Computers in Biology and Medicine 146 (2022) 105616

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Scene-graph-driven semantic feature matching for monocular digestive endoscopy

Zhuoyue Yang [a], Junjun Pan [a,b,c,*], Ranyang Li [a], Hong Qin [d]

[a] *State Key Laboratory of Virtual Reality Technology and Systems,Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, China*
[b] *Peng Cheng Lab, 2 Xingke 1st Street, Nanshan District, Shenzhen, 518000, China*
[c] *Faculty of Media and Communication, Bournemouth University, Bournemouth, UK*
[d] *Department of Computer Science, Stony Brook University, New York, USA*

ARTICLE INFO

Keywords:
Semantic description
Scene graph
Deformation-aware metric design
Dense feature correspondence
Structure from motion

ABSTRACT

Background and objective: Registration of the preoperative 3D model with the video of the digestive tract is the key task in endoscopy surgical navigation. Accurate 3D reconstruction of soft tissue surfaces is essential to complete registration. However, existing feature matching methods still fall short of desirable performance, due to the soft tissue deformation and smooth but less-textured surface.

Methods: In this paper, we present a new semantic description based on the scene graph to integrate contour features and SIFT features. Firstly, we construct the semantic feature descriptor using the SIFT features and dense points in the contour regions to obtain more dense point feature matching. Secondly, we design a clustering algorithm based on the proposed semantic feature descriptor. Finally, we apply the semantic description to the structure from motion (SfM) reconstruction framework.

Results: Our techniques are validated by the phantom tests and real surgery videos. We compare our approaches with other typical methods in contour extraction, feature matching, and SfM reconstruction. On average, the feature matching accuracy reaches 75.6% and improves 16.6% in pose estimation. In addition, 39.8% of sparse points are increased in SfM results, and 35.31% more valid points are obtained for the DenseDescriptorNet training in 3D reconstruction.

Conclusions: The new semantic feature description has the potential to reveal more accurate and dense feature correspondence and provides local semantic information in feature matching. Our experiments on the clinical dataset demonstrate the effectiveness and robustness of the novel approach.

1. Introduction

For patients with gastrointestinal diseases, digestive endoscopy is still the most effective way of diagnosis and treatment. However, digestive endoscopic surgery has strict requirements for surgeons' experience and skills, due to the narrow field of view and lack of depth perception. In recent years, with the rapid development of VR/AR techniques, an increasing number of researchers choose AR based surgical navigation technology to solve the difficulties mentioned above. Technically, the registration of the preoperative 3D model with the endoscopic video of the digestive tract is the key task. And surgical scene reconstruction using a monocular endoscope is the first step to complete the registration. Technically, one effective way to handle the 3D reconstruction based on monocular video is the structure from motion (SfM) [25,41]. In natural scenes, SfM methods have been developed in recent years with a number of theories and applications to tackle the problem of 3D reconstruction of rigid objects under constant illumination [35]. However, in the endoscopy surgical scene, two main issues affect SfM reconstruction performance. One is the soft tissue deformation, which violates the static scene assumption. The other is the smooth and repetitive soft tissue texture, which usually results in sparse features and wrong feature matching. At present, the state-of-the-art work [23] used self-supervised information provided by SfM to train the network and tried to solve the difficulty of feature matching in endoscopic surgery. However, if SfM fails to provide reasonable results, especially when tissue deformation occurs, it may not work correctly. Moreover, it

* Corresponding author. State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, China.

E-mail address: pan_junjun@buaa.edu.cn (J. Pan).

https://doi.org/10.1016/j.compbiomed.2022.105616
Received 4 February 2022; Received in revised form 11 April 2022; Accepted 11 May 2022
Available online 16 May 2022

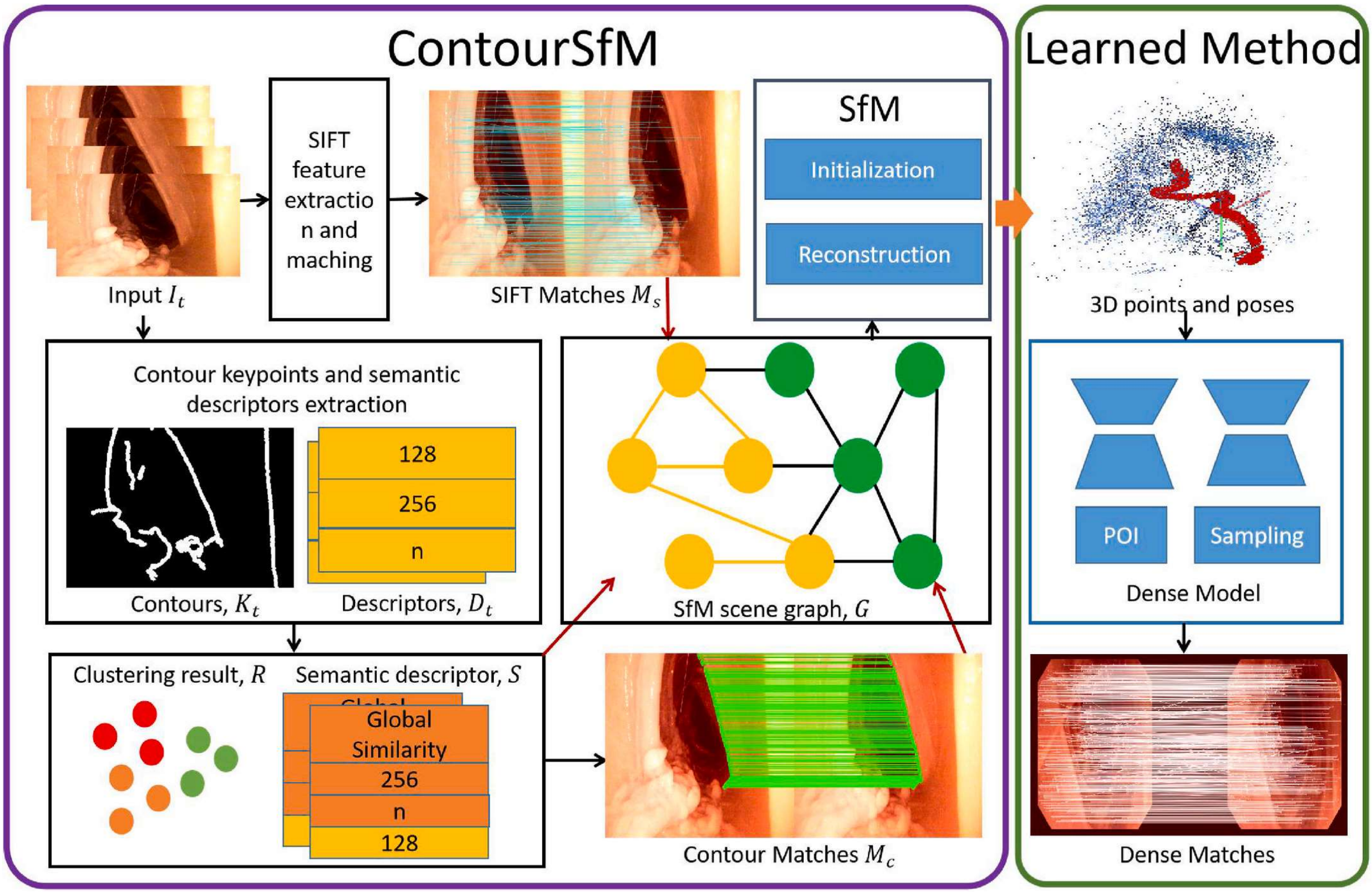


Fig. 1. The framework of the proposed method. Each node in the semantic scene graph contains 2 features. Different color lines between nodes represent different feature matching correspondences generated by SIFT, contour points.

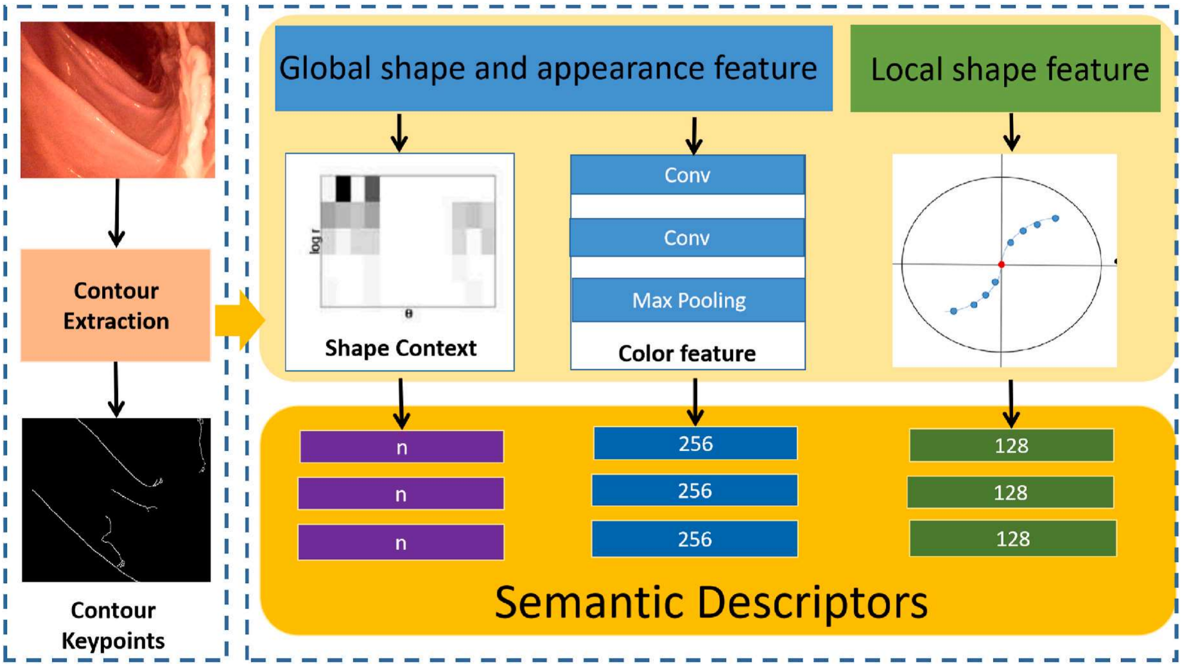


Fig. 2. The pipeline of the dense semantic descriptors extraction.

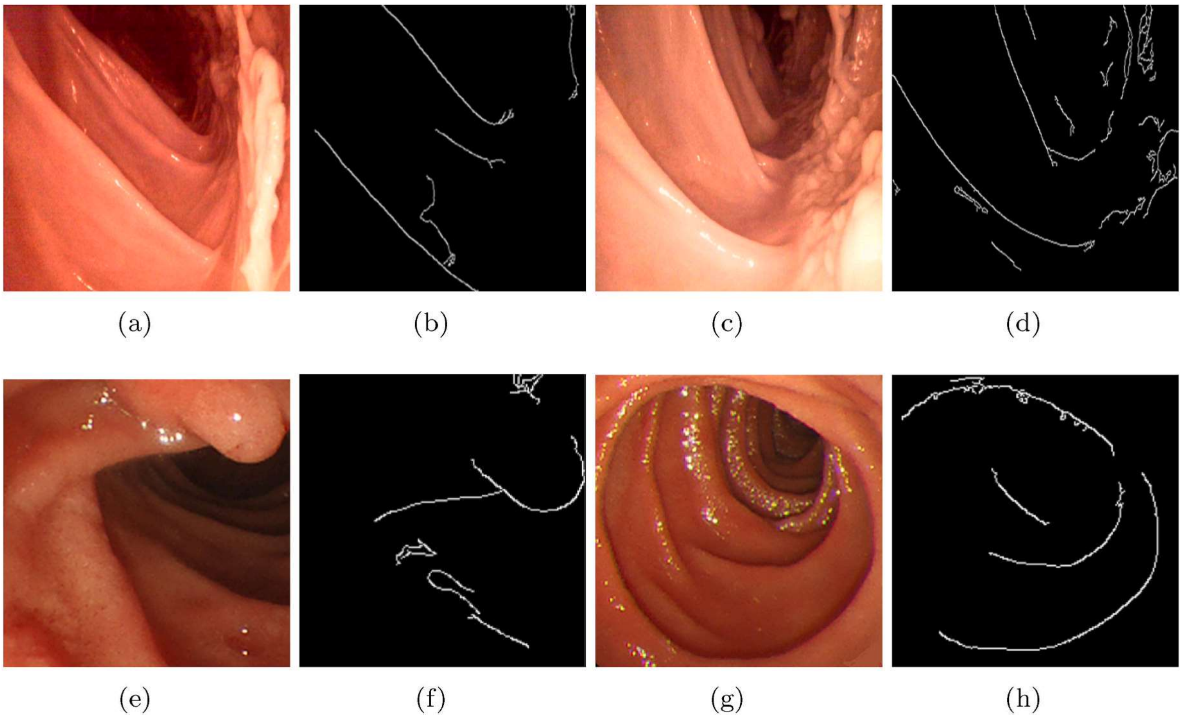


Fig. 3. Illustrations of the result of contour extraction in endoscopy images. (a) (c) (e) and (g) are the original images, and the others are the extracted contours. (a) and (c) are from the phantom dataset. (e) and (g) are from the real dataset.

is difficult to generate a large number of stable 3D points in traditional SfMs due to highlight reflection and soft tissue deformation.

In this paper, we present a new semantic description based on the scene graph to explore new SfM methods for endoscopy. Firstly, a semantic feature descriptor, which combines SIFT feature and contour feature, is designed to obtain denser matching. The contours of the soft tissue seldomly change in different lighting conditions. So contour feature matching is more robust, less affected by highlights, and has fewer noise points compared with the traditional method. Secondly, we design a clustering algorithm based on the proposed semantic feature descriptor. Then we utilize the scene graph to manage the feature matching and construct the semantic description. It increases the number of stable feature points and improves the accuracy of pose estimation in SfM with a denser 3D point cloud. Finally, the point cloud and the pose are used as the input of a deep neural network and provide more accurate self-supervised information. The innovative contributions are summarized as follows:

- We design a semantic feature description by combining SIFT feature, contour feature through the scene graph construction. Dense features result from stable dense points near contours and the semantic feature description emphasizes the local relationship among contours, which provides critical semantic information for better reconstruction.
- We propose a clustering algorithm based on semantic feature description to speed up the initialization of structure from motion. Meanwhile, the clustering results can guide the priority reconstruction of the physiological structure of interest. It provides more accurate data sets for the training by DenseDescriptorNet afterward.
- We implement a new SfM framework with semantic descriptors. It simplifies the complex soft tissue deformation into a multitude of localized rigid reconstruction sub-tasks, which improve the reconstruction performance. Experiments confirm that it can obtain denser point clouds with more accurate pose estimation.

2. Related works

2.1. Point and line features

Feature matching is a key technology in image reconstruction. The most popular point features include SIFT [6], SURF [1], ORB [34], and some of their variants. However, in endoscopic images, only a few matching relationships could be extracted using the above features due to the existence of highlight and noise. Basically, dense feature correspondences can enhance the density of sparse reconstruction and the accuracy of camera trajectory estimation [22]. In recent years, researchers started to shift their foci on dense point features based on learning methods, for instance, a local image descriptor named DAISY [39], universal correspondence network (UCN) [3] and a self-super-vised interest point detection and description (SuperPoint) [7]. However, the lack of large labeled data sets is a critical challenge.

At present, the state-of-the-art work [23] uses a self-monitoring strategy to train the network and tries to solve the difficulty of feature matching in endoscopic surgery. Self-supervised information refers to the point cloud results and camera poses obtained by SfM. The sparse point cloud results from SfM is reprojected to the image planes and is used as sparse feature matching ground truth. The accuracy of this method depends on the correct feature matching and camera poses. If SfM fails to provide reasonable results, especially when tissue deformation occurs, the latest learning-based surface reconstruction method may not work correctly [23].

There are also many methods to extract line features from natural scenes, for example, edge lines (EDlines) [5] and line segment detector (LSD) [13]. Most of the line segments extracted directly from the endoscopic images are seldom due to the highlight. These small line segments can neither accurately represent the overall contour nor exist stably. In addition, there are some methods to extract and match contours in natural scenes, such as [14,21]. The contour of soft tissue is too complex to be described by simplified mathematical symbols. Fortunately, a chain code method is designed to encode any geometric structure. Through this method, the coding operation and basic

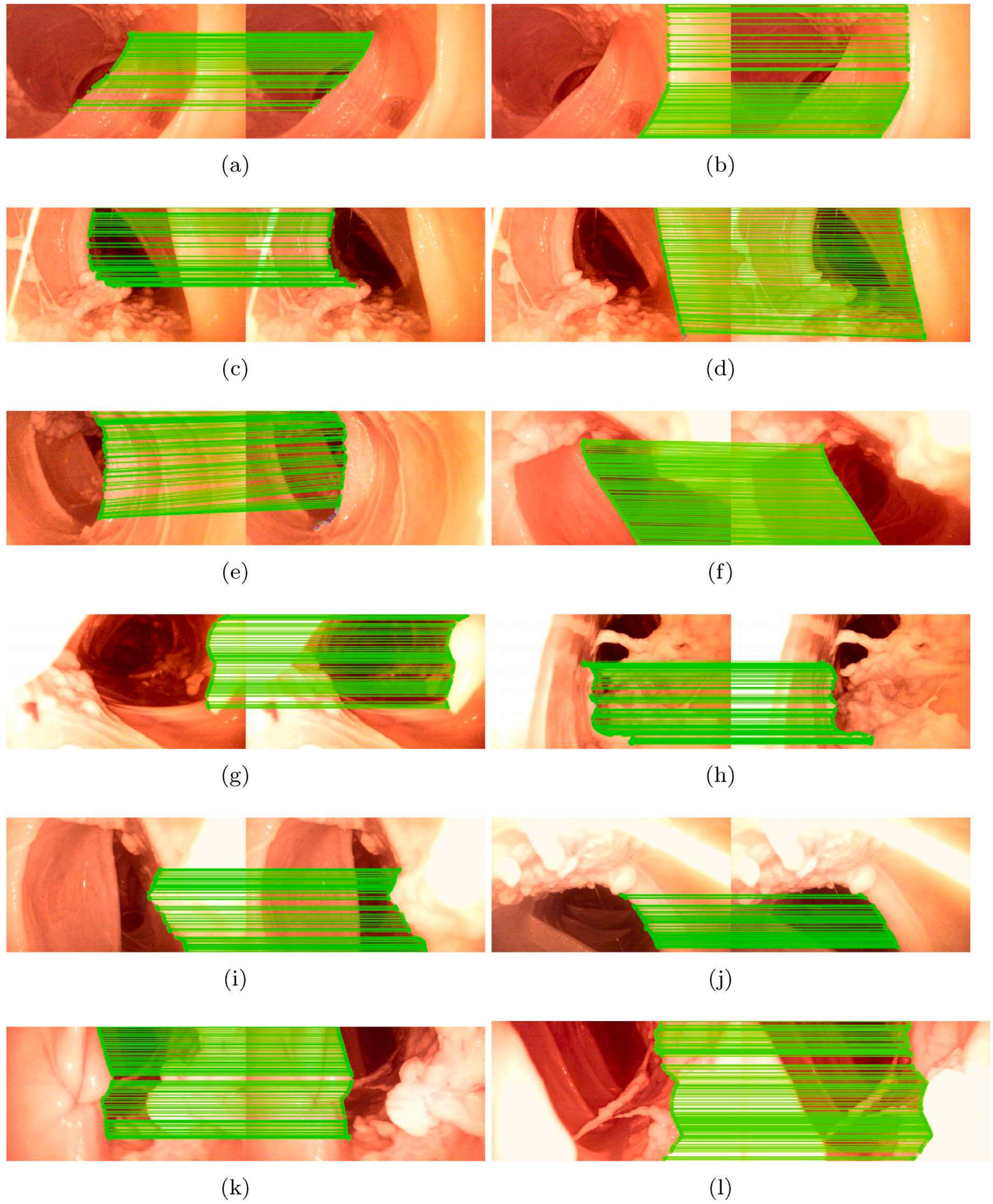
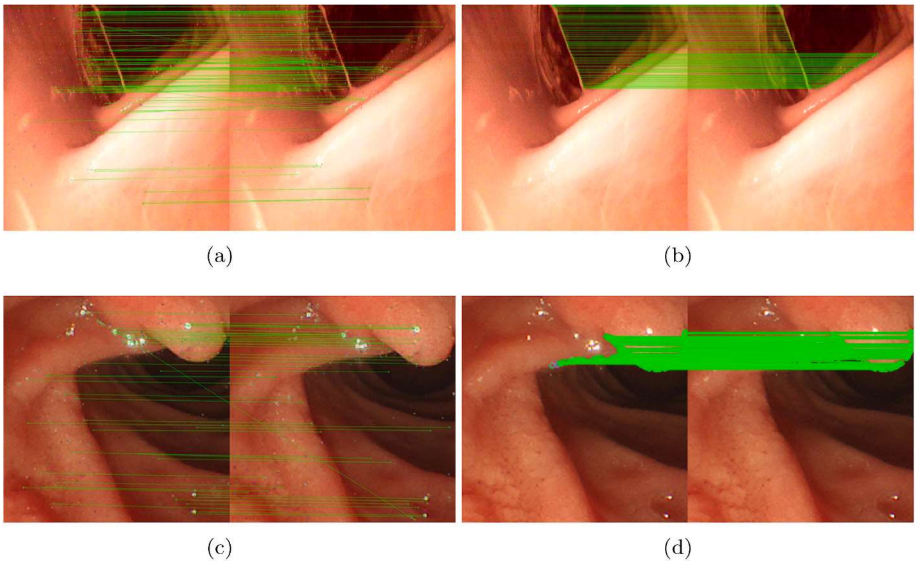


Fig. 4. Examples of our semantic feature matching. (a)–(f) indicate that our method can perform contour matching on the regular contour of the intestine. (g)–(l) show that our method can also perform contour matching on the irregular convex structure of the soft tissue in the intestine.

operation of this method are described [10]. Researchers focused on the nature of chain code, and the steps of rotating, expanding, and smoothing line structure, and studied the steps of determining the similarity between two contours through related technologies [11]. Based on the chain code theory, the recent trend is to use biological abstraction to determine how to move and encode information around boundaries. This method has many applications in image generation [26] and image compression [8,9,17].

2.2. *Structure from motion*

At present, the widely used tracking and reconstruction technologies include simultaneous localization and mapping (SLAM) [24], structure from motion (SfM) [4,19,37], non-rigid structure from motion (NR-SfM) [30–32], shape from template (SfT) [18], optical flow [29], and neural network technology [28,33,42]. According to the tracking method, the location and tracking methods can be divided into two categories: manual tag-based location and unmarked location. Labeled methods are mainly used in rigid organ surgery. SfM and SLAM don’t utilize manual



(a)

(b)

(c)

(d)

Fig. 5. Comparision of semantic feature matching and SIFT feature matching. (a) and (c) are the results of SIFT feature matching. (b) and (d) are the results of contour feature matching.

Table 1
Results of contour extraction.

Sequence	Contour point num	Contours	Method
Seq.1	43.6623	0.769	baseline [38]
	315.868	**2.37865**	**ours**
Seq.2	158.852	2.778	baseline [38]
	318.754	**3.049**	**ours**
Seq.3	351.471	20.563	baseline [38]
	480.966	**6.827**	**ours**
Seq.4	318.733	7	baseline [38]
	484.8	**4**	**ours**
Seq.5	86	0.8	baseline [38]
	280	**2.3**	**ours**
Seq.6	36.9	23.5	baseline [38]
	74.19	**1.99**	**ours**

Table 2
Comparision of different contour selection factors.

Strategy	Accuracy	Highlight factor
Length	**86.6%**	**0.01%**
Area	79.3%	46.8%

Table 3
Results of contour matches in each image pair.

Sequence	Average contour matches
Seq.1	337.68
Seq.2	230.496
Seq.3	182.994
Seq.4	215.456
Seq.5	134.8268
Seq.6	160.5196

Table 4
Feature matches accuracy in semantic feature matching.

Sequence	Baseline	**Ours**
Seq.1	50.5%	**74%**
Seq.2	49.8%	**70.8%**
Seq.3	61.5%	**89.5%**
Seq.4	54.3%	**79.66%**
Seq.5	51.2%	**71.33%**
Seq.6	42%	**68.5%**

tags. SLAM can meet the real-time requirements of tracking task in intraoperative navigation. SfM has better performance in reconstruction which is important for registration. The traditional feature-based SfM utilized threshold strategy to distinguish rigid points from non-rigid points. NR-SfM learns the deformation model from observations, while SfT assumes the defined template and estimate the deformation of each image. These methods all depend on feature extraction. Although the optical flow method does not rely on feature extraction, the changing light does not meet the luminosity consistency assumption.

2.3. *Scene graph*

In SfM, we use a graph that contains the matching relationship between images and features as the input for the tasks afterward. However, the underlying features might not contain specific semantic information. Johnson et al. [16,27,43] first defined a scene graph as a directed graph representation that contains objects and their attributes and relationships. Based on the similarity of the two structures in the above methods, we combine the semantic feature through a scene graph to obtain more robust and dense feature matching.

3. Methods

Given an image sequence, our method can offer dense correspondences with the semantic description for feature matching in order to improve the accuracy of SfM and increase the density of the point cloud. The pipeline is illustrated in Fig. 1. First, we performed feature extraction on the input image, including SIFT features and contour features. The contour feature extraction can obtain the keypoints and three types of descriptors on the contour. The contour descriptors are used as the semantic description of each image for clustering. The clustering result is used as guidance for initialization. Both the contour feature matching pair and the SIFT matching pair are added to the scene graph and provide more accurate pose estimation. The details are shown in Algorithm 1.

Algorithm 1. Scene-graph Driven ContourSfM

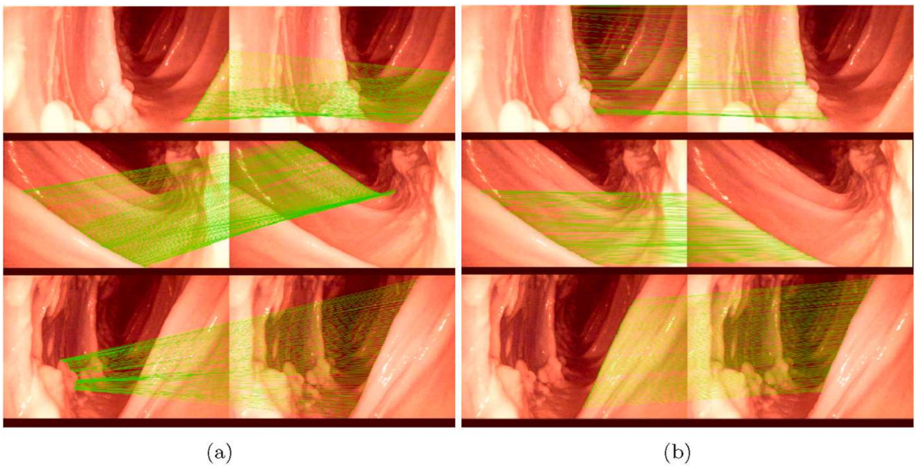


Fig. 6. The influence of different semantic descriptors on feature matching. (a) are the results only using the global shape descriptor. (b) are the results using the global shape and appearance descriptor.

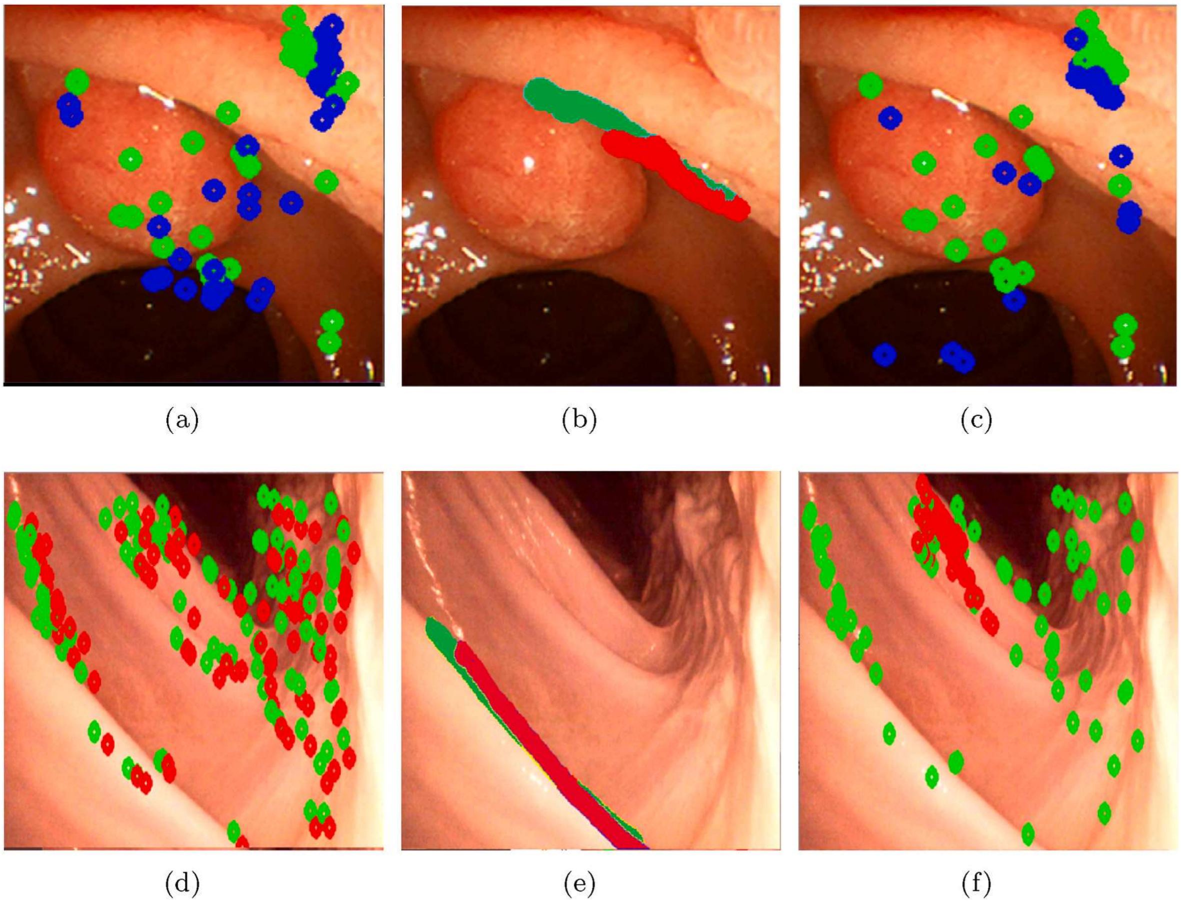


Fig. 7. Comparison of semantic feature matching performance in endoscopy. We use reprojection error as the measurement. The green points are the extracted SIFT points. The blue and red points are the reprojected points. (a) and (d) are the results of semantic feature matching. (b) and (e) are the results of contour matching. (c) and (f) are the results of SIFT feature matching. The distance between the blue point and the green point is closer in (a) and (d), which means the camera pose estimated by our semantic feature matching is more accurate.

Table 5
Improvement of pose estimation accuracy with semantic feature matching.

Sequence	Improved percentage
Seq.1	25%
Seq.2	16.42%
Seq.3	19.8474%
Seq.4	5.47%

Table 6
Reprojection error in semantic feature matching.

Example	Semantic features	Contour	SIFT
1	19.8474	25.7558	138.6640
2	25.8629	27.4980	1199.3329
3	454.693	410.362	5.16245e+19
4	4.91149e-13	5.7567e-13	5.756743e-13

Table 7
Comparison of semantic description performance in clinical endoscopy in the SfM task, × means the method is not available because the soft tissue surface is not Lambert surface. The percentage in brackets is the change rate in point clouds based on [35].

Dataset	**SR point number**	**Registered ratio**	**Methods**
Seq.1	**4614 (+65.3%)**	**140/206**	**ours**
	2791	80/206	[35]
	413	17/206	[12]
Seq.2	**932 (+10.55%)**	**115/193**	**ours**
	843	68/193	[35]
	×	×	[12]
Seq.3	**12021(+5.66%)**	**1565/2001**	**ours**
	11376	1964/2001	[35]
	×	×	[12]
Seq.4	**70**	**10/15**	**ours**
	0	0	[35]
	×	×	[12]
Seq.5	**985(+27.425%)**	**93/97**	**ours**
	773	86/97	[35]
	×	×	[12]
Seq.6	**10062(+90.2%)**	**1021/1400**	**ours**
	5288	724/1400	[35]
	×	×	[12]

Algorithm 1 Scene-graph Driven ContourSfM

```
Input: Image sequence of N images Q = $\{I_1, I_2, ..., I_t, ..., I_n\}$.
Output: Dense feature matches $M_d$.
Conotour keypoints set K and semantic descriptors set D. $K \leftarrow \phi$, $D \leftarrow \phi$.
for all $I_t \in Q$ do
    Extract and store SIFT features for $I_t$.
    Extract contour keypoints $K_t$ and contour descriptor $D_t$ for $I_t$ based on Algorithm 2.
    $K = K \cup K_t$, $D = D \cup D_t$.
end for
Generate the semantic descriptor $S$ with $D$ and similarity measurement based on Section 3.2.
Clustering with $S$ and obtain clustering result $R$.
SIFT feature matching and obtain SIFT feature matches $M_s$.
Contour feature matching based on $S$ and obtain contour feature matches $M_c$ based on Section 3.3.
Generate semantic scene graph $G$ with $M_s$,$M_c$ and $R$ based on Section 3.4.
Initialization with $G$ based on Section 3.5 in SfM .
Incremental reconstruction in SfM and obtain 3D point clouds $P$ and camera poses $C$.
Obtain dense feature matches $M_d$ with $P$ and $C$ based on dense model in Section 3.6.
return $M_d$
```

3.1. Contour dense keypoints and descriptors

To accurately and densely describe the soft tissue contour, we propose a new semantic feature descriptor, which mainly includes dense contour features and SIFT feature [6]. The overall process of extracting key points and descriptors on contours is shown in Fig. 2. We obtain the dense keypoint set of soft tissue contours with a series of matrix operations. And we utilize three descriptors to describe the shape and appearance of one contour. Firstly, the shape context method [2] is used as a global shape descriptor to describe the distribution of all 2D contour keypoints. Secondly, a descriptor based on color information to express the appearance characteristics of the contour is designed to distinguish the similar contour by the surrounding colors. Finally, we devise a local shape descriptor to distinguish the shape details between contours. The details are shown in Algorithm 2.

Algorithm 2. Contour Keypoints and Descriptors Extraction in Single Frame.

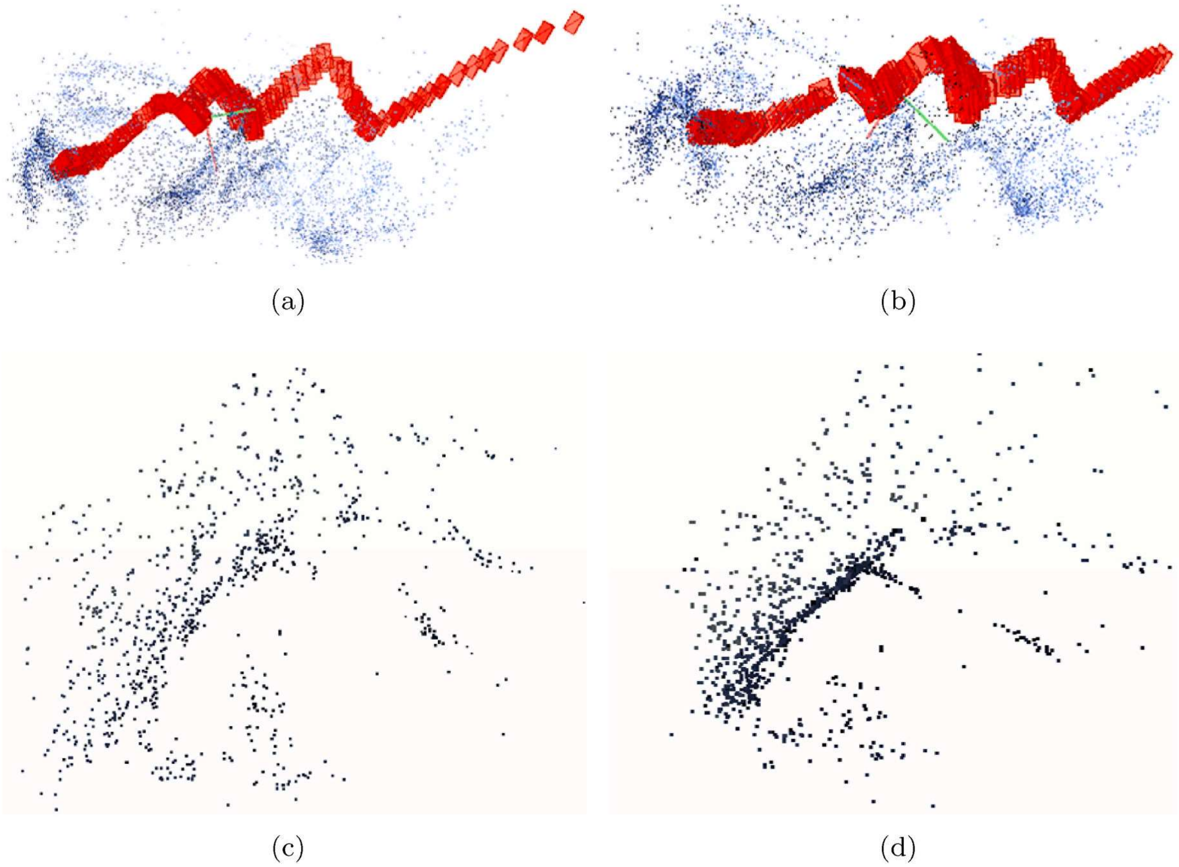


Fig. 8. Qualitative comparison of traditional SfM results and our contourSfM results. (a) and (c) are the results of SfM [35]. (b) and (d) are the results of our contourSfM. Our method has more accurate pose estimation and clearer contour reconstruction.

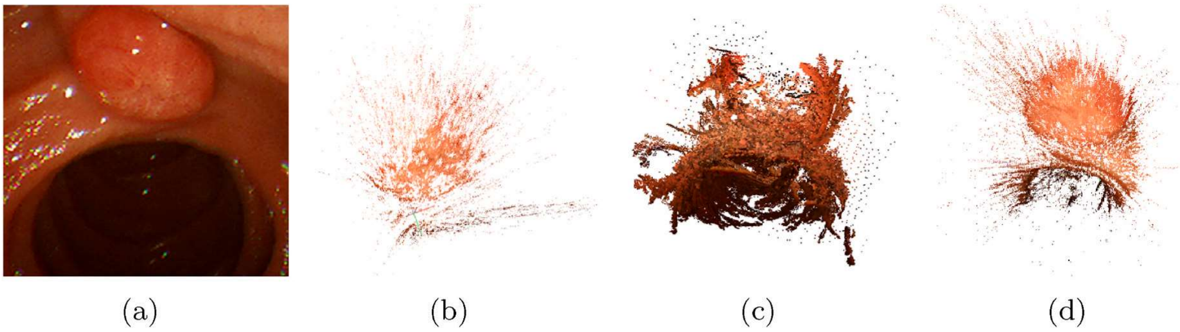


Fig. 9. Qualitative comparison of different SfM methods' influence on dense reconstruction results in endoscopy. (a) The original image is captured by the endoscope. (b) The dense point cloud resulting from the original SfM [36]. (c) The dense point cloud is obtained from Ref. [12]. (d) The dense point cloud resulting from our new method.

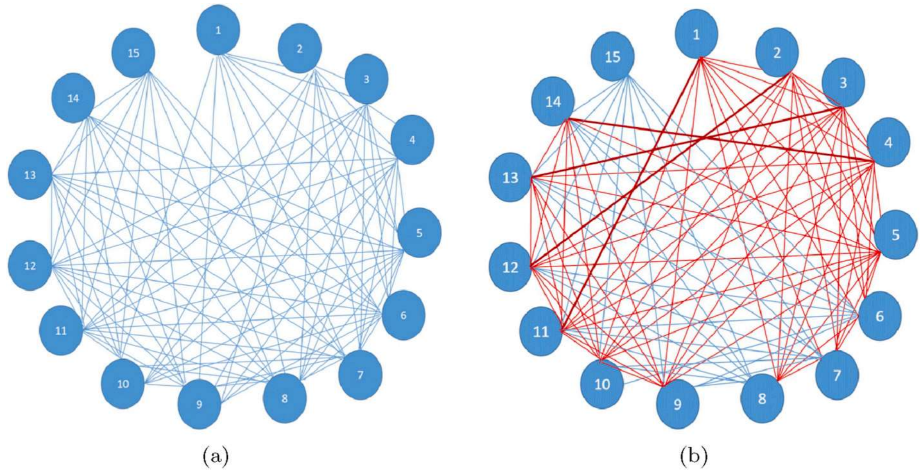


Fig. 10. An illustration of scene graph construction. (a) is the scene graph constructed by SIFT feature matching. (b) is the scene graph augmented by our contour feature matching. Using the scene graph in (a) can not be successfully initialized in some small samples, but by using the scene graph we proposed, SfM can initialize and construct a contour point cloud.

Table 8
Comparision of different SfM results for deep learning. Our contourSfM provides more valid points for the learning-based method.

No.	Images	3D points	Valid points	Methods
Seg.1	86	773	314	[35]
	93	**985**	**526(+67.52%)**	**ours**
Seg.2	745	9434	2829	[35]
	745	**9595**	**3481(+23.04%)**	**ours**
Seg.3	724	5288	513	[35]
	1021	**10062**	**665(+29.6%)**	**ours**
Seg.4	283	2378	212	[35]
	179	**2753**	**310(+46.22%)**	**ours**
Seg.5	80	2791	1225	[35]
	140	**4614**	**1351(+10.2%)**	**ours**
Seg.6	1964	11376	526	[35]
	1565	**12021**	**1391(+164.44%)**	**ours**

Algorithm 2 Contour Keypoints and Descriptors Extraction in Single Frame.

```
Input: Input image $I_t$.
Output: Contour set $\mathbf{B}$, global shape descriptors $S_g$, global
    appearance descriptors $A$, local shape descriptors $S_l$.
$\mathbf{B} \leftarrow \phi$, $S_g \leftarrow \phi$, $S_l \leftarrow \phi$, $A \leftarrow \phi$
$I_d \leftarrow downsampling(I_t)$
$I_c \leftarrow canny(I_d)$
$I_b \leftarrow threshold(I_c)$
$\mathbf{B} \leftarrow findContours(I_b)$
$\mathbf{B} \leftarrow selectContours(\mathbf{B})$
for each $b \in \mathbf{B}$ do
    $s_g \leftarrow shapeContext(b)$
    $a \leftarrow appearance(b)$
    $s_l \leftarrow shapeEncoding(b)$
    $S_g = S_g \cup s_g$, $S_l = S_l \cup s_l$, $A = A \cup a$
end for
return $\mathbf{B}$, $S_g$, $A$, $S_l$
```

Specifically, for dense contour keypoint extraction, we utilize morphological closed operation and hole-filling processing to enhance edges and obtain the stable dense point set. The dense point set is represented by $\mathbf{B} = \{b_1, b_2, \ldots, b_n\}$. We define the shape context of each contour point as s_g. The shape context of the key points composes the descriptor of this contour. For the appearance descriptor, the 256-dimensional descriptor (i.e., local appearance) can be obtained by performing convolution operations and maximum pooling operations on the pixels around the main contours. For the local shape descriptor, a local coordinate system is set up for each contour point. The 128-bit descriptor vector (i.e., local contour shape) can be formed by encoding the relative positions of 64 adjacent points in four quadrants as binary patterns.

3.2. Soft tissue surface deformation clustering

We compute the global similarity G between two images $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ based on the shape and the appearance descriptor, where G $(\mathbf{I}_i, \mathbf{I}_{i+1}) = S(\mathbf{I}_i, \mathbf{I}_{i+1}) + D(\mathbf{I}_i, \mathbf{I}_{i+1})$, S represents the global shape similarity function, D represents the global appearance similarity function. We introduce the appearance descriptor of each contour in last section. The global appearance of an image is constructed by the collection of contours and their appearance descriptors. So the global appearance similarity between images is computed using cosine similarity, as shown in Eq. (1),

$$\mathrm{D}(\mathbf{I}_i, \mathbf{I}_{i+1}) = \frac{1}{m}\sum argmax \frac{\mathbf{v}\cdot\mathbf{u}}{\|\mathbf{v}\|\|\mathbf{u}\|}, \tag{1}$$

where $\mathbf{v}$ represents a descriptor in $\mathbf{I}_i$, and $\mathbf{u}$ is a descriptor in $\mathbf{I}_{i+1}$, m represents the number of contours. For feature compression, we use the shape context method [2] on dense point sets ($\mathbf{B}_i$ and $\mathbf{B}_{i+1}$) to get the global shape similarity, as shown in Eq. (2),

$$\begin{aligned}\mathrm{S}(\mathbf{I}_i, \mathbf{I}_{i+1}) = 1 - \frac{1}{|\mathbf{B}_i|}\sum \arg\min C(b_i, \alpha) - \\ \frac{1}{|\mathbf{B}_{i+1}|}\sum \arg\min C(\beta, b_{i+1}),\end{aligned} \tag{2}$$

where C denotes the matching cost function defined in Ref. [2], $\mathbf{B}_i$ represents the dense point set in $\mathbf{I}_i$, $\mathbf{B}_{i+1}$ represents the dense point set in $\mathbf{I}_{i+1}$, b_i represents a point in $\mathbf{B}_i$, b_{i+1} represents a point in $\mathbf{B}_{i+1}$, $\alpha = T(b_{i+1})$, $\beta = T(b_i)$, 1 is the empirical value, and T denotes the transform function in Ref. [2]. Finally, the endoscopic image sequences are clustered based on the basic sequential algorithm scheme (BSAS).

3.3. Feature matching with semantic feature descriptors

After constructing semantic descriptors and defining similarity

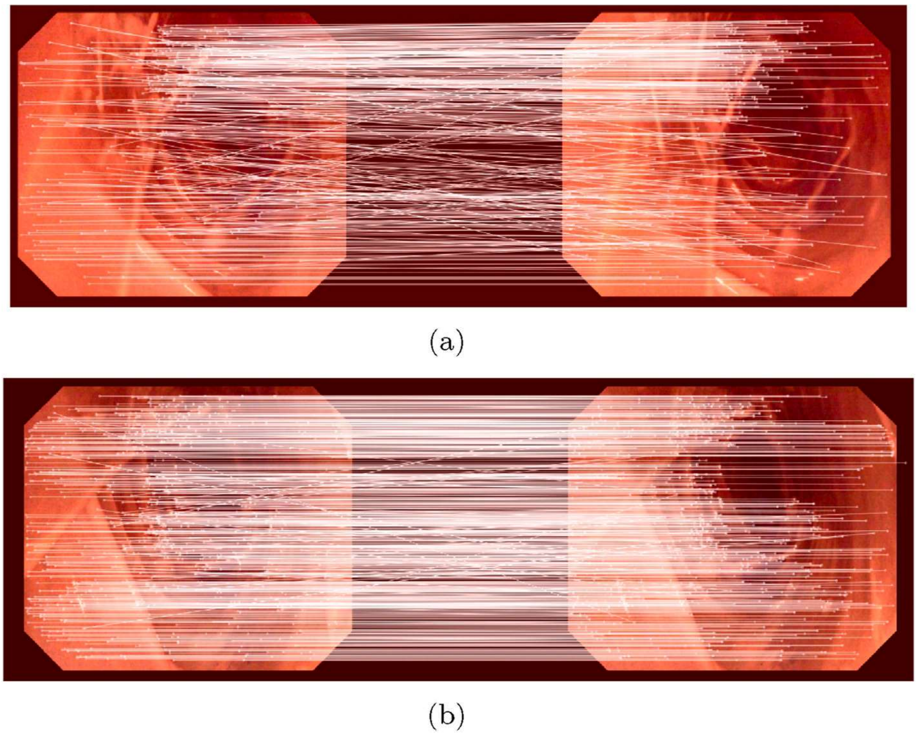


Fig. 11. Deep learning-based feature matching based on different SfM results. (a) is the result from the baseline. (b) is the result of our improved data. Our method provides more valid self-supervised information and promotes the performance of deep learning feature matching.

measurement between semantic features, we perform feature matching between contours of different pictures. A video can be regarded as a sequence of images arranged according to the shooting time. To ensure the overlap of the scene, the interval between each image pair is within 20 frames. First, the global shape descriptor is used to traverse and match the main contours of each pair of pictures, and the candidate contour matching with a higher similarity coefficient is selected. Next, we employ local color features to refine the matching relationship of the contours. In addition, the contour extraction may be incomplete due to changes in the viewing angle. Therefore, the local contour shape descriptor is used to determine the correspondence between the local contour and the global contour. After the above steps, the point-to-point correspondence between the two contours is generated.

In addition, there are some factors that affect the result and accuracy of contour matching, such as the number of points contained in the contour and the length of the contour. Specifically, the two contours may contain the same number of points, but the lengths of the contours are different. It may lead to a number of mismatches. Therefore, the length of the contour is a key property in selecting the candidate contours and the sampling step is also necessary when generating the matching between different points. According to the point number of the shorter contour, we uniformly downsample the longer contour. For example, if there are two contours A and B, and the point number of A is twice the point number of B, then a sample is taken at an interval of one point.

3.4. Semantic feature description with scene graph

In order to obtain the dense feature matching of soft tissue when deformation occurs, we construct a new semantic feature description based on a scene graph, which includes SIFT descriptor, dense point descriptor around contours and their associative feature correspondences. The proposed approach distinguishes the deformed scene with multiple constructed scene graphs. A single scene graph $\mathbf{G} = (\mathbf{O}, \mathbf{E})$ is used to represent the matching relationship between images categorized into the same group using the clustering method, **O** is a set of nodes, and each node represents the semantic feature descriptor of an image. Moreover, **E** is a set of edges, and each edge represents the matching relationship between semantic feature descriptors. Existing methods usually use a single feature descriptor, but our semantic feature descriptor contains two different feature descriptors (SIFT feature descriptor, dense contour descriptor). With the semantic information of contours, it can construct the denser corresponding relationship among point-point sets through the scene graph.

3.5. SfM augmented by semantic feature description

To fully utilize the semantic information among local frames, the semantic feature description **G** is applied to our SfM framework. First, we enhance SfM's performance by utilizing robust and dense correspondences provided by the semantic description. Since our method has distinguished the deformed scene, many outliers can be filtered. Then we improve the non-iterative solution named EPnP [20] in the SfM with dense points. We define the specific cost function E that combines the two types of geometric entities, where $\mathrm{E} = E_s + E_d$, E_s represents the reprojection error of SIFT features, E_d represents the reprojection error of contour points. The camera pose parameters $\theta = \{\mathrm{R,t}\}$ are optimized at each frame with a bundle adjustment (BA) strategy [40].

Generally, the rapid movement of the camera and the deformation of soft tissue will lead to the failure of reconstruction due to the undesirable performance of feature matching. Our semantic description utilizes stable contour features to provide robust and dense feature matching relationships. Moreover, clustering can be used as a priori information to classify the scene information in advance and actively divide it into multiple templates. This can speed up the initialization process, from the original random search to the guided search.

3.6. Data improvement for DenseDescriptorNet training

The dense model called DenseDescriptorNet is based on [23]. The input is an image of the digestive tract scene taken through an endoscope. The sparse 3D reconstructions and camera poses can be obtained from SfM. It obtains the dense features of each image through the DenseNet network and then uses the POI convolutional layer to generate feature matching. Specifically, the feature matching problem is transformed into a key point positioning problem. The keypoint in each picture selects the position with the largest response in the corresponding target heat map.

This method has better results in nasal endoscopy scenes. However, the digestive tract scene is more complicated, and the soft tissues are easily deformed. Through analysis of the acquired features, it can be found that some true values come from the vicinity of the highlight, as shown in Fig. 5 (a) and (c). The use of contourSfM increases the feature matching of the real physical structure. It can add more effective 3D points in the limited data and increase the number of real values. Moreover, it can increase the accuracy of pose estimation together with the sift feature.

4. Results

4.1. Dataset preparation

We employ two datasets for validation. One is the phantom dataset and the other is the real surgery dataset. These two data sets were created by our laboratory. These sequences are representative image sequences in the surgical navigation stage. All experiments are conducted on a workstation with 1 NVIDIA RTX2080ti, with 8 GB memory. Our method is based on COLMAP [35] framework and is implemented using C++ and OpenCV [15].

Phantom dataset. The phantom dataset consists of 4 videos. The phantom data were collected from the carcass of a pig. The average number of images contained in each video is 5200. The content in the phantom picture is mainly the large intestine and rectum. These images mainly simulate the different conditions, including the anatomic structures such as the fat, walls, and folds in the intestine. A sequence can be divided into different small segments to verify different situations.

Real surgery dataset. The real dataset consists of 2 videos from ERCP (Endoscopic Retrograde Cholangiopancreatography). Each video in our own dataset contains 190–210 images. These images mainly include the appearance of the duodenal papilla.

4.2. Contour extraction

We use the traditional method [38] as the baseline, and collect the number of main contours and the number of points contained in each contour as indicators. All results are shown in Table 1. We can find that our method extracted more feature keypoints and longer contours, as shown in the second column. The qualitative results are illustrated in Fig. 3. It can maintain the accuracy and completeness of the contours and avoid incorrect extraction by filtering out small contours such as highlights. As shown in Table 2, we found that using the length of the contour used as an indicator to select the main contours can improve the matching accuracy (increase by 7.3%) and reduce the impact of the highlight (only account for 0.01%). The matching accuracy is calculated by the number of correct contour matching pairs divided by the number of all contour matching pairs. The highlight factor means the number of matching errors caused by highlights divided by the total matching error pairs.

4.3. Feature matching performance with contours

Fig. 4 qualitatively shows the performance of our contour keypoints and descriptors in the task of pair-wise feature matching. The subfigures are endoscopic pictures obtained at different times. The contour shapes and positions in the endoscope image are different, but our method can find the similar contour in most cases, regardless of whether the inner wall of the intestine is regular or irregular. In order to show the matching relationship clearly, we draw partial contours in each subfigure. According to the results in Table 3, the average number of point-to-point correspondences (average contour matches) using dense contour points is 130–330. Average contour matches are calculated by M/P, M presents the total point-wise matches and P is the number of total image pairs. We utilize this metric to show the improvement of our method in point-to-point matching performance. We use the normalized cross-correlation method as the baseline. The accuracy of our contour feature matching is shown in Table 4. Our method can effectively filter out contours that match incorrectly. Our method increases the number of matches and permits the images containing the same soft tissue contour to inter-connect more closely. This description can reflect the local relationship, which makes local matching and local reconstruction more robust.

The comparison between SIFT feature matching and contour feature matching is shown in Fig. 5. The images in (a) and (b) are from the phantom dataset. (c) and (d) are images obtained in real surgery. The SIFT feature used before has a large number of highlight points (white points in Fig. 5 (a) and (c)). Our feature matching is mainly based on contours, which can effectively find the soft tissue structure in the endoscope. The use of contour features can restore a large number of real 3D points with actual semantics. In Fig. 5 (b), our method can recognize specific structural contours. As shown in Fig. 5 (d), our method emphasizes the contours of the intestinal cavity. Fig. 6 illustrates the comprehensive use of global contour shape features and color appearance features. The subfigures in Fig. 6 (a) are the results only depending on the global shape descriptor. The subfigures in Fig. 6 (b) are the improved results. Using color features can significantly improve the accuracy of feature matching and effectively distinguish contours with similar shapes. This is also consistent with the experience of using the observation from eyes to distinguish digestive endoscopy images.

4.4. Pose estimation with semantic feature matching

Fig. 7 shows two specific examples of the accuracy comparision of semantic features, contour features, and SIFT features. The measurement to evaluate the estimated homography matrix is reprojection error. The reprojection error is a geometric error between a projected point and a measured one. In Fig. 7, green points are SIFT points and other points are the re-projection matching points. It can be seen that each pair of green and other points in the first column are evenly distributed and closed. And the outline of the second column is basically coincident. When SIFT points are used alone, the distance between blue points and green points is far, which indicates that the estimated pose is incorrect. The semantic features improved the accuracy of the pose estimation. By using semantic features, we could increase the accuracy of pose estimation by about 5%–25%, as shown in Table 5. Some specific examples of reprojection errors are shown in Table 6.

4.5. SfM performance with semantic feature description

We compared the sparse and dense reconstruction using our proposed semantic descriptor with COLMAP [12,35]. We use the number of registered views (Registered ratio) and the number of sparse reconstruction points (SR points) to evaluate the performance of our method in the task of SfM in endoscopy. Our method has the following two advantages. The first advantage is that our method can obtain more 3D points. Table 7 shows the performance comparison in two datasets. The number of 3D points restored by our method is 90.2% higher than [35] in Seq.6.

In SfM tasks, the numbers of point clouds are larger using our method. Our method increases the number of 3D points, reconstructs more details, reduces the number of outliers, and makes the restored 3D structure clearer, as shown in Fig. 8 (b) and (d). Fig. 9 qualitatively shows the comparison results of dense reconstruction. The result shown in Fig. 9 (c) using [12] contains many outliers.

The second advantage is that our method can successfully initialize and recover the 3D points on the contour when the SIFT feature initialization fails, as shown in Seq.4 in Table 7. We analyze the scene graph based on SIFT feature matching and contour feature matching on small data sets, as shown in Fig. 10. As shown in the scene graph, the nodes in the figure represent the image, and the connecting lines between nodes represent the existence of feature matching. The scene with contour matching is shown in Fig. 10 (b). The red line indicates that contour matching is added between two nodes, and the bold red line indicates that feature matching has not occurred in SIFT feature matching. Through analysis, it can be found that although these 15 images are the same scene, most of the points extracted by SIFT are noise points, which can not exist stably. In addition, the number of SIFT feature matching is also unbalanced, ranging from hundreds to tens, which can not meet the initialization conditions. Our method can increase the number of feature matching, and these feature points can exist stably. Moreover, the 3D points restored by these feature matching are less affected by the change of light. When the matching number of contour features provided by our method meets the initialization conditions, the reconstruction could depend on the contour, otherwise, the reconstruction still depends on the combination of SIFT and contour features.

4.6. Data improvement for DenseDescriptorNet training

Theoretically, the ground truth determines the performance of the network. Therefore, we take the improved reconstructed point clouds as the input of the pipeline. As shown in Table 8, our method has more valid points for deep learning, compared with the original feature matching results. In the best case, valid points are increased by 160% as shown in Seg.6. The reason is that our contour matching can provide better pose and matching information. In other cases, it increased on average 35.31%. In order to show the visual effect of our improved method, we selected some frames in the video. Fig. 11 shows the dense matching results from the learning-based method.

4.7. Limitations

The framework we propose can generate more reconstructed points, estimate more accurate poses, and provide better true values for deep learning. But it is not without limitations. At present, the contour extraction relies on traditional methods, because the rapid movement of the camera may lead to image blurring, resulting in the originally extracted matches being screened out. Fortunately, in practice, surgeons also need to repeatedly confirm and observe the key area that needs to be reconstructed, and the system can continue to extract contours.

5. Conclusion and future work

In this paper, we have combined SIFT feature and dense point feature into a new semantic feature description solely based on the scene graph. We evaluated our SfM system on the endoscope dataset, and several advantages are highlighted. Firstly, our semantic feature description improves the effect of feature matching. Secondly, the application of semantic feature description to SfM improves the accuracy of pose estimation and increases the density of point clouds. Moreover, our method provides better results and promotes the development of learning-based methods. Nevertheless, our method is not without limits. Due to noise and image blurring, the contour extraction may be incomplete and the matching error may increase.

Several research topics are still ongoing, since the current SfM application is offline, it could not support real-time mapping and localization yet. We plan to transfer this work into the existing simultaneous localization and mapping (SLAM) system to make it more effective and robust in real-time endoscopic surgery navigation. As a valuable case study, we expect this method could improve the surgical scene reconstruction in practice. We plan to conduct animal experiments on pig carcasses and verify the effectiveness of our method.

CRediT authorship contribution statement

Zhuoyue Yang: Methodology, Software, Data curation, Investigation, Writing – original draft. **Junjun Pan:** Conceptualization, Writing – review & editing. **Ranyang Li:** Data curation, Resources. **Hong Qin:** Conceptualization, Writing - review.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.61872020, 62172437, U20A20195), Research Unit of Virtual Human and Virtual Surgery, Chinese Academy of Medical Sciences (2019RU004) and Global Visiting Fellowship of Bournemouth University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105616.

References

[1] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Understand. 110 (2008) 346–359, https://doi.org/10.1016/j.cviu.2007.09.014.

[2] S. Belongie, J. Malik, J. Puzicha, in: Shape Context: A New Descriptor for Shape Matching and Object Recognition, NIPS, 2000, pp. 831–837, in: https://proceedings.neurips.cc/paper/2000/file/c44799b04a1c72e3c8593a53e8000c78-Paper.pdf.

[3] C.B. Choy, J. Gwak, S. Savarese, M. Chandraker, in: Universal Correspondence Network, NIPS, 2016, pp. 2414–2422. https://cvgl.stanford.edu/projects/ucn/.

[4] T. Collins, D. Pizarro, S. Gasparini, N. Bourdel, P. Chauvet, M. Canis, L. Calvet, A. Bartoli, Augmented reality guided laparoscopic surgery of the uterus, IEEE Trans. Med. Imag. 40 (2020) 371–380, https://doi.org/10.1109/TMI.2020.3027442.

[5] C.T. Cuenyt Akinlar, Edlines: a real-time line segment detector with a false detection control, Pattern Recogn. Lett. 32 (2011) 1633–1642, https://doi.org/10.1016/j.patrec.2011.06.001.

[6] L.G. David, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110, https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[7] D. DeTone, T. Malisiewicz, A. Rabinovich, in: Superpoint: Self-Supervised Interest Point Detection and Description, CVPRW, 2018, pp. 337–33712, https://doi.org/10.1109/CVPRW.2018.00060.

[8] K. Dhou, A new chain coding mechanism for compression stimulated by a virtual environment of a predator-cprey ecosystem, Future Generat. Comput. Syst. 102 (2020) 650–669, https://doi.org/10.1016/j.future.2019.08.021.

[9] K. Dhou, C. Cruzen, in: An Innovative Employment of the Netlogo Aids Model in Developing a New Chain Code for Compression, ICCS, 2021, pp. 17–25, https://doi.org/10.1007/978-3-030-77961-0_2, 2021.

[10] H. Freeman, On the encoding of arbitrary geometric configurations, IRE Transactions on Electronic Computers EC- 10 (1961) 260–268, https://doi.org/10.1109/TEC.1961.5219197.

[11] H. Freeman, Computer processing of line-drawing images, ACM Comput. Surv. 6 (1974) 57–97, https://doi.org/10.1145/356625.356627.

[12] S. Fuhrmann, F. Langguth, M. Goesele, in: Mve-a Multi-View Reconstruction Environment, GCH, 2014, pp. 11–18, https://doi.org/10.1016/j.cag.2015.09.003.

[13] R.G.V. Gioi, J. Jakubowicz, J.M. Morel, G. Randall, Lsd: a line segment detector, Image Process. Line 2 (2012) 35–55, https://doi.org/10.5201/ipol.2012.gjmr-lsd.

[14] J.H. Han, J.S. Park, Contour matching using epipolar geometry, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 358–370, https://doi.org/10.1109/34.845378.

[15] Itseez, Open source computer vision library. https://github.com/itseez/opencv, 2015.

[16] Y. Jianwei, L. Jiasen, L. Stefan, B. Dhruv, P. Devi, in: Graph R-Cnn for Scene Graph Generation, ECCV, 2018, pp. 670–685, https://doi.org/10.1007/978-3-030-01246-5_41.

[17] D. Khaldoon, C. Christopher, A highly efficient chain code for compression using an agent-based modeling simulation of territories in biological beavers, Future Generat. Comput. Syst. 118 (2021) 1–13, https://doi.org/10.1016/j.future.2020.12.016.

[18] J. Lamarca, J.M.M. Montiel, Camera tracking for slam in deformable maps, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, https://doi.org/10.1007/978-3-030-11009-3_45.

[19] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G.L. Gallia, R.H. Taylor, G.D. Hager, Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data, IEEE Trans. Med. Imag. 37 (2018) 2185–2195, https://doi.org/10.1109/TMI.2018.2833868.

[20] V. Lepetit, F. Moreno-Noguer, P. Fua, Epnp: an accurate o(n) solution to the pnp problem, Int. J. Comput. Vis. 81 (2009) 155–166, https://doi.org/10.1007/s11263-008-0152-6.

[21] H. Leventic, T. Keser, K. Vdovjak, A fast one-pixel wide contour detection method for shapes contour traversal in binary images, in: 2018 International Conference on Smart Systems and Technologies, SST), 2018, pp. 11–14, https://doi.org/10.1109/SST.2018.8564595.

[22] X. Liu, M. Stiber, J. Huang, M. Ishii, G.D. Hager, R.H. Taylor, M. Unberath, in: Reconstructing Sinus Anatomy from Endoscopic Video–Towards a Radiation-free Approach for Quantitative Longitudinal Assessment, MICCAI, 2020, pp. 3–13, https://doi.org/10.1007/978-3-030-59716-0_1.

[23] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G.D. Hager, R.H. Taylor, M. Unberath, in: Extremely Dense Point Correspondences Using a Learned Feature Descriptor, CVPR, 2020, pp. 4846–4855, https://doi.org/10.1109/CVPR42600.2020.00490.

[24] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, J.M.M. Montiel, Live tracking and dense reconstruction for handheld monocular endoscopy, IEEE Trans. Med. Imag. 38 (2019) 79–89, https://doi.org/10.1109/TMI.2018.2856109.

[25] S. Mills, L. Szymanski, R. Johnson, in: Hierarchical Structure from Motion from Endoscopic Video, IVCNZ, 2014, pp. 102–107, https://doi.org/10.1145/2683405.2683411.

[26] E.J.F. do Nascimento, T.M. Castro, A.C.S. Abreu, F.A. Lira, A.H. Souza, in: Procedural Generation of Isometric Racetracks Using Chain Code for Racing Games, SBGames, 2021, pp. 136–143, https://doi.org/10.1109/SBGames54170.2021.00025, 2021.

[27] J. ohnson, R. Krishna, M. Stark, L.J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, in: Image Retrieval Using Scene Graphs, CVPR, 2015, pp. 3668–3678, https://doi.org/10.1109/CVPR.2015.7298990.

[28] K.B. Ozyoruk, G.I. Gokceler, T.L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, et al., Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos, Med. Image Anal. 71 (2021), 102058, https://doi.org/10.1016/j.media.2021.102058.

[29] J. Pan, W. Liu, P. Ge, F. Li, W. Shi, L. Jia, H. Qin, Real-time segmentation and tracking of excised corneal contour by deep neural networks for dalk surgical navigation, Comput. Methods Progr. Biomed. 197 (2020), 105679, https://doi.org/10.1016/j.cmpb.2020.105679.

[30] S. Parashar, Y. Long, M. Salzmann, P. Fua, A closed-form solution to local non-rigid structure-from-motion, arXiv preprint doi:https://ui.adsabs.harvard.edu/abs/2020arXiv201111567P/abstract, , 2020.

[31] S. Parashar, D. Pizarro, A. Bartoli, Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 2442–2454, https://doi.org/10.1109/TPAMI.2017.2760301.

[32] S. Parashar, M. Salzmann, P. Fua, in: Local Non-rigid Structure-From-Motion from Diffeomorphic Mappings, CVPR, 2020, pp. 2059–2067, https://doi.org/10.1109/CVPR42600.2020.00213.

[33] D. Recasens, J. Lamarca, J.M. Fácil, J. Montiel, J. Civera, Endo-depth-and-motion: localization and reconstruction in endoscopic videos using depth networks and photometric constraints, IEEE Rob. Autom. Lett. 6 (2021) 7225–7232, https://doi.org/10.1109/LRA.2021.3095528.

[34] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, in: Orb: an Efficient Alternative to Sift or Surf, ICCV, 2011, pp. 2564–2571, https://doi.org/10.1109/ICCV.2011.6126544.

[35] J.L. Schönberger, J.M. Frahm, in: Structure-from-motion Revisited, CVPR, 2016, pp. 4104–4113, https://doi.org/10.1109/CVPR.2016.445.

[36] J.L. Schönberger, E. Zheng, J.M. Frahm, M. Pollefeys, in: Pixelwise View Selection for Unstructured Multi-View Stereo, ECCV, 2016, pp. 501–518, https://doi.org/10.1007/978-3-319-46487-9_31.

[37] S. Skuratovskyi, I. Gorovyi, V. Vovk, D. Sharapov, Outdoor mapping framework: from images to 3d model, in: 2019 Signal Processing Symposium (SPSympo), IEEE, 2019, pp. 296–399, https://doi.org/10.1109/SPS.2019.8882019.

[38] S. Suzuki, K. be, Topological structural analysis of digitized binary images by border following, Comput. Vis. Graph Image Process 30 (1985) 32–46, https://doi.org/10.1016/0734-189X(85)90016-7.

[39] E. Tola, V. Lepetit, P. Fua, Daisy: an efficient dense descriptor applied to wide-baseline stereo, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 815–830, https://doi.org/10.1109/TPAMI.2009.77.

[40] B. Triggs, P. Mclauchlan, R. Hartley, Bundle ajustment – a modern synthesis, in: Vision Algorithms: Theory and Practice, 2000, pp. 298–372, https://doi.org/10.1007/3-540-44480-7_21.

[41] A.R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, K. Miki, in: 3d Reconstruction of Whole Stomach from Endoscope Video Using Structure-From-Motion, EMBC, 2019, pp. 3900–3904, https://doi.org/10.1109/EMBC.2019.8857964.

[42] H. Yao, R.W. Stidham, Z. Gao, J. Gryak, K. Najarian, Motion-based camera localization system in colonoscopy videos, Med. Image Anal. 73 (2021), 102180, https://doi.org/10.1016/j.media.2021.102180.

[43] L. Yikang, O. Wanli, Z. Bolei, S. Jianping, Z. Chao, W. Xiaogang, in: Factorizable Net: and Efficient Subgraph-Based Framework for Scene Graph Generation, ECCV, 2020, pp. 335–351, https://doi.org/10.1007/978-3-030-01246-5_21.