Visual Modalities Based Multimodal Fusion for Surgical Phase Recognition

Bogyu Park[1], Hyeongyu Chi[1], Bokyung Park[1], Jiwon Lee[1], Sunghyun Park[2], Woo Jin Hyung[1,2], and Min-Kook Choi[1(✉)]

[1] Vision AI, Hutom, Seoul, Republic of Korea
{bgpark,hyeongyuc96,bokyung,jiwon,wjhyung,mkchoi}@hutom.io
[2] Yonsei University College of Medicine, Seoul, Republic of Korea
{GODON,wjhyung}@yush.ac

Abstract. We propose visual modalities-based multimodal fusion for surgical phase recognition to overcome the limitation of the diversity of information such as the presence of tools. Through the proposed methods, we extracted a visual kinematics-based index related to the usage of tools such as movement and the relation between tools in surgery. In addition, we improved recognition performance using the effective fusion method which is fusing CNN-based visual feature and visual kinematics-based index. The visual kinematics-based index is helpful for understanding the surgical procedure as the information related to the interaction between tools. Furthermore, these indices can be extracted in any environment unlike kinematics in robotic surgery. The proposed methodology was applied to two multimodal datasets to verify that it can help to improve recognition performance in clinical environments.

Keywords: Surgical workflow · Surgical phase recognition · Multimodal learning · Visual kinematics-based index

1 Introduction

Surgical workflow analysis using a computer-assisted intervention (CAI) system based on machine learning or deep learning has been extensively studied [1–10]. In particular, surgical phase recognition can help optimize surgery by activating communication between surgeons and staffs, not only for smooth teamwork, but also for efficient use of resources throughout the entire surgical procedure [11]. Moreover, it is valuable for monitoring the patient after surgery and educational materials through the classification of stereotyped surgical procedures [1]. However, phase recognition is a challenging task that involves many interactions between the actions of the tools and the organs. In addition, surgical video analysis has limitations such as video quality (i.e. occlusion and illumination change) and unclear annotations at event boundaries [2,3].

Many studies that performed surgical workflow analysis have limitations due to performing analysis using only CNN-based visual features and information

X. Li et al. (Eds.): MMMI 2022, LNCS 13594, pp. 11–23, 2022.
https://doi.org/10.1007/978-3-031-18814-5_2

for the presence of tools in video. In this paper, to overcome this limitation, we introduce a visual modality-based multimodal fusion method that improves the performance of phase recognition by using interactions between the recognized tools. The proposed method extracts indices related to tools used in surgery and fuses them with visual features extracted from CNN. We demonstrate the effectiveness of proposed tool-related indices to improve performance by the VR simulator-based dataset and the collected gastrectomy dataset.

We have the following contributions:

- We propose a method to extract a visual kinematics-based index related to tools that are helpful in surgical workflow analysis from visual modality such as semantic segmentation map.
- In addition, it shows that it can be applied in environments where it is difficult to extract the kinematics of tools in a system unlike robotic surgery.
- We propose a fusion method that improves recognition performance by effectively aggregating the visual kinematics-based index and visual features.

2 Related Works

Phase Recognition. In early machine learning-based research, a statistical analysis of temporal information using Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW) was conducted [4]. Since then, as the use of deep learning has become more active, EndoNet [5] that recognizes tool existence through CNN-based feature extraction had been studied. MTRCNet-CL [6], which combines CNN and LSTM to perform multi-tasks, was also performed. Furthermore, a multi-stage TCN (MS-TCN)-based surgical workflow analysis study that performs hierarchically processes using temporal convolution was also performed [10]. Each stage was designed to refine the values predicted by the previous stage to return more accurate predictions. Previous studies had been conducted using only video information for analysis or additionally using only the presence of tools in the video. On the other hand, the proposed method uses a method of fusing visual features and indices related to tools.

Surgical Workflow Dataset. Datasets published to perform surgical workflow recognition include actual surgical videos like Cholec80 [5], toy samples for action recognition of a simple level such as JIGSAWS [12] and MISAW [13], and synthetic data generated from VR simulators PETRAW [14]. In the case of the JIGSAWS and MISAW, kinematic information of the instrument from the master-slave robotic platform was provided, so that more precise tool movements could be analyzed. However, in laparoscopic surgery, it was difficult to use kinematic information owing to the absence of a surgery robot. There was a limit to extracting and applying actual kinematic information due to security issues of the robotic surgery device. To address these problems, we use a method of generating tool-related indices from visual modality to replace kinematic information.

Multimodal Learning. The various modalities (i.e., video, kinematics) created in the surgical environment have different information about the surgical

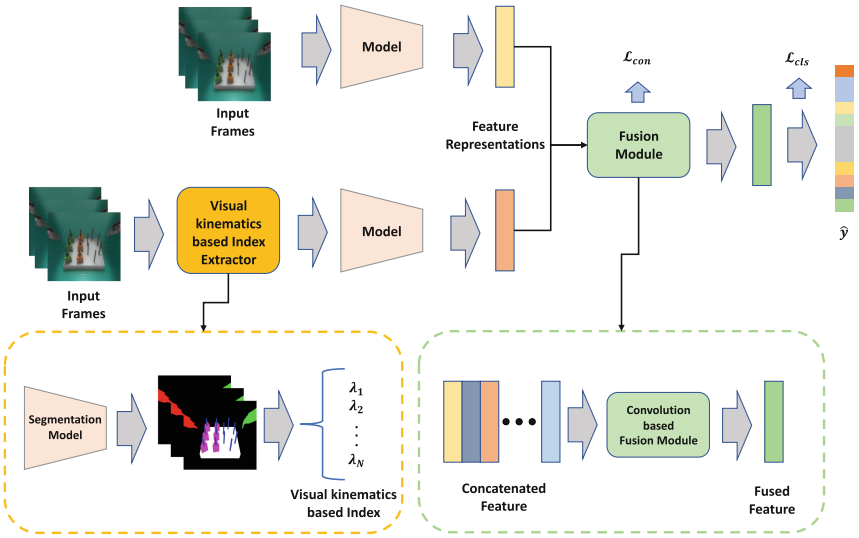

Fig. 1. Proposed visual modalities-based multimodal fusion method. The visual kinematics-based index and frame sequence extracted for the input frame sequence is used as input to the models for each modality. The feature representations of each modality are used as input to the fusion model for joint training.

workflow. Multimodal learning aims to improve performance by using mutual information between each modality. However, researches on multimodal learning in surgical workflow analysis were still insufficient [5,12–15]. In particular, there was a limitation because of related to data that is difficult to access or extract such as the kinematics of surgical tools. We propose a method to effectively achieve performance improvement by fusing various information generated from vision modalities through virtual or real data.

3 Methods

In this section, we propose an extraction manner of a visual kinematics-based index and a visual modality-based multimodal feature fusion method. We used two visual modalities: video and visual kinematics-based index. The visual kinematics-based index expresses the movement and relationship of surgical tools extracted from the semantic segmentation mask. To improve the phase recognition performance, we applied convolutional feature fusion to enhance the interaction of features extracted from visual modalities. The overall learning structure is shown in Fig. 1.

3.1 Visual Kinematics-based Index

A visual kinematics-based index was defined as an index expressing the relationship between tools and the movement of tools. These indices helped to

understand the impact of the action of tools in surgical procedures. Actually, according to previous studies, surgical instrument index which included kinematics extracted from surgical robot or video was used to analyze the skill level of surgeon who performed surgery for all or part of the operation [15–21]. However, indices such as kinematics were extracted from the robot system and were hard to access. To solve this problem, we extracted the visual kinematics-based index by recognizing the tools from the semantic segmentation mask.

Types of Visual Kinematics-based Index. The visual kinematics-based index was consist of two types which are movement or relation between tools. Movement index was measured as {path length, velocity, centroids, speed, bounding box, economy of area} [21]. Movement index measurement is as follows:

$$PL = \sum_{t}^{T} \sqrt{(D(x,t))^2 + (D(y,t))^2}, \quad D(x,t) = x_t - x_{t-1}. \tag{1}$$

$$s = \frac{PL}{T}, \quad v(x) = \frac{x_t - x_{t-\Delta}}{\Delta}. \tag{2}$$

$$EOA = \frac{bw \times bh}{W \times H}. \tag{3}$$

where PL is path length in the current time frame t and T is the time range for computing index. The path length consists of two types which are cumulative path length and partial path length. $D(x,t)$ measures the difference of x coordinate between the previous and current time frame. x and y mean centroids of an object in the frame. Centroids are average positional values for X- and Y-coordinate in the semantic segmentation mask. s is the speed for time range T, and v is the velocity for the direction of X or Y at time interval Δ. bw and bh are the width and height of the bounding box, and W and H are the width and height of the image. Bounding box (BBox) is consist of four values such as top, left, box width, box height (bx, by, bw, bh).

Relation index was measured as {IoU, gIoU, cIoU, dIoU} [21–23]. gIoU, cIoU, and dIoU are modified versions of IoU. The index of IoU family is related to how close two objects are to each other. We considered $\{\lambda_1, ..., \lambda_N\}$ to train phase recognition model by index combination experiments. λ denotes a visual kinematics-based index.

3.2 Feature Fusion

The feature representation for each modality has different information regarding surgical workflow. The representation extracted from the video is related to the overall action in the scene, and the representation extracted from the visual index is related to the detailed movement of each tool. We designed a convolution-based feature fusion module for the interaction of representations to improve recognition performance. For performance comparison, a simple linear feature fusion method and a convolution-based feature fusion method were introduced.

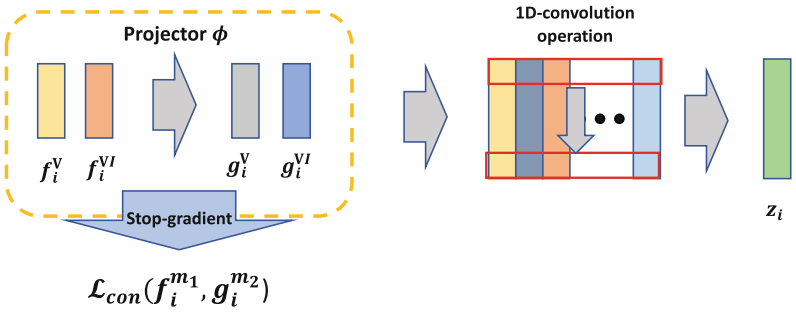

Fig. 2. An illustration of convolution-based feature fusion module. Before feature fusion, enhancement for feature representation is performed by stop-gradient strategy. After then, features are aggregated by 1D-convolutional operation.

Linear Feature Fusion. For each feature representation from modality, the linear fusion module is as follows:

$$f_i^m = \eta(\theta_m(x_i^m)), \quad m \in \{V, VKI\}. \tag{4}$$

$$z_i = \psi(\text{concat}(f_i^V, f_i^{VKI})). \tag{5}$$

where f_i^m is a d-dimensional projected feature for each modality, x_i^m is ith input data of modality m, and θ_m is a deep neural network based recognition model for each modality. V and VKI denote video and visual kinematics-based index. η and ψ are fusion blocks based on Multi-Layer Perceptron (MLP) layers for generating features of another view and aggregating features, respectively. The concatenated feature is aggregated to d-dimensional feature z_i as the input classification layer.

Convolution Based Feature Fusion. Linear fusion module is not an effective approach due to the simple late-fusion method based on a vanilla fully-connected layer. The proposed convolution-based feature fusion module is effective in enhancing interaction between features for phase recognition. The proposed method is processed in 2 steps; 1) Stop gradient-based representation enhancement, 2) Convolutional feature aggregation as shown in Fig. 2.

$$g_i^m = \phi(f_i^m) \tag{6}$$

We apply the stop gradient-based approach proposed in [24] to close the representations of modality with different views and to speed up the learning convergence speed. g_i^m with the same dimension and different view is generated through a projector composed of MLP in Eq. 6. [24] used contrastive loss to learn similarity between representations. According to [24], the contrastive loss is defined as:

$$\mathcal{D}(a_i, b_i) = (\sum_{j=1}^{d} |a_{i,j} - b_{i,j}|^p)^{1/p} \tag{7}$$

$$\mathcal{L}_{con}(f_i^{m_1}, g_i^{m_2}) = \frac{1}{2}\mathcal{D}(f_i^{m_1}, \text{stopgrad}(g_i^{m_2})) + \frac{1}{2}\mathcal{D}(\text{stopgrad}(f_i^{m_1}), g_i^{m_2}) \quad (8)$$

where a_i and b_i are the feature representations of different views, p is the order of a norm and m_1, m_2 are consist with one of $\{V, VKI\}$. Unlike [24], the similarity is calculated using pairwise distance through the experiments. Fused feature representation z_i is forwarded by convolution-based feature fusion as follows:

$$z_i = \Theta(\text{concat}(g_i^V, g_i^{VKI})) \quad (9)$$

where Θ is a 1D convolution-based feature fusion block for kernel size k, z_i is used as input of classifier h to predict $\hat{y}$. Recognition loss $\mathcal{L}_{cls}$ is computed by cross-entropy loss and then total loss is defined as Eq. 11.

$$\mathcal{L}_{cls} = \text{CrossEntropyLoss}(\hat{y}, y), \quad \hat{y} = h(z_i) \quad (10)$$

$$\mathcal{L}_{total} = \mathcal{L}_{con} + \mathcal{L}_{cls} \quad (11)$$

4 Experiment Results

4.1 Base Setting

Dataset. We validated the proposed methods using two different datasets. 1) PETRAW [14] was released at challenge of MICCAI 2021. PETRAW dataset consisted of the pair which are video, kinematics of arms, and semantic segmentation mask generated from VR simulator. Training and test data were constructed with 90 and 60 pairs, respectively. The PETRAW had four tasks such as Phase(3), Step(13), Left action(7), and Right action(7); values in parentheses are the number of classes. 2) The 40 surgical videos for gastrectomy surgery which is called G40 were collected with da Vinci Si and Xi devices between January 2018 and December 2019. We constructed a 30:10 training and evaluation set by considering the patient's demographic data such as {age, gender, pre_BMI, OP_time, Blood_loss, and length of surgery}. According to [3], G40 dataset was annotated for ARMES based 27 surgical phases by consensus of 3 surgeons. G40 consisted of video and semantic segmentation mask with 31 classes, including tools and organs for {harmonic ace, bipolar forceps, cadiere forceps, grasper, stapler, clip applier, suction irrigation, needle, gauze, specimen bag, drain tube, liver, stomach, pancreas, spleen, and gallbladder}. Each instrument consisted of a head, wrist, and body parts[1].

Model. To train models for various modalities, we used Slowfast50 [25] with α, β, and τ for video and Bi-LSTM [26] for kinematics and visual kinematics based index. The segmentation model was trained to predict semantic segmentation masks for generating an index. We used UperNet [27] with Swin Transformer [28] as backbone network.

[1] Please refer supplementary material for class definition details and segmentation results on G40.

Evaluation Metrics. We used various evaluation metrics which are accuracy of whole correctly classified samples, the average version of recall, precision, and F-1 score for classes each task to compare phase recognition results. All metrics were computed frame-by-frame. In all tables, we selected the best models by the average F1 score of tasks.

4.2 Performance Analysis

Table 1. Best combination experiments for visual kinematics based index on PETRAW. $\{\lambda_1, ..., \lambda_N\}$ are indicated in order by cumulative path length(1), partial path length(2), velocity(3), speed(4), EOA(5), centroids(6), IoU(7), gIoU(8), dIoU(9) and cIoU(10). The best combination is selected by mF1-score.

N	Best combination	Phase	Step	Action(L)	Action(R)	Avg.
1	λ_1	88.28	66.68	29.82	29.16	53.48
2	λ_1, λ_2	90.41	67.57	32.62	32.19	55.70
3	$\lambda_1, \lambda_2, \lambda_3$	90.87	68.74	33.12	33.36	56.52
4	$\lambda_1, \lambda_2, \lambda_4, \lambda_6$	90.96	68.85	32.67	33.66	56.53
5	$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_6$	**91.47**	68.85	**34.18**	**34.03**	**57.13**
6	$\lambda_1, \lambda_2, \lambda_3, \lambda_5, \lambda_8, \lambda_{10}$	89.30	67.77	31.71	32.80	55.40
7	$\lambda_1, \lambda_2, \lambda_3, \lambda_6, \lambda_7, \lambda_8, \lambda_{10}$	89.69	**69.02**	34.06	33.09	56.47
8	except λ_8 and λ_{10}	90.48	68.51	32.74	33.19	56.23
9	except λ_{10}	91.03	68.24	33.04	32.34	56.16
10	ALL	89.90	68.31	33.69	33.31	56.30

Important Feature Selection. We extracted various visual kinematics-based indices, and then what kinds of index pairs are positively affected by performance was evaluated on PETRAW in Table 1. λ_1 and λ_2 were related to performance improvement in all cases, and λ_3 was also significantly affected by performance. Figure 3 shows cumulative counts of the index for each combination of best and worst performance. In best combination, $\{\lambda_1, \lambda_2, \lambda_3, \lambda_6\}$ were mostly used but, λ_6 was also related to achieve worst performance. We used $N = 5$ due to achieve the best performance in that combination. The index of the bounding box was included in all combination experiments because that is influenced by performance improvement in Table 2. The bounding box can be synergy by using other indices because it has the positional information (bx, by) and the information of object size (bw, bh). All indices with a bounding box obtained better performance compared to those not used it.

Performance on PETRAW. We used an Adam optimizer with an initial learning rate of 1e-3, an L2 weight decay of 1e-5, a step scheduler for Bi-LSTM and convolution-based fusion method, and a cosine annealing scheduler with a

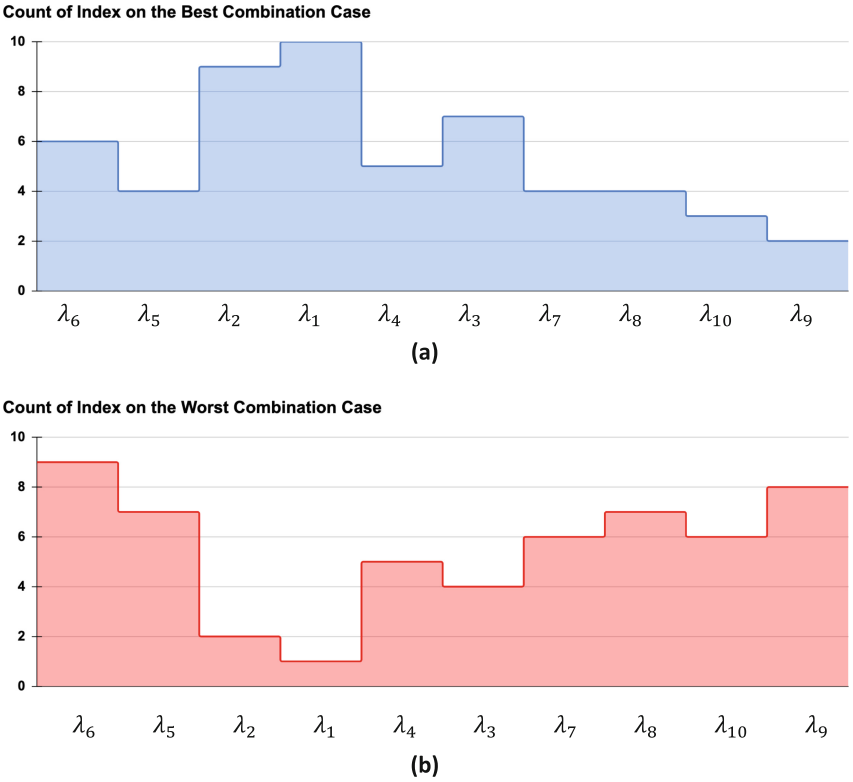

Fig. 3. The histogram of the visual kinematics-based index for best and worst performance. (a) Cumulative counts of each index on the combination of best performance (b) Cumulative counts of each index on the combination of worst performance.

warmup scheduler during 34 epochs for slowfast and linear fusion method. A batch size of 128 was used in all experimental environments. The learning rate decay rate was applied at 0.9 every five epochs for step scheduler. According to [25], α, β, and τ were set $\{4, 8, 4\}$ in slowfast. The hidden layer size and output dimension of Bi-LSTM were set at 256 and 256, respectively. Projected feature size d set 512 for both fusion modules, and convolution kernel size k was 3. To address data imbalance, all networks used class-balanced loss [29] and trained for 50 epochs. We also used train and test datasets which were subsampled by 5 fps. The clip size was 8, and the time range T was the same as the clip size.

Table 3 shows mF1 performances for each modality on PETRAW dataset. The baselines, including video and kinematics, were compared to the visual kinematics-based index. Especially, performances of phase and step by visual kinematics based index were achieved similar performance compared to kinematics based performance. It verified that visual kinematics based index can be

Table 2. Evaluation for impact of bounding box. Each row is the performance using a single index. The value in parentheses is the improvement in adding the bounding box, and the bold is the most significant improvement.

Index	Phase	Step	Action(L)	Action(R)	Avg.
BBox only	55.63	27.75	19.57	20.16	30.78
λ_1	83.98(+4.30)	61.37(+5.31)	**9.99(+19.82)**	**10.19(+18.97)**	41.38(+12.10)
λ_2	42.68(+16.84)	14.82(+15.10)	16.32(+9.02)	14.26(+10.74)	22.02(+12.93)
λ_3	**35.88(+22.95)**	13.47(+16.08)	14.78(+8.73)	13.67(+10.07)	19.45(+14.46)
λ_4	35.88(+21.61)	8.75(+20.51)	11.87(+10.95)	10.29(+12.30)	**16.70(+16.34)**
λ_5	36.63(+19.91)	15.38(+13.76)	14.32(+7.91)	14.17(+7.56)	20.13(+12.29)
λ_6	48.58(+6.77)	20.83(+6.59)	17.80(+2.61)	18.11(+3.33)	26.33(+4.82)
λ_7	34.55(+20.75)	7.59(+19.99)	10.16(+10.56)	10.17(+10.54)	15.62(+15.46)
λ_8	34.54(+20.58)	7.18(+20.38)	10.01(+11.01)	10.16(+10.92)	15.47(+15.72)
λ_9	34.22(+21.39)	**6.82(+20.68)**	10.18(+9.98)	10.15(+11.83)	15.34(+15.97)
λ_{10}	33.80(+21.14)	7.10(+20.10)	10.06(+12.02)	10.16(+11.59)	15.28(+16.21)

helpful to recognize the actions of tools in Tables 1, 2, and 3[2]. Furthermore, the proposed fusion technique achieved improved performance compared to baseline. Our fusion methodology was useful for fusing the representations by enhancing the interactions between features.

Performance on G40. As like setting of PETRAW, we used the same setting of training models. However, the initial learning rate was set 1e-2, weighted cross-entropy loss was used for slowfast, and a cosine annealing scheduler was used for all experiments. A batch size of 64 was used in all experimental environments, and all networks were trained for 50 epochs. The sampling rate was set 1 fps for train and test datasets. The clip size was 32, and the time range T was the same as the clip size. It also improved performance by using the visual kinematics-based index on G40 in Table 4. That is, the visual kinematics-based index was available to replace the kinematics in actual surgery.

Table 3. Performance change for each modality on PETRAW. {V, K, VKI} denote video, kinematics and visual kinematics based index.

Model	Modality	Phase	Step	Action(L)	Action(R)	Avg.
Slowfast50	V	98.13	96.15	79.52	78.72	88.13
Bi-LSTM	K	96.79	80.52	78.10	77.01	83.11
Bi-LSTM	VKI	91.47	68.85	34.18	34.03	57.13
Linear Fusion	V+K	98.26	96.13	80.45	**81.86**	86.14
Conv. Fusion	V+K	**98.59**	**96.43**	**82.57**	81.83	**89.85**
Linear Fusion	V+VKI	98.21	96.28	79.93	79.17	85.12
Conv. Fusion	V+VKI	98.23	96.38	79.87	78.98	88.36

[2] Please refer to supplementary material for additional experimental results of Accuracy, mPrecision, mRecall, and mF1 on PETRAW.

Table 4. Performance change of each modality on G40. mPrecision, mRecall, and mF1 are measured by the average of results for each class.

Model	Modality	Accuracy	mPrecision	mRecall	mF1
Slowfast50	V	63.37	55.40	59.10	55.49
Bi-LSTM	VKI	50.53	40.32	36.79	34.80
Linear Fusion	V+VKI	**69.71**	56.58	58.83	56.76
Conv. Fusion	V+VKI	67.71	**56.75**	**60.19**	**57.41**

4.3 Ablation Study

Visual Kinematics Based Index for Organs. The surgical procedure was related to the interaction between tools and organs. Therefore, relation indices of tools and organs can be helped for recognition performance. We evaluated the performance change by involving a relation index between tools and organs. We used λ_8 and λ_{10} measured between tools and organs for considering the relationship. The comparison is shown in Table 5. Those indices were validated to help recognize the surgical procedure by improved performance.

Table 5. The comparative results for including indices of organs on G40. We compared by adding the relation index between tools and organs, including the liver, stomach, pancreas, spleen, and gallbladder.

Model	Index	Accuracy	mPrecision	mRecall	mF1
Bi-LSTM	tools only	52.58	41.40	40.76	39.46
Bi-LSTM	add organs	**53.72**	**44.04**	**41.10**	**40.67**

Change of Semantic Model. We evaluated the change in performance regarding segmentation models. We considered three models, DeeplabV3+ [30], UperNet [27], and OCRNet [31]. UperNet used Swin Transformer [28] as backbone network and HRNet [32] for OCRNet. We used the basic setting of MMSegmentation [33] to train models during 100 and 300 epochs on PETRAW and G40, respectively. According to accurate segmentation results, the performance was improved in Table 6.

Table 6. Performance change for various segmentation models on PETRAW. The values in table are mF1-score for each task.

Seg. Model	Target Model	mIoU	Phase	Step	Action(L)	Action(R)	Avg.
DeeplabV3+	Bi-LSTM	**98.99**	89.91	61.83	24.33	22.40	49.62
OCR-HRNet	Bi-LSTM	98.98	**92.06**	68.67	31.71	**35.02**	56.86
Swin-UperNet	Bi-LSTM	98.94	91.47	68.85	**34.18**	34.03	**57.13**

Table 7. Performance change for various segmentation models on G40.

Seg. Model	Target Model	mIoU	Accuracy	mPrecision	mRecall	mF1
DeeplabV3+	Bi-LSTM	85.14	50.20	39.96	37.58	36.69
OCR-HRNet	Bi-LSTM	86.45	50.40	39.66	40.30	38.39
Swin-UperNet	Bi-LSTM	**87.64**	**52.58**	**41.40**	**40.76**	**39.46**

5 Conclusion

We proposed a visual modalities-based feature fusion method for recognizing surgical procedures. We extracted a visual kinematics-based index from a visual modality such as a semantic segmentation map and trained the model using the indices and visual features from CNN. We validated that our approach helped to recognize the surgical procedure in simple simulation (PETRAW) and actual surgery (G40). In addition, the visual kinematics-based index is expected to be helpful in non-robotic surgery like laparoscopic surgery due to generating them from visual modality. For further study, we will consider evaluating by extracting a visual kinematics-based index from other visual modalities such as the object detection model.

Acknowledgement. "This research was funded by the Ministry of Health & Welfare, Republic of Korea (grant number : 1465035498 / HI21C1753000022)."

References

1. Zisimopoulos, O., et al.: DeepPhase: surgical phase recognition in CATARACTS Videos. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 265–272. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_31
2. Klank, U., Padoy, N., Feussner, H., Navab, N.: Automatic feature generation in endoscopic images. Int. J. Comput. Assist. Radiol. Surg. **3**(3), 331–339 (2008). https://doi.org/10.1007/s11548-008-0223-8
3. Hong, S., et al.: Rethinking generalization performance of surgical phase recognition with expert-generated annotations. arXiv preprint. arXiv:2110.11626 (2021)
4. Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O., Navab, N.: Statistical modeling and recognition of surgical workflow. Med. Image Anal. **16**(3), 632–641 (2012)

5. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2016)
6. Jin, Y.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. **59**, 101572 (2020)
7. Lecuyer, G., Ragot, M., Martin, N., Launay, L., Jannin, P.: Assisted phase and step annotation for surgical videos. Int. J. Comput. Assist. Radiol. Surg. **15**(4), 673–680 (2020). https://doi.org/10.1007/s11548-019-02108-8
8. Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., Jannin, P.: Automatic data-driven real-time segmentation and recognition of surgical workflow. Int. J. Comput. Assist. Radiol. Surg. **11**(6), 1081–1089 (2016). https://doi.org/10.1007/s11548-016-1371-x
9. Loukas, C.: Video content analysis of surgical procedures. Surg. Endosc. **32**(2), 553–568 (2017). https://doi.org/10.1007/s00464-017-5878-1
10. Czempiel, T., et al.: TeCNO: surgical phase recognition with multi-stage temporal convolutional networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 343–352. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_33
11. Maier-Hein, L., et al.: Surgical data science for next-generation interventions. Nat. Biomed. Eng. **1**(9), 691–696 (2017)
12. Gao, Y., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): a surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2cai, vol. 3 (2014)
13. Huaulmé, A., et al.: Micro-surgical anastomose workflow recognition challenge report. Comput. Methods Programs Biomed. **212**, 106452 (2021)
14. Huaulmé, A., et al.: Peg transfer workflow recognition challenge report: does multi-modal data improve recognition? arXiv preprint. arXiv:2202.05821 (2022)
15. Khalid, S., Goldenberg, M., Grantcharov, T., Taati, B., Rudzicz, F.: Evaluation of deep learning models for identifying surgical actions and measuring performance. JAMA Netw. Open **3**(3), e201664–e201664 (2020)
16. Funke, I., Mees, S.T., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3D convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1217–1225 (2019). https://doi.org/10.1007/s11548-019-01995-1
17. Hung, A.J., Chen, J., Jarc, A., Hatcher, D., Djaladat, H., Gill, I.S.: Development and validation of objective performance metrics for robot-assisted radical prostatectomy: a pilot study. J. Urol. **199**(1), 296–304 (2018)
18. Lee, D., Yu, H.W., Kwon, H., Kong, H.J., Lee, K.E., Kim, H.C.: Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. J. Clin. Med. **9**(6), 1964 (2020)
19. Liu, D., et al.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9522–9531 (2021)
20. Birkmeyer, J.D., et al.: Surgical skill and complication rates after bariatric surgery. N. Engl. J. Med. **369**(15), 1434–1442 (2013)
21. Oropesa, I., et al.: Eva: laparoscopic instrument tracking based on endoscopic video analysis for psychomotor skills assessment. Surg. Endosc. **27**(3), 1029–1039 (2013). https://doi.org/10.1007/s00464-012-2513-z
22. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 658–666 (2019)

23. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence **34**, 12993–13000 (2020)
24. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021)
25. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
26. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**(11), 2673–2681 (1997)
27. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 418–434 (2018)
28. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
29. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)
30. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
31. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 173–190. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_11
32. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)
33. MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020)