 5: Phantom data. (a) Snapshot of the external cylinder surface and two small FoV skin images.(b) Snapshot of the internal cylinder surface and two small FoV stomach images.

### *4.1. Accuracy of the DOF-based SfM scheme*

#### *4.1.1. Phantom description and data acquisition*

The two phantoms used for the surface recovery tests consist of half-cylinders with precisely known internal and external diameters and carrying each an orange sphere whose diameter $d_{gt}$ equals 40.14 $mm$ (see Fig. 5). Epithelial tissue images were printed on paper sheets and glued onto the cylinders. Due to the printing process, the epithelial textures are much more contrasted than those visible directly in the images. A camera equipped by a 12 mm focal length objective is used to acquire sequences of images with a size of $780 \times 580$ pixels. As in medical scenes where the acquisition is done close to the epithelial tissue, the camera/phantom surface distance was short so that each image only visualize a small object region. In these experiments, a state-of-the-art SfM method, namely COLMAP [18] which uses SIFT features [4], is placed in ideal conditions to find numerous correspondences. Thus, the accuracy of the proposed DOF-based SfM method can be evaluated through comparison with COLMAP[3]. The parameters of COLMAP were set to the default values as given in [18]. Even published in 2016, COLMAP remains among the reference SfM methods in terms of accuracy and robustness. More recent publications, instead of improving strongly the accuracy and robustness, adapted the SfM principle to be, for

[3]The code can be downloaded at `https://colmap.github.io/`

instance, more suitable for different acquisition devices [39].

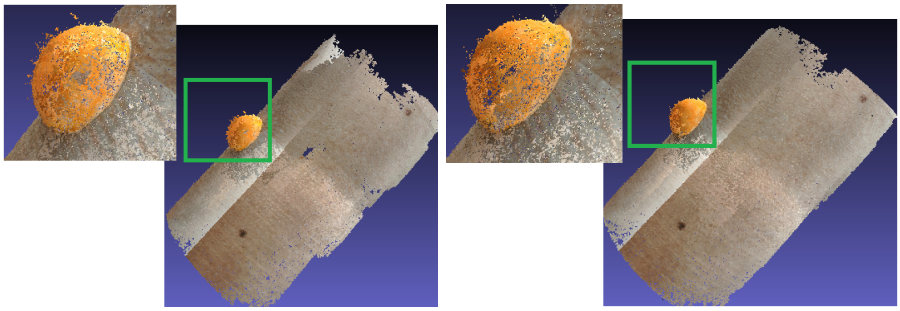
*Skin surface phantom.* The external cylinder surface with diameter $D_{gt} = 159.45mm$ is lined by skin images (see the left image in Fig. 5(a)). This phantom roughly simulates the shape of arm or leg parts. As in dermatology, the epithelium is on an external body surface and the camera which is close to the simulated tissues acquired a sequence of 621 images (two of the latter are shown on the right in Fig. 5(a)).

*Internal hollow organ phantom.* The inner surface of this phantom (see Fig. 5(b)) is covered with paper sheet printings of stomach images acquired during gastroscopies. The internal diameter $D_{gt}$ of the cylinder equals 191.8 $mm$. A sequence of 265 small FoV images was acquired for this phantom.

#### *4.1.2. Phantom reconstruction results*

For the result evaluation, the 3D point clouds are first separated in two independent surface parts, namely the cylinder part and the sphere part. Then, a fitting technique is separately applied to each part to obtain the equations of the reconstructed cylinder and sphere surfaces. For each reconstruction, both the diameters of the cylinder and sphere surfaces (denoted by $D$ and $d$, respectively), and information relating to inlier and outlier points are calculated. A point of a reconstructed cloud is considered as an outlier when its distance to the estimated phantom surface is greater than $0.005.D$ (i.e., 0.5% of the cylinder diameter).

Four criteria are used to evaluate the accuracy of the reconstruction methods:
- The **outlier rate** (in %) corresponds to the ratio of the outlier number over the whole 3D point number of the cloud.
- The **mean outlier error** (in $mm$) gives the mean distance between outlier points and the fitted phantom surface.
- The **3D phantom shape accuracy** is assessed by comparing the diameter ratio $D/d$ of the reconstructed cylinder and sphere surfaces with their ground truth $D_{gt}/d_{gt}$. This criterion is defined by (19) in which $p = 100\%$ and $p = 0\%$



(a) COLMAP (SfM + MVS) (b) Proposed DOF-based SfM

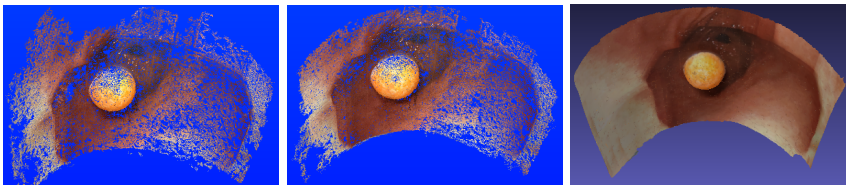
Figure 6: SfM results obtained for the skin phantom shown in Fig. 5(a). (a) 3D point cloud obtained with COLMAP [18] and zoom on the sphere region delineated by the green rectangle. (b) 3D point cloud obtained with the proposed method and zoom on the sphere.

indicate a perfect and a completely wrong shape, respectively.

$$p = \left(1 - \frac{|D_{gt}/d_{gt} - D/d|}{D_{gt}/d_{gt}}\right) \times 100\%. \tag{19}$$

- For COLMAP, the **computation time** criterion includes the total time of the SfM and MVS parts, while for the proposed method it only corresponds to the SfM part (it is recalled that COLMAP provides dense point clouds using the SfM step followed by the MVS step, whereas the proposed SfM method leads directly to dense point clouds). Experiments were performed on a HP Pavillion laptop with an Intel Core i5 1.60GHz and 16GB RAM and NVIDIA GeForce 940MX GPU.

*Reconstruction accuracy for the skin surface phantom.* For this test, the cloud computed by the proposed DOF-based SfM method includes 558397 3D points (see Fig. 6(b)). Among them, 20010 points are outliers (the outlier rate is 3.58%). The point cloud of COLMAP consists of 591126 points (see Fig. 6(a)) and the number of outlier points is 25359 (the outlier rate is 4.29%). The mean outlier error of the proposed SfM method is $7.5mm$, while the same value is $6.92mm$ for COLMAP. $D_{gt}/d_{gt} = 3.972$ is the ground truth of this phantom. The diameter ratios obtained with COLMAP and with the proposed DOF-based SfM method are 3.9462 ($p = 99.35\%$) and 3.945 ($p = 99.32\%$), respectively.



(a) COLMAP (SfM+MVS) (b) DOF-based SfM (c) Textured surface

Figure 7: Internal phantom reconstruction results (same viewpoint as in 5(b)). (a) Point cloud given by COLMAP [18]. (b) Point cloud obtained with the proposed SfM method. (c) Textured triangle mesh obtained with the cloud in (b).

The computation time of the DOF-based SfM method is 147.8 minutes, while COLMAP requested 150.5 minutes (115.5 minutes for the SfM-part and 35 minutes for the MVS-part) to reconstruct the surface.

It is visible that, compared to the COLMAP feature-based method, the outlier rate is significantly smaller for the proposed SfM method. The computation time and the ability of the object shape preservation are nearly the same for both methods. A visual comparison of Figs. 6(a) and 6(b) shows that the point cloud obtained with the proposed method covers a greater cylinder surface than that of COLMAP (see the cylinder borders).

*Reconstruction accuracy for the internal hollow organ wall phantom.* For this phantom, the proposed method constructed a cloud of 233639 points in 79.15 minutes. COLMAP requested 44.1 minutes for computing a cloud of 173262 points. The outlier rate and the mean outlier error of the proposed method are 5.17% and 5.65 $mm$, respectively. For the COLMAP approach, these values are 6.84% and 8.298$mm$. The ground truth ratio $D_{gt}/d_{gt}$ equals 4.78 for this phantom. The diameter ratio obtained with COLMAP is $D/d = 4.87$ ($p = 98.29\%$), while $D/d = 4.72$ ($p = 98.76\%$) for the proposed method. The point cloud computed by the proposed SfM method (see Fig. 7(b)) is much denser than that of COLMAP (see Fig. 7(a)), while the outlier rate and the mean distance error are slightly lower than those of COLMAP.

Globally, the results obtained with the two phantoms highlight the accuracy

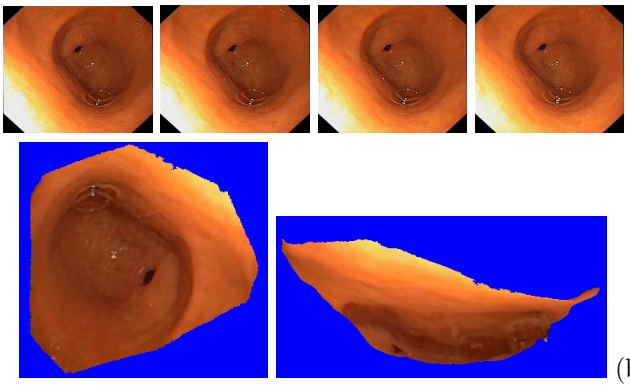


Figure 8: Robustness tests on gastroscopic data.(a) Images acquired from four viewpoints and used to show the impact of the homologous point grouping efficiency on the robustness of a SfM method. (b) Two viewpoints of the pyloric antrum region surface reconstructed with the DOF-based SfM method.

of the proposed method since it is quite similar to that of COLMAP which has a high precision in presence of contrasted textures (the diameter ratios obtained by the two SfM methods are very close). However, reconstructing accurately surfaces with an SfM technique based on a DOF is a first important result (such accuracy obtained with a DOF-based SfM scheme is an original result).

### *4.2. Robustness of the DOF-based SfM scheme*

Four gastroscopic images of size $640 \times 482$ pixels (image set $\{\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4\}$, see Fig. 8(a)) are used to illustrate the importance of a robust homologous point group determination in the frame of SfM applied to medical scenes. Both the proposed and the COLMAP SfM methods were applied to these images. It can be seen in Fig. 8(b) that the DOF-based approach led to a realistic pyloric antrum shape, while COLMAP was unable to reconstruct a surface.

| Image pair | $(\mathbf{I}_2^{ref}, \mathbf{I}_1)$ | $(\mathbf{I}_2^{ref}, \mathbf{I}_3)$ | $(\mathbf{I}_2^{ref}, \mathbf{I}_4)$ |
|---|---|---|---|
| SIFT matches | 33 | 23 | 18 |
| OF correspondences | 410 | 406 | 404 |

Table 2: Number of correspondences determined between each image pair using SIFT features and DOF.

| Image groups | $\mathbf{I}_1$, $\mathbf{I}_2^{ref}$, $\mathbf{I}_3$ | $\mathbf{I}_2^{ref}$, $\mathbf{I}_3$, $\mathbf{I}_4$ | $\mathbf{I}_2^{ref}$, $\mathbf{I}_3$, $\mathbf{I}_4$ | $\mathbf{I}_1$, $\mathbf{I}_2^{ref}$, $\mathbf{I}_3$, $\mathbf{I}_4$ |
|---|---|---|---|---|
| OF | 400 triplets | 389 triplets | 372 triplets | 368 quadruplets |
| SIFT | 12 triplets | 10 triplets | 14 triplets | 7 quadruplets |

Table 3: Number of homologous point triplets and quadruplets obtained with the SIFT features and DOF matches in Table 2.

These differences in terms of reconstruction performances can be explained by the number of homologous points obtained with the SIFT and DOF approaches. Table 2 gives the number of correct homologous points obtained by COLMAP and the proposed SFM scheme for three image pairs. For COLMAP, only few tens of homologous points were found on these image pairs, while more than 400 correspondences were established using the DOF method.

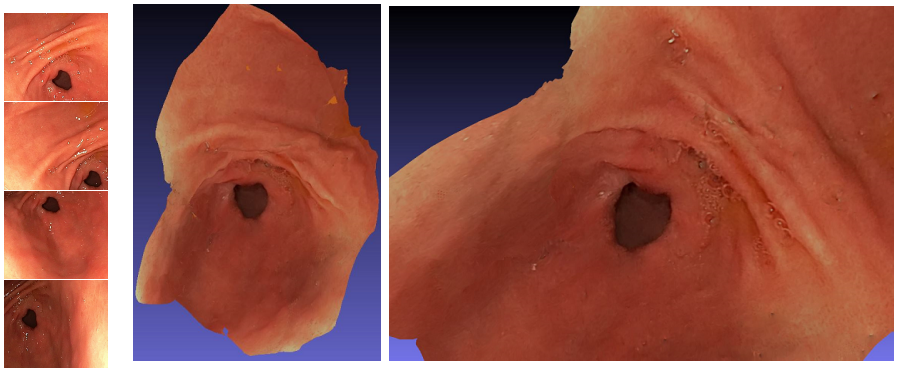
As seen in Table 3 numerous point triplets and 368 point quadruplets obtained with the DOF fields enable to construct a part of the pyloric antrum region (stomach). On the contrary, only very few homologous point triplets and quadruplets were obtained with SIFT features. For this reason the COLMAP software completely failed in the reconstruction of the pyloric antrum region.

### *4.3. Tests on various medical scenes*

In endoscopy, video-sequences are often not archived because they enable only an easy diagnosis during the examination itself. Video-sequences are also not optimal for lesion evolution assessment between two examinations since their visual comparison is very difficult. The proposed DOF-based SfM method can improve the efficiency of such medical examinations.

#### *4.3.1. 3D mosaicing of the pyloric antrum in gastroscopy*

The inner stomach wall is scanned by an endoscope to find lesions like cancers or inflammations (chronic inflammations in the pylorc antrum region often lead to cancers). The reference white light (WL) color modality is classically complemented by the narrow band imaging (NBI) modality in which inflammations can be earlier detected as in WL. A gastroscope often allow for switching between



(a) (b)

Figure 9: Pyloric antrum region surface obtained with the DOF-based SfM method in the white light modality. (a) Four colour images of the sequence of 191 frames. (b) Pyloric antrum region under to viewpoints.

the two modalities so that both WL and NBI video-sequences can be acquired. Constructing FoV extended WL and NBI surfaces of the pyloric antrum has several advantages: *i)* it will lead to an original and medically interesting way to document a gastroscopic examination, *ii)* it will be a new information exchange media between gastroenterologists and other specialists, and *iii)* the comparison of two 3D mosaics built with video-sequences acquired at some weeks or month intervals will allow for an inflammation or cancer follow-up.

Fig. 9.(a) shows four images among 191 frames of a WL video-sequence. No textures and only few structures are visible in these images which are classically affected by reflections. The lack of textures impedes feature-based approaches to establish numerous correspondences, while the reflections lead to wrong matches. The pyloric antrum region surface constructed with the DOF-based SfM approach is presented under two viewpoints (i.e., at different orientations and scales) in Fig. 9.(b). With such surfaces, gastroenterologists are able to virtually navigate into the stomach after the examination. It is also noticeable that most of the reflections do not appear in the 3D mosaics since these illumination effects are not systematically present for all viewpoints on a same 3D point (viewpoints without reflections can be chosen during the surface

texturing).

A 3D mosaicing example of a sequence of 214 images acquired in the pyloric antrum in the NBI modality was also reconstructed [4].

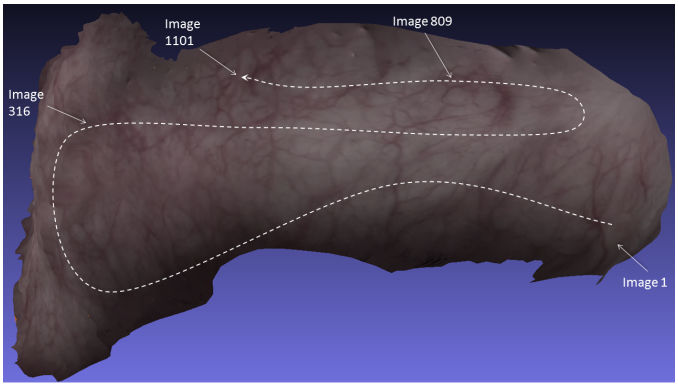
*4.3.2. 3D bladder wall mosaicing*

In cystoscopy, urologists scan the bladder wall with the endoscope's distal tip maintained close to the epithelial tissue. Urologists have to mentally reconstruct bladder parts in order to be able to know the position and orientation of the cystoscope into the organ. To do so, the endoscopist comes regularly back to landmarks like the meatus of the urethra or the ureters. Locating the exact position of the instrument in the bladder in this way is only possible during the cystoscopic examination. This is one reason why cystoscopic video sequences are usually not archived to document an examination. Building 2D or 3D mosaics of large bladder parts is a solution to archive important information for examination traceability and lesion evolution assessment. Moreover, controlling exactly the trajectory of the cystoscope is very difficult. Consequently, it is never obvious whether a region of interest (e.g., with potential lesions) was completely scanned or not. Constructing large FoV mosaics is also a solution for this issue since non scanned regions are visible when a cartography of organ parts can be robustly done.

Fig. 10(a) shows a 2D mosaic constructed with a sequence of 1101 images acquired in the WL modality. This mosaic, obtained with a robust 2D image registration technique [40], has several drawbacks. On the one hand, all images are placed in a 2D mosaicing plane defined in this example by "image 1" acting as reference. The mosaic path is materialized by the white curve, the pixels of image $I_i$ being added to the current mosaic built with information of images $I_1$ to $I_{i-1}$. As visible in Fig. 10(a), the first ($I_1$) and last ($I_{1101}$) images correspond to ellipses with very different areas in pixels. The resolution of an image in the mosaic plane depends strongly on the viewpoint changes of the endoscope

[4] See the video available at `https://github.com/CRAN-BioSiS-Imaging/PR2020`


(a)




(b)

Figure 10: Same bladder surface part represented by a 2D and a 3D mosaic. (a) 2D mosaic constructed with 1101 images. (c) 3D mosaic showing the bladder wall under the viewpoint of the 2D mosaic in (b).

between images, leading to both strong image distortion and significant resolution losses (only image $I_1$ is without resolution loss). On the other hand, due to accumulating registration errors, images which should be overlapped are in different (non-overlapping) places in the 2D mosaic plane. Thus, images $I_{809}$ to $I_{1101}$ should partly overlap the previous images of the sequence. The gaps without bladder texture in the 2D map are not due to tissue areas which were not scanned by the endoscope, but to an accumulating registration error that grows during the map construction. These registration errors are generally difficult to

correct, even with sophisticated techniques as described in [16]. In 2D mosaics it is not possible to distinguish between mosaicing errors and tissues that were not scanned by the endoscope.

Fig. 10(b) shows the bladder surface reconstructed using the proposed DOF-based SfM method applied to the sequence of 1101 images. The orientation of the curved surface (the 3D shape is not very perceptible from this viewpoint) was chosen so that the 3D mosaic content can be visually compared to that of the 2D mosaic in Fig. 10(a). It can be seen that the surface is without gaps and that the bladder part was actually completely scanned by the endoscope. The dashed white line approximately represents the endoscope trajectory position computed for the 2D mosaic (this trajectory is not computed in the proposed SfM method). Also, contrary to Fig. 10(a) where the resolutions is strongly affected by the uncontrolled viewpoint changes of the camera, the resolution onto the surface is by far more constant since the image textures are projected on a realistic bladder shape surface part. The 3D mosaic can be archived and used as information media exchange and patient follow-up.

In urology, the fluorescence modality is a complementary modality often used to detect some bladder cancers in an early stage (with some cystoscopes it is possible to switch between the WL and fluorescence modalities). A 3D bladder surface was constructed with 84 fluorescence images [5].

#### *4.3.3. 3D skin mosaicing in dermatology*

In dermatology, lesions like pressure ulcers or cancers have to be acquired with a high resolution. Pressure ulcers are usually lesions that are widespread over several images. Besides the fact that extended FoV images are required to represent them with a high resolution, it is also important to assess the lesion area evolution between two examinations. This evolution assessment is more precise on a 3D surface than on a 2D mosaic. Moreover, in countries like France, there is a lack of dermatologists in the countryside. A nurse often

[5] See the video available at `https://github.com/CRAN-BioSiS-Imaging/PR2020`

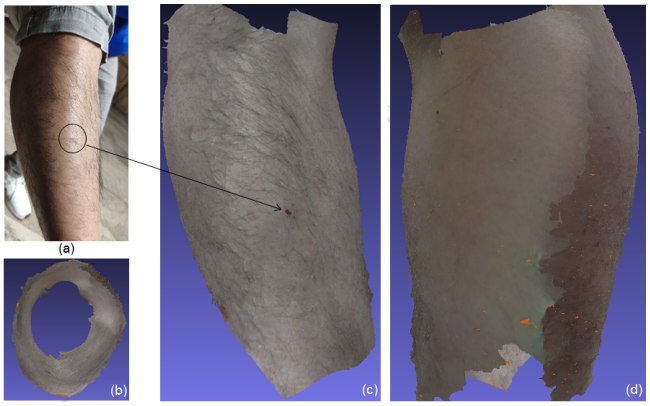



Figure 11: Leg reconstruction. (a) Snapshot of the leg. (b), (c), (d) Surface under three viewpoints.

takes a few images of a wound at the patient's home and transmits them to a dermatologist in the city. An alternative would be to acquire a video-sequence of the interesting skin part and to transmit the data before or after the 3D surface construction. This would also allow for a virtual navigation around the body part under consideration without the presence of the patient.

Fig. 11 shows the surface construction of a part of a leg seen in the snapshot given in Fig. 11(a). The circle in Fig. 11(a) encompasses a small wound which is clearly visible in the top view of the leg given in Fig. 11(b). In the bottom view of the leg (see Fig. 11(d)), the vertical light gradient is due to the camera viewpoint differences between the first and last images (the number of the last image is 130), which close the loop trajectory required for 360 degree scan of the leg (see the cross-sectional view of the leg in Fig.11(b)). This colour differences, which do not affect the robustness of the proposed SfM technique owing to the used illumination invariant OF method, can be corrected.

## 5. Conclusion: global discussion and perspectives

An algorithm can be considered as being robust when it provides appropriate results for very different scene contents and acquisition conditions. In this paper, surface construction tests were presented on very different data. Surfaces

with almost no textures (gastroscopy), with rather few textures (cystoscopy) and with more textures (dermatology) were successfully reconstructed with the proposed DOF-based SfM method. These surfaces were reconstructed for hardly controllable camera trajectories and under strongly varying illumination conditions. Moreover, surface construction tests were conducted for very different imaging modalities, i.e. WL for all medical applications, NBI in gastroscopy and fluorescence in cystoscopy. All surface construction tests were performed with the constant algorithm parameters given in this paper. For the tested medical applications the described 3D reconstruction algorithm led systematically to consistent 3D shapes (in accordance with the anatomy of the organ), without discontinuities of textures or structures, as well as with an acceptable resolution regardless of the location observed on the surface. The robustness of the proposed solution is not only related to the fact that the OF-method provides numerous homologous points even without contrasted textures and structures. This robustness is also due to the image grouping algorithm which allows for a 3D point reconstruction using a large number of homologous points seen in images taken under different viewpoints.

In the medical context of this work, the purpose of the proposed SfM method was not to construct very precise surfaces because hollow organs never have the same shape between two examinations or from one patient to another. However, the results on phantoms show that the precision of the proposed DOF-based SfM method can closely approach that of a method based on the detection of features (as with SIFT). The less accurate (non subpixel accuracy) homologous point matching with OF methods is compensated by the numerous correspondences provided by the flow field (the matched points are by far more numerous than those obtained with feature methods working with a subpixel accuracy). Although the accuracy of the proposed DOF-based SfM can only approach that of state-of-the-art SfM methods when features can be detected, its robustness make it potentially interesting for other (non-medical) scenes with few textures. Further test will be made to assess precisely the appropriateness of the method to other scenes.

Apart from the strengths mentioned above, the proposed method still has limitations that can be corrected. In endoscopy, the quality of numerous images is affected by defocussing/refocussing, motion blur, and floating objects, etc. The proposed SfM algorithm should be associated to a more complete pre-processing step to improve image selection [41]. Moreover, constructing the 3D surfaces in less than an hour allows for a second more reliable diagnosis after the medical examination, patient follow-up, examination traceability and information exchange between various specialists. However, this surface can currently not be available during the examination itself. The 3D reconstruction can be speeded-up from the algorithmic point of view (e.g., when available, combining feature information and OF information, see the very preliminary work in [42]) and from the informatics point of view (code optimizing and parallelization).

In this work, the medical scenes were almost rigid: the pyloric antrum region presents rather moderate surface deformations, the bladder is filled with an isotonic solution which rigidifies the organ surface, and the skin surface can be considered as completely rigid. A natural extension of the proposed reconstruction algorithm would be to adapt it to non-rigid scenes. Non-rigid structure from motion (NRSfM, [43]) could be an appropriate principle for the development of a 3D reconstruction pipeline of medical scenes, for instance for facilitating the diagnosis of the Barret's esophagus.

**Acknowledgments**

This work was partially funded by the Agence Nationale de la Recherche in the frame of the EMMIE (Endoscopie MultiModale pour les lésions Inflammatoires de l'Estomac) project (ANR-15-CE17-0015).

**References**

[1] Z. Zhang, A flexible new technique for camera calibration, IEEE Trans. Pattern Analasis Machine Intelligence 22 (11) (2000) 1330–1334.

[2] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, 2nd Edition, Cambridge University Press, New York, NY, USA, 2004.

[3] S. Agarwal, N. Snavely, S. M. Seitz, R. Szeliski, Bundle adjustment in the large, in: ECCV, Part II, 2010, pp. 29–42.

[4] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[5] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comp. Vision and Image Understanding 110 (3) (2008) 346–359.

[6] Y. Furukawa, J. Ponce, Accurate, dense, and robust multiview stereopsis, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1362–1376.

[7] M. Jancosek, T. Pajdla, Multi-view reconstruction preserving weakly-supported surfaces, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011, pp. 3121–3128.

[8] J. Schönberger, E. Zheng, J. M. Frahm, M. Pollefeys, Pixelwise view selection for unstructured multi-view stereo, in: European Conference on Computer Vision (ECCV), 2016, pp. 501–518.

[9] M. M. Kazhdan, M. Bolitho, H. Hoppe, Poisson surface reconstruction, in: Eurographics Symposium on Geometry Processing, 2006, pp. 61–70.

[10] H. Vu, P. Labatut, J. Pons, R. Keriven, High accuracy and visibility-consistent dense multiview stereo, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 889–901.

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

[12] M. James, S. Robson, Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application, Journal of Geophysical Research 117 (F03017) (2012) 1–17.

[13] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebni, M. Pollefeys, Building rome on a cloudless day, in: ECCV, Vol. 6314, LNCS, 2010, pp. 368–381.

[14] D. J. Crandall, A. Owens, N. Snavely, D. P. Huttenlocher, SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion, IEEE Trans. on Pat. Anal. and Mach. Intell. 35 (12) (2013) 2841–2853.

[15] A. Behrens, T. Stehle, S. Gross, T. Aach, Local and global panoramic imaging for fluorescence bladder endoscopy, in: Int. Conf. of the IEEE Engineering in Medicine and Biology Society, 2009, pp. 6990–6993.

[16] T. Weibel, C. Daul, D. Wolf, R. Rösch, F. Guillemin, Graph based construction of textured large field of view mosaics for bladder cancer diagnosis, Pattern Recognition 45 (12) (2012) 4138–4150.

[17] M. A. Fischler, R. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.

[18] J. L. Schönberger, J. M. Frahm, Structure-from-motion revisited, in: CVPR, 2016, pp. 4104–4113.

[19] N. Shevchenko, J. Fallert, H. Stepp, H. Sahli, A. Karl, T. Lueth, A high resolution bladder wall map: Feasibility study, in: 34th Int. Conf. of the IEEE Engineering in Medicine and Biology Society, 2012, pp. 5761–5764.

[20] A. Ben-Hamadou, C. Daul, C. Soussen, Construction of extended 3D field of views of the internal bladder wall surface: a proof of concept, 3D Research 7 (3) (2016) 95:1–95:23.

[21] C. Wu, S. Narasimhan, B. Jaramaz, A multi-image shape-from-shading framework for near-lighting perspective endoscopes, International Journal of Computer Vision 86 (2-3) (2010) 211–228.

[22] A. E. Kaufman, J. Wang, 3D surface reconstruction from endoscopic videos, in: Visualization in Medicine and Life Sciences, Springer, 2008, pp. 61–74.

[23] T. Zhao, Q.and Price, S. Pizer, M. Niethammer, R. Alterovitz, J. Rosenman, The endoscopogram: A 3D model reconstructed from endoscopic

video frames, in: Medical Image Computing and Computer-Assisted Intervention(MICCAI), Vol. 9900, LNCS, 2016, pp. 439–447.

[24] O. G. Grasa, E. Bernal, S. Casado, I. Gil, J. M. M. Montiel, Visual SLAM for handheld monocular endoscope, IEEE Trans. Medical Imaging 33 (1) (2014) 135–146.

[25] T. Soper, M. Porter, E. J. Seibel, Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance, IEEE Transactions on Biomedical Engineering 59 (6) (2012) 1670–1680.

[26] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, A. K. E. Bowden, 3D reconstruction of cystoscopy videos for comprehensive bladder records, Biomedical Optics Express 8 (4) (2017) 2106–2123.

[27] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet2.0: Evolution of optical flow estimation with deep networks, in: CVPR, 2017, pp. 2462–2470.

[28] C. Bailer, B. Taetz, D. Stricker, Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation, in: ICCV, 2015, pp. 4015–4023.

[29] Y. Hu, R. Song, Y. Li, Efficient coarse-to-fine patchmatch for large displacement optical flow, in: CVPR, 2016, pp. 5704–5712.

[30] D.-H. Trinh, C. Daul, On illumination-invariant variational optical flow for weakly textured scenes, Comput. Vis. and Image Under. 179 (2019) 1–18.

[31] L. Álvarez, J. Weickert, J. S. Pérez, Reliable estimation of dense optical flow fields with large displacements, Int. J. of Comput. Vis. 39 (1) (2000) 41–56.

[32] A. Bruhn, J. Weickert, C. Schnörr, Lucas/kanade meets horn/schunck: Combining local and global optic flow methods, International Journal of Computer Vision 61 (3) (2005) 211–231.

[33] D. Sun, S. Roth, M. J. Black, Secrets of optical flow estimation and their principles, in: CVPR, 2010, pp. 2432–2439.

[34] D.-H. Trinh, C. Daul, W. Blondel, D. Lamarque, Mosaicing of images with few textures and strong illumination changes: Application to gastroscopic scenes, in: IEEE Int. Conf. on Image Processing, 2018, pp. 1263–1267.

[35] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: ECCV, 2004, pp. 25–36.

[36] K. Yoon, I. Kweon, Adaptive support-weight approach for correspondence search, IEEE Trans. on Pat. Anal. and Mach. Intell. 28 (4) (2006) 650–656.

[37] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, Journal of Mathematical Imaging and Vision 40 (1) (2011) 120–145.

[38] M. Drulea, S. Nedevschi, Motion estimation using the correlation transform, IEEE Trans. Image Processing 22 (8) (2013) 3260–3270.

[39] S. Nousias, M. I. A. Lourakis, C. Bergeles, Large-scale, metric structure from motion for unordered light fields, in: CVPR, 2019, pp. 3292–3301.

[40] S. Ali, C. Daul, E. Galbrun, F. Guillemin, W. Blondel, Anisotropic motion estimation on edge preserving riesz wavelets for robust video mosaicing, Pattern Recognition 51 (2016) 425–442.

[41] S. Ali, F. Zhou, B. Braden, et al., An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Science Reports 60 (2748). doi:doi.org/10.1038/s41598-020-59413-5.

[42] T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, C. Daul, 3D surface reconstruction using dense optical flow combined to feature matching: Application to endoscopy, in: Colloque GRETSI, Lille, France, 2019.

[43] S. Kumar, A. Cherian, Y. Dai, , H. Li, Scalable dense non-rigid structure-from-motion: A grassmannian perspective, in: CVPR, 2018, pp. 254–263.