

ARTICLE

## KST-Mixer: Kinematic Spatio-Temporal Data Mixer For Colon Shape Estimation

Masahiro Oda<sup>a,b</sup> and Kazuhiro Furukawa<sup>c</sup> and Nassir Navab<sup>d</sup> and Kensaku Mori<sup>b,e</sup>

<sup>a</sup>Information and Communications, Nagoya University, Nagoya, Japan; <sup>b</sup>Graduate School of Informatics, Nagoya University, Nagoya, Japan; <sup>c</sup>Department of Endoscopy, Nagoya University Hospital, Nagoya, Japan; <sup>d</sup>Technical University of Munich, München, Germany; <sup>e</sup>Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan

### ARTICLE HISTORY

Compiled September 14, 2022

### ABSTRACT

We propose a spatio-temporal mixing kinematic data estimation method to estimate the shape of the colon with deformations caused by colonoscope insertion. Endoscope tracking or a navigation system that navigates physicians to target positions is needed to reduce such complications as organ perforations. Although many previous methods focused to track bronchoscopes and surgical endoscopes, few number of colonoscope tracking methods were proposed. This is because the colon largely deforms during colonoscope insertion. The deformation causes significant tracking errors. Colon deformation should be taken into account in the tracking process. We propose a colon shape estimation method using a Kinematic Spatio-Temporal data Mixer (KST-Mixer) that can be used during colonoscope insertions to the colon. Kinematic data of a colonoscope and the colon, including positions and directions of their centerlines, are obtained using electromagnetic and depth sensors. The proposed method separates the data into sub-groups along the spatial and temporal axes. The KST-Mixer extracts kinematic features and mix them along the spatial and temporal axes multiple times. We evaluated colon shape estimation accuracies in phantom studies. The proposed method achieved 11.92 mm mean Euclidean distance error, the smallest of the previous methods. Statistical analysis indicated that the proposed method significantly reduced the error compared to the previous methods.<sup>1</sup>

### KEYWORDS

Colon; Colonoscope tracking; Kinematic data estimation; Multi layer perceptron

## 1. Introduction

CT colonography (CTC) is performed to find colonic polyps from CT images. If colonic polyps or early-stage cancers are found in a CTC, a colonoscopic examination is performed to endoscopically remove them. A physician controls the colonoscope based on its camera view during a colonoscope examination. However, its viewing field is limited and unclear because the camera is often covered by fluid or the colonic wall. Furthermore, the colon changes shape significantly during colonoscope insertion. Physicians require much skill and experience to estimate how the colonoscope travels inside the

---

CONTACT Masahiro Oda. Email: moda@mori.m.is.nagoya-u.ac.jp

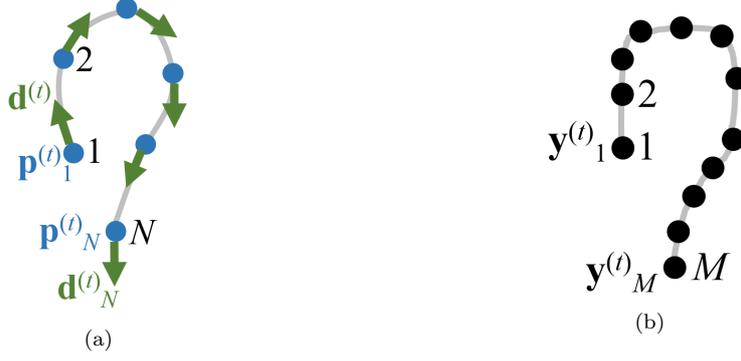
<sup>1</sup>Codes and data of the proposed method are available at: <https://github.com/modafone/stmix>

colon. Inexperienced physicians overlook polyps or cause such complications as colon perforation. A colonoscope navigation system is needed that guides the physician to the polyp position. A colonoscope tracking method is necessary as its core.

Many tracking methods of endoscopes are proposed (Peters 2008; Deligianni 2008; Rai 2008; Deguchi 2009; Visentini-Scarzanella 2017; Wang 2021; Banach 2021; Gildea 2006; Schwarz 2006; Mori 2005; Luo 2015; Liu 2013; Yao 2021; Ching 2010; Fukuzawa 2015; Oda 2017). They can be classified into the image-based, the sensor-based, and the hybrid methods. The bronchoscope is the main application of bronchoscope tracking. Many researchers proposed image- and sensor-based methods in recent several decades. Image-based tracking methods estimate the camera movements based on image registrations. Registrations between temporally continuous bronchoscopic images (Peters 2008) or between real and virtualized bronchoscopic images (Deligianni 2008; Rai 2008; Deguchi 2009) are used for tracking. Recent approaches use deep learning-based depth estimation results to improve image-based tracking accuracy (Visentini-Scarzanella 2017; Wang 2021; Banach 2021). Sensor-based tracking methods use small sensors to obtain bronchoscope position (Gildea 2006; Schwarz 2006). Hybrid methods of them use both image and sensor information to accurately estimate bronchoscope position (Mori 2005; Luo 2015). Some research groups propose tracking methods for colonoscope. In colonoscope tracking, the image-based method (Liu 2013) is difficult to apply because unclear colonoscopic images appear frequently. Unclear image removal and disparity map estimation are utilized to improve tracking accuracy (Yao 2021). Electromagnetic (EM) sensors are used to obtain colonoscope shapes (Ching 2010; Fukuzawa 2015). Unfortunately, they cannot guide physicians to polyp positions because they cannot map the colonoscope shape to a colon in a CT volume, which may contain polyp detection results. Combining the colonoscope position or shape information with polyp position information, which can be detected in a CT volume taken prior to colonoscope insertion, is essential in colonoscope navigation. Such navigation is called colonoscope-CT-based navigation.

A few colonoscope tracking methods are applicable to perform colonoscope-CT-based navigation. Reference (Oda 2017) used colonoscope shape measured by an EM sensor and CT volume for colonoscope tracking. The method obtains two curved lines representing the colon and colonoscope shapes to estimate the colonoscope position on a CT volume coordinate system. The method enables real-time tracking regardless of the colonoscopic image quality. However, the method does not consider colon deformation caused by colonoscope insertion. Such deformation caused significant tracking error. Large tracking errors were observed at the transverse and sigmoid colons, which are significantly deformed by a colonoscope insertion. To reduce tracking errors, estimation methods of the shape of the colon with deformations caused by colonoscope insertion were proposed (Oda 2018a,b). One of them (Oda 2018a) uses the shape estimation network that has a long short-term memory (LSTM) layer (Hochreiter 1997) to estimate the colon shape and the other (Oda 2018b) uses regression forests. Estimation accuracies of them need to be improved to perform accurate colonoscope tracking.

We propose a novel shape estimation method of the colon for colonoscope tracking using a Kinematic Spatio-Temporal data Mixer (KST-Mixer). The proposed method estimates the colon from time-series shape data of the colonoscope, which is inserted into the colon. The KST-Mixer extracts kinematic features from the data of the colonoscope using simple multi-layer perceptrons (MLPs). The extracted features are mixed along the spatial and temporal axes in spatio-temporal mixing blocks to estimate the colon shape dynamically. Because the KST-Mixer has a simple processing flow, it provides estimation results in a short processing time. It is suitable to be used in real-time



**Figure 1.** Examples of (a) colonoscope and (b) colon shapes.

colonoscope navigation systems.

Contributions of this paper are summarized as: **(1)** to propose a novel colon shape estimation method that utilizes spatial and temporal kinematic features extraction and mixing processes, **(2)** to enable short computation time in estimation used in real-time colonoscope tracking, and **(3)** to achieve the smallest shape estimation error among the previous methods.

## 2. Method

### 2.1. Overview

The proposed method estimates the colon shape from the colonoscope shape. They are time-series data measured at a specific time interval. The KST-Mixer is trained to estimate a colon shape from time-series colonoscope shapes. After the training, a trained model estimates a colon shape during a colonoscope insertion.

### 2.2. Colon and colonoscope shape representation

We represent the colon and colonoscope shapes as point sets. The colonoscope shape (Fig. 1 (a)) is a set of 3D positions  $\mathbf{p}_n^{(t)}$  and 3D directions  $\mathbf{d}_n^{(t)}$ , that is represented as

$$\mathbf{X}^{(t)} = \{\mathbf{p}_n^{(t)}, \mathbf{d}_n^{(t)}; n = 1, \dots, N\}, \quad (1)$$

where  $t$  ( $t = 1, \dots, T$ ) is the index of time,  $T$  is the total number of time frames, and  $N$  is the number of points in the colonoscope shape.  $\mathbf{p}_n^{(t)}$  is a point aligned along the colonoscope centerline.  $\mathbf{p}_1^{(t)}$  corresponds to the position of the colonoscope tip.  $\mathbf{d}_n^{(t)}$  is a tangent direction of the colonoscope tube at  $\mathbf{p}_n^{(t)}$ .

The colon shape (Fig. 1 (b)) is a set of 3D points  $\mathbf{y}_m^{(t)}$  that is represented as

$$\mathbf{Y}^{(t)} = \{\mathbf{y}_m^{(t)}; m = 1, \dots, M\}, \quad (2)$$

where  $M$  is the number of points in the colon shape.  $\mathbf{y}_m^{(t)}$  is a point aligned along the colon centerline.  $\mathbf{y}_1^{(t)}$  and  $\mathbf{y}_M^{(t)}$  correspond to the cecum and the anus positions, respectively.

### 2.3. Kinematic spatio-temporal data mixer (KST-Mixer)

#### 2.3.1. Overview of KST-Mixer

The KST-Mixer estimates a colon shape from time-series colonoscope shapes. Its architecture is based on MLPs that are repeatedly applied across the spatial or temporal axes. This architecture is inspired by the MLP-Mixer (Tolstikhin 2021), which classifies images utilizing spatial locations and image features. The MLP-Mixer has competitive image classification performance to current methods such as Vision Transformers (ViT) (Dosovitskiy 2021) and provides a short processing time. We utilize the MLP-based architecture to process kinematic data in shape estimation tasks.

#### 2.3.2. Data preparation

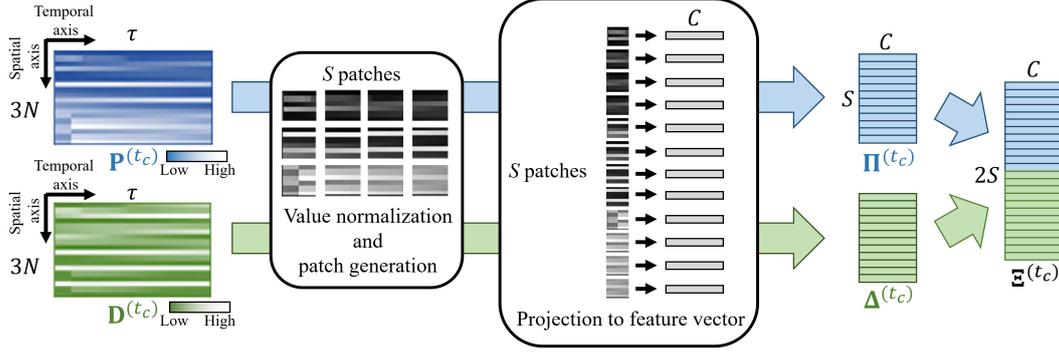
To generate input data of the KST-Mixer, we rearrange a time-series colonoscope shape data as a matrix with spatial and temporal axes. Components of the 3D point and direction are represented as  $\mathbf{p}_n^{(t)} = (p_{n,x}^{(t)}, p_{n,y}^{(t)}, p_{n,z}^{(t)})$  and  $\mathbf{d}_n^{(t)} = (d_{n,x}^{(t)}, d_{n,y}^{(t)}, d_{n,z}^{(t)})$ . From  $\mathbf{p}_n^{(t)}$ , we define the positional matrix ( $3N \times \tau$ ) of time period  $t = t_c, \dots, t_c - \tau + 1$  as

$$\mathbf{P}^{(t_c)} = \begin{pmatrix} p_{1,x}^{(t_c)} & \cdots & p_{1,x}^{(t_c-\tau+1)} \\ p_{1,y}^{(t_c)} & \cdots & p_{1,y}^{(t_c-\tau+1)} \\ p_{1,z}^{(t_c)} & \cdots & p_{1,z}^{(t_c-\tau+1)} \\ \vdots & \ddots & \vdots \\ p_{N,x}^{(t_c)} & \cdots & p_{N,x}^{(t_c-\tau+1)} \\ p_{N,y}^{(t_c)} & \cdots & p_{N,y}^{(t_c-\tau+1)} \\ p_{N,z}^{(t_c)} & \cdots & p_{N,z}^{(t_c-\tau+1)} \end{pmatrix}, \quad (3)$$

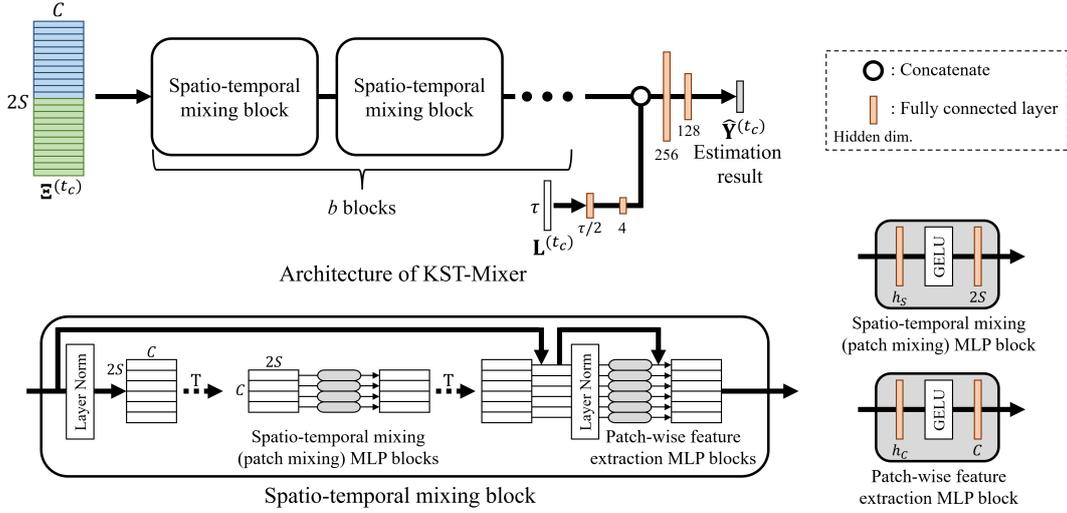
where  $\tau$  is time length. We also define the directional matrix as  $\mathbf{D}^{(t_c)}$  similarly.

Values in  $\mathbf{P}^{(t_c)}$  are normalized to take values in the range  $[0, 1]$ . We regard the normalized matrix as a 2D image to generate  $S$  non-overlapping and homogeneous sized image patches. The image size is  $(3N, \tau)$  and the size of each patch is  $(s_1, s_2)$ . From them, the number of patches is calculated as  $S = \frac{3N\tau}{s_1s_2}$ . Each patch contains spatially and temporally local data. Each patch is projected to a feature vector of hidden dimension  $C$ . As the result, we obtain a input matrix of positional data  $\mathbf{\Pi}^{(t_c)} \in \mathbb{R}^{S \times C}$ . The order of feature values in the matrix is sensitive to both spatial and temporal axes. Therefore, such a position embedding technique as ViT employs is not necessary. We also obtain a input matrix of directional data  $\mathbf{\Delta}^{(t_c)}$  from  $\mathbf{D}^{(t_c)}$ . We make a matrix of colonoscope shape data  $\mathbf{\Xi}^{(t_c)} \in \mathbb{R}^{2S \times C}$  consisting of  $\mathbf{\Pi}^{(t_c)}$  and  $\mathbf{\Delta}^{(t_c)}$  elements. This process is illustrated in Fig. 2.

We generate additional input data of the KST-Mixer from the insertion length of the colonoscope. The insertion length of colonoscope at time  $t$  is  $l^{(t)}$ . A set of insertion lengths in the period of time  $t = t_c, \dots, t_c - \tau + 1$  is represented as  $\mathbf{L}^{(t_c)} = (l^{(t_c)} \dots l^{(t_c-\tau+1)})^T$ , which is a column vector of size  $\tau \times 1$ . The insertion length data  $\mathbf{L}^{t_c}$  is used in the process of the KST-Mixer.



**Figure 2.** Data preparation process from positional and directional 2D-vector matrixes  $\mathbf{P}^{(t_c)}$ ,  $\mathbf{D}^{(t_c)}$  to  $\Xi^{(t_c)}$ .



**Figure 3.** Architecture of KST-Mixer. It estimates colon shape  $\hat{\mathbf{Y}}^{t_c}$  from colonoscope shape data  $\Xi^{t_c}$  and insertion length data  $\mathbf{L}^{t_c}$ . It extracts colonoscope shape features in multiple spatio-temporal mixing blocks. Features are then combined with insertion length feature. Fully connected layers process combined features to generate estimation results.

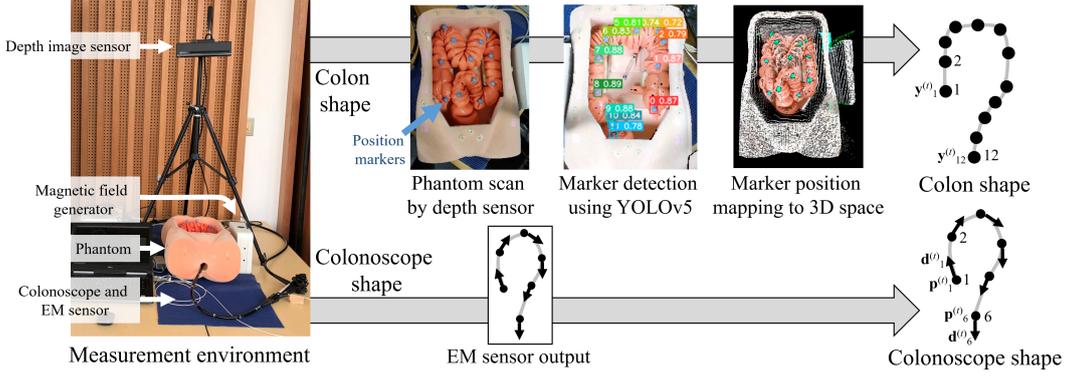
### 2.3.3. Architecture of KST-Mixer

The input of the KST-Mixer are  $\Xi^{(t_c)}$  and  $\mathbf{L}^{(t_c)}$ . The KST-Mixer outputs an estimated colon shape  $\hat{\mathbf{Y}}^{(t_c)}$ . The architecture of the KST-Mixer is shown in Fig. 3. It has  $b$  spatio-temporal mixing blocks. Each of them consists of two MLP blocks similarly to the MLP-Mixer (Tolstikhin 2021). The first is the spatio-temporal feature mixing MLP block (patch mixing MLP block). In the block, an input patch-wise feature vector is transposed and processed by the MLP block. The second MLP block is the patch-wise feature extraction MLP block.

Each MLP block has two fully connected layers and an activation function. Dropout with a probability of  $d_1$  is performed. Operation in the spatio-temporal mixing MLP block can be represented as

$$\mathbf{U}_{*,i} = \mathbf{I}_{*,i} + \mathbf{w}_2 \sigma(\mathbf{w}_1 N(\mathbf{I})_{*,i}), \quad (4)$$

$$\mathbf{O}_{j,*} = \mathbf{U}_{j,*} + \mathbf{w}_4 \sigma(\mathbf{w}_3 N(\mathbf{U})_{j,*}), \quad (5)$$



**Figure 4.** Colonoscope and colon shapes measurement environment. Processing flow of measured data is also illustrated.

where  $\mathbf{I}, \mathbf{O}$  are input and output feature vectors,  $\mathbf{w}_{1,\dots,4}$  are weight parameters of fully connected layers, and  $i = 1, \dots, C, j = 1, \dots, 2S$ .  $N$  is a layer normalization function (Ba 2016).  $\sigma$  is an GELU activation function (Hendrycks 2016).  $*$  indicates the row or the column vectors where operations are applied. Equation (4) is the calculation in the patch mixing MLP block. The calculation is performed for each column of  $\mathbf{I}$ . The number of hidden units of the first fully connected layer in this block  $h_S$  is used to control patch mixing. Equation (5) is the calculation in the patch-wise feature extraction MLP block. The calculation is performed for each rows of  $\mathbf{U}$ . The number of the hidden units of the first fully connected layer in this block  $h_C$  is used to control feature extraction from the patch.

After the processes of the spatio-temporal mixing blocks, feature values are mapped to a vector. It is combined with feature values calculated from the insertion length data  $\mathbf{L}^{(t_c)}$  and then processed by some fully connected layers. Dropout with a probability of  $d_2$  is performed here. The last layer outputs an estimated colon shape  $\hat{\mathbf{Y}}^{(t_c)}$ .

### 3. Experimental Setup

We evaluated the colon shape estimation accuracy of the proposed method in a phantom study. We used a colon phantom (colonoscopy training model type I-B, Koken, Tokyo, Japan), a CT volume of the phantom, a colonoscope (CF-Q260AI, Olympus, Tokyo, Japan), an EM sensor (Aurora 5/6 DOF Shape Tool Type 1, NDI, Ontario, Canada), and a depth image sensor (Kinect v2, Microsoft, WA, USA). We measured colonoscope and colon shapes in our measurement environment shown in Fig. 4. We assume the colonoscope tip is inserted up to the cecum when colonoscope tracking starts because physicians observe and treat the colon while retracting the colonoscope after its insertion up to the cecum. The colonoscope was moved from the cecum to the anus.

#### 3.1. Colonoscope shape measurement

We measured colonoscope shapes using the EM sensor. The EM sensor is strap-shaped with six sensors at its tip and points along its strap-shaped body (one sensor is 6 DOF and remaining are 5 DOF). Each sensor provides a 3D position and a 3D/2D direction

along the colonoscope by inserting the sensor into the colonoscope working channel. The measured colonoscope shape is  $\mathbf{X}^{(t)} = \{\mathbf{p}_n^{(t)}, \mathbf{d}_n^{(t)}; n = 1, \dots, 6\}$  at every time  $t$ .

### 3.2. Colon shape measurement

We measured colon shapes from the colon phantom using the depth image sensor. Twelve position markers were attached to the surface of the colon phantom to scan its shape. The depth image sensor scanned colon shapes during the colonoscope insertions to the colon phantom. We automatically detect marker positions using YOLOv5 (Ultralytics 2020) from the scanned color and depth images. Then, the detection results were manually corrected. The detected markers are described as  $\mathbf{Y}^{(t)} = \{\mathbf{y}_m^{(t)}; m = 1, \dots, 12\}$  at every time  $t$ , used as colon shape.  $\mathbf{y}_1^{(t)}$  and  $\mathbf{y}_{12}^{(t)}$  correspond to the cecum and the anus positions, respectively.

### 3.3. Training and testing of KST-Mixer

We measured both  $\mathbf{X}^{(t)}$  and  $\mathbf{Y}^{(t)}$  during colonoscope insertions to the colon phantom. The shape recording frequency was six times per second.  $\mathbf{X}^{(t)}$  and  $\mathbf{Y}^{(t)}$  belong to the EM and depth image sensor coordinate systems. We registered them in the CT coordinate system using the iterative closest point (ICP) algorithm (Besl 1992) and manual registrations. Registered shape data was used to train and test the KST-Mixer. Parameters used in the trainings were:  $\tau = 18$ ,  $(s_1, s_2) = (6, 3)$ ,  $b = 7$ ,  $d_1 = 0.1$ ,  $d_2 = 0.3$ ,  $h_S = 64$ ,  $h_C = 128$ , 50 minibatch size, and 200 training epochs. Mean squared error was used as the loss function in training. We implemented the KST-Mixer using the Keras build in TensorFlow 2.4.0 running on a Windows PC equipped with a NVIDIA RTX A6000 GPU. The KST-Mixer used 2.5 GBytes of GPU memory in trainings.

In the test step, we provide colonoscope shapes for testing to the trained KST-Mixer. We obtain estimated colon shape  $\hat{\mathbf{Y}}^{(t_c)}$  of current time  $t_c$  from it.

## 4. Experimental Results

We measured colonoscope and colon shapes during eight colonoscope insertions and recorded 1,388 shape pairs. An engineering researcher operated the colonoscope. Leave-one-colonoscopy-insertion-out cross validation was performed for evaluation. We used mean Euclidean distance (MED) (mm) between  $\mathbf{Y}^{(t)}$  and  $\hat{\mathbf{Y}}^{(t)}$  as an evaluation metric. It is defined as

$$\text{MED} = \frac{1}{12T} \sum_{t=1}^T \sum_{m=1}^{12} \|\hat{\mathbf{y}}_m^{(t)} - \mathbf{y}_m^{(t)}\|. \quad (6)$$

We compared the MED of the proposed method with previous colon shape estimation methods, including SEN (LSTM-based method) (Oda 2018a) and regression forests-based method (Oda 2018b). Table 1 shows results of the comparison. The proposed method achieved the smallest MED among the methods. Statistical analysis of the results indicated that the proposed method significantly reduced MED compared to the SEN (Oda 2018a) ( $p < 0.05$  with paired  $t$ -test of MED values). We compared

**Table 1.** Mean and standard deviation of MED calculated in cross-validations performed using proposed and previous methods.

Method	MED (Mean $\pm$ S.D.) (mm)
KST-Mixer (Proposed)	<b>11.92 <math>\pm</math> 1.75</b>
SEN (LSTM-based method) (Oda 2018a)	12.58 $\pm$ 2.08
Regression forests-based method (Oda 2018b)	13.08 $\pm$ 1.55

**Table 2.** Computation times of proposed and previous methods in estimation of one colon shape.

Method	Computation time (msec.)
KST-Mixer (Proposed)	7.3
SEN (LSTM-based method) (Oda 2018a)	2.9
Regression forests-based method (Oda 2018b)	8.9

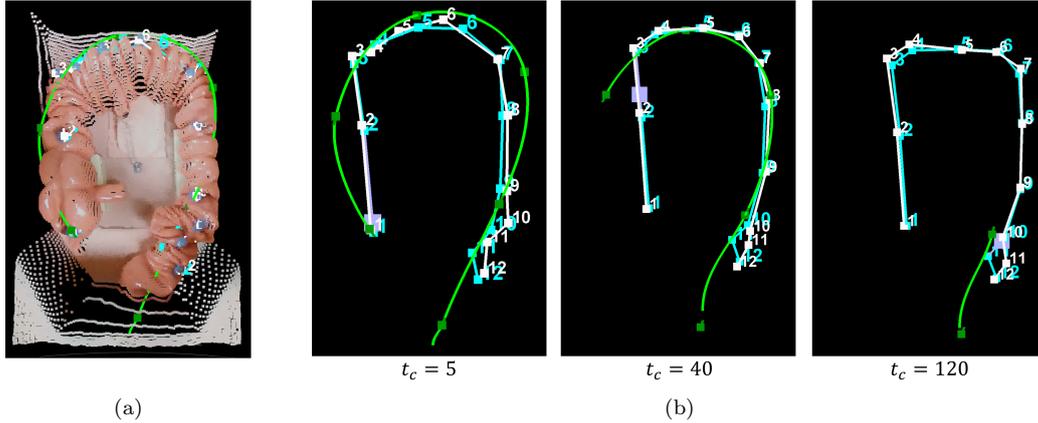
computation times in estimations of one colon shape among these methods. The results are shown in Table 2. From the results, both the proposed and previous methods have real-time performances. Examples of colon shape estimation results are in Fig. 5. The figure shows that the differences between the ground truths and estimated colon shapes were tiny.

## 5. Discussion

The proposed KST-Mixer achieved the smallest error in the colon shape estimation among the previous methods. We designed the KST-Mixer to extract features from kinematic data using simple MLP blocks. Extracted features are then mixed in the spatio-temporal feature mixing MLP blocks to generate spatially and temporally global features. This structure is quite effective in processing time-series kinematic data because we achieved better estimation results than the LSTM-based method (Oda 2018a). The proposed method improves colonoscope tracking accuracy by accurately estimating deformed colon shape during colonoscope insertions. Furthermore, the computation time of the proposed method was short enough to be used in real-time applications.

The application of the proposed method is not limited to colon shape estimation alone. It can be applied to estimations of elastic organs in diagnosis and treatment. Organ shape estimation is essential in surgical assistances by computers. Accurate organ shape estimation contributes to the generation of real-time surgical navigation information and the automation of surgical assistance robots.

Although we have obtained promising results in colon shape estimation, many challenges are still remain for application of the proposed method to colonoscope tracking. Such challenges include (1) *collecting data containing variations of operators and colon shapes*, (2) *collecting in-vivo data*, (3) *development of intuitive visualization method of tracking result*, and (4) *development of a colonoscope that have embedded EM sensors*. (1) *collecting data containing variations of operators and colon shapes* is necessary to improve robustness of the method to real situations. Colonoscope movements have variations among physicians depending on their years of experience. Furthermore, colon shapes also have variations among patients. Colon and colonoscope shapes data containing such variations is necessary to achieve better estimation model. We will measure shape data under operations of colonoscope by physicians of various years of



**Figure 5.** (a) Measurement results of depth image and EM sensors. (b) Colon shape estimation results of three frames. Colonoscope shapes are points on curved green lines. Colon shapes (ground truth) are white numbered points. Estimated colon shapes by proposed method are blue numbered points.

experience. We also measure shape data using many colon phantoms and 3D printed phantoms that have variation of the shapes. (2) *collecting in-vivo data* is necessary to improve the proposed method from phantom level to clinically applicable level. (3) *development of intuitive visualization method of tracking result* is required that presents deformed colon shapes in real-time during colonoscope insertions. Such visualization helps physicians to understand how the colonoscope traveling in the colon. (4) *development of a colonoscope that have embedded EM sensors* is required to perform tracking in clinical situations.

## 6. Conclusions

This paper proposed a colon shape estimation method using the KST-Mixer from kinematic data. The KST-Mixer extracts kinematic features and mixes them along the spatial and temporal axes in multiple MLP blocks. We evaluated the method's estimation accuracy in the colon shape estimation from colonoscope shapes in the phantom study. The proposed KST-Mixer achieved the smallest estimation error in the comparative experiments. Future work includes improvement of the data number using other phantoms, evaluating the method using shape data measured during colonoscope operations by physicians, application to colonoscope tracking, and application to the human colon.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Parts of this research were supported by the MEXT/JSPS KAKENHI Grant Numbers 21K12723, 17H00867 and the JSPS Bilateral International Collaboration Grants.

## References

- Peters, T., Cleary, K.: Image-guided interventions: technology and applications. Springer, Germany (2008)
- Deligianni, F., Chung, A., Yang, G.Z.: Predictive camera tracking for bronchoscope simulation with condensation. MICCAI, LNCS 3749, pp. 910–916 (2005)
- Rai, L., Helferty, J.P., Higgins, W.E.: Combined video tracking and image-video registration for continuous bronchoscopic guidance. *Int J CARS* **3**, 3–4 (2008)
- Deguchi, D., Mori, K., Feuerstein, M., Kitasaka, T., Maurer C.R. Jr, Suenaga, Y., Takabatake, H., Mori, M., Natori, H.: Selective image similarity measure for bronchoscope tracking based on image registration. *Med Image Anal* **3**(14), 621–633 (2009)
- Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., Koto, S.: Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. *Int J CARS* **12**, 1089–1099 (2017)
- Wang, C., Hayashi, Y., Oda, M., Kitasaka, T., Takabatake, H., Mori, M., Honma, H., Natori, H., Mori, K.: Depth-based branching level estimation for bronchoscopic navigation. *Int J CARS* **16**, 1795–1804 (2021)
- Banach, A., King, F., Masaki, F., Tsukada, H., Hata, N.: Visually navigated bronchoscopy using three cycle-consistent generative adversarial network for depth estimation. *Med Image Anal* **73**, 102164 (2021)
- Gildea, T.R., Mazzone, P.J., Karnak, D., Meziane, M., Mehta, A.: Electromagnetic navigation diagnostic bronchoscopy: a prospective study. *American J Respiratory and Critical Care Medicine* **174**(9), 982–989 (2006)
- Schwarz, Y., Greif, J., Becker, H.D., Ernst, A., Mehta, A.: Real-time electromagnetic navigation bronchoscopy to peripheral lung lesions using overlaid CT images: the first human study. *Chest* **129**(4), 988–994 (2006)
- Mori, K., Deguchi, D., Akiyama, K., Kitasaka, T., Maurer C.R. Jr, Suenaga, Y., Takabatake H., Mori, M., Natori, H.: Hybrid bronchoscope tracking using a magnetic tracking sensor and image registration. MICCAI, LNCS 3750, pp. 543–550 (2005)
- Luo, X., Wan, Y., He, X., Mori, K.: Observation-driven adaptive differential evolution and its application to accurate and smooth bronchoscope three-dimensional motion tracking. *Med Image Anal* **24**, 282–296 (2015)
- Liu, J., Subramanian, K.R., Yoo, T.S.: An optical flow approach to tracking colonoscopy video. *Comput Med Imag Graphics* **37**(3), 207–223 (2013)
- Yao, H., Stidham, R.W., Gao, Z., Gryak, J., Najarian, K.: Motion-based camera localization system in colonoscopy videos. *Med Image Anal* **73**, 102180 (2021)
- Ching, L.Y., Moller, K., Suthakorn, J.: Non-radiological colonoscope tracking image guided colonoscopy using commercially available electromagnetic tracking system. In: 2010 IEEE Conference on Robotics, Automation and Mechatronics (2010)
- Fukuzawa, M., Uematsu, J., Kono, S., Suzuki, S., Sato, T., Yagi, N., Tsuji, Y., Yagi, K., Kusano, C., Gotoda, T., Kawai, T., Moriyasu, F.: Clinical impact of endoscopy position detection unit (UPD-3) for a non-sedated colonoscopy. *World J Gastroenterology* **21**(16), 4903–4910 (2015)
- Oda, M., et al.: Robust colonoscope tracking method for colon deformations utilizing coarse-to-fine correspondence findings. *Int J CARS* **12**(1), 39–50 (2017)
- Oda, M., Roth, H., Kitasaka, T., Furukawa, K., Miyahara, R., Hirooka, Y., Goto, H., Navab, N., Mori, K.: Colon shape estimation method for colonoscope tracking using recurrent neural networks. MICCAI, LNCS 11073, pp. 176–184 (2018)
- Oda, M., Kitasaka, T., Furukawa, K., Miyahara, R., Hirooka, Y., Goto, H., Navab, N., Mori, K.: Machine learning-based colon deformation estimation method for colonoscope tracking. *SPIE Med Imag*, **10576**, 1057619 (2018)
- Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
- Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: MLP-Mixer: An all-MLP

architecture for vision. arXiv:2105.01601 (2021)

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. arXiv:1607.06450 (2016)

Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv:1606.08415 (2016)

Ultralytics YOLOv5, <https://github.com/ultralytics/yolov5>. Last accessed 27 Feb. 2022

Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. IEEE PAMI **14**, 239–256 (1992)