An Improved Kinect-Based Real-Time Gesture Recognition Using Deep Convolutional Neural Networks for Touchless Visualization of Hepatic Anatomical Models in Surgery

Jiaqing Liu[1], Kotaro Furusawa[1], Tomoko Tateyama[2], Yutaro Iwamoto[1], and Yen-wei Chen[1]
[1] Graduate School of Information Science and Engineering, Shiga, Japan
[2] Department of Computer Science Hiroshima Institute of Technology, Hiroshima, Japan
Email: {gr0302kv, is0326fp}@ed.ritsumei.ac.jp, t.tateyama.es@cc.it-hiroshima.ac.jp, yiwamoto@fc.ritsumei.ac.jp, chen@is.ritsumei.ac.jp

***Abstract*—Visualization of three-dimensional (3D) medical images is an important tool in surgery, particularly during the operation. However, it is often challenging to review a 3D anatomic model while maintaining a sterile field in the operating room. Thus, there is a great interest in touchless interaction using hand gestures to reduce the risk of infection during surgery. In this paper, we propose an improved real-time gesture-recognition method based on deep convolutional neural networks that works with a Microsoft Kinect device. A new multi-view RGB-D dataset consisting of 25 hand gestures was constructed for deep learning. The nine gestures that were associated with the high recognition accuracies were selected for the touchless visualization system. A deep network architecture, AlexNet, was used for the hand gesture recognition. The recognition accuracy was about 96.5%, which was much higher than that in our previous systems. We further demonstrated that this technique facilitates touchless real-time visualization of hepatic anatomical models during surgery. This system is expected to ultimately lead to better patient outcomes by enhancing the ability to visualize medical images in 3D during surgery.**

***Index Terms*—hand gestures recognition, deep learning technique, surgery aid system**

I. Introduction

Understanding a patient's anatomical hepatic structure is important and essential for successful liver surgery [1], [2]. However, ease of use and hygiene standards make it challenging to use traditional systems while operating. Touchless three-dimensional (3D) visualization during an operation is an open problem with significant potential to address these issues [3].

A few touchless surgery-support systems using the Microsoft Kinect device have been developed [4]-[6]. However, many such systems rely on similar approaches, such as the use of the basic gesture interface of Kinect to interact with medical images [7], which is not sufficient for robust and real-time touchless visualization.

Thus, in this study, we focus on developing a more intuitive of operation scheme and meeting the requirements for interacting with 3D medical images in the operating room. In our previous system [8] (the first version), we used HOG as features and SVM as a classifier to recognize nine different hand gestures from the depth images. The average recognition accuracy was 87.5%, which is sufficient, but the speed was only eight frames per second, indicating that the system could not achieve real-time gesture recognition. In the second version of the system [9], we implemented high-level Kinect APIs provided by Microsoft to automatically recognize three different hand states, which were combined into hand movements that can be used to interact with medical images. In addition to hand states and movements, depth information was used to respond only to users in a predefined range (the operating range of 2.5-3.5 m) while actions or gestures out of the operating range were considered noise and disregarded; this allowed the system to only respond to the gestures of the surgeon, which is particularly important in an operating room. This version of the system, however, was not able to handle complicated interactions and flexibility in the interactions. In the third version of the system [10], we implemented gesture recognition based on deep learning using only a simple LeNet [11]. This touchless hepatic surgery-support system was trained with an existing dataset of limited hand gestures [12]. Experimental testing revealed that the system had a rapid response time, but the accuracy was only 84.3%. Since deep learning usually requires a very large amount of labeled data to return acceptable results, we constructed a new multi-view RGB-D dataset (MaHG-RGBD) of 15 participants performing 25 different hand gestures [13] including front views as well as views from a degree angle, which are relevant as space is often limited especially in the operating room. After building the dataset, AlexNet [14] was used to recognize the hand gestures. The nine hand gestures that were most robustly recognized were applied for touchless visualization of the 3D medical images. The experimental results demonstrate that this version the

Manuscript received January 29, 2019; revised May 13, 2019.

doi: 10.18178/joig.7.2.45-49

system outperforms previous versions in terms of both the recognition time and the accuracy.

Herein, the proposed version of our touchless visualization system is presented in Section 2, including a detailed description of the hand pre-processing and hand gesture recognition using deep learning. Experimental results are reported in Section 3 and Section 4 provides a conclusion.

II. Proposed System

First, we describe the proposed system in detail. The diagram in Fig. 1 summarizes the architecture of the system include its two modules: an interaction module and a visualization module. When the Kinect sensor detects that a user's hand enters an available state (i.e., the nearest user's right hand is positioned 45 cm above the waist), it performs a fast and flexible hand gesture recognition using the trained deep learning model. The classified gesture and the associated hand movement are processed by a command module within the interaction module. This information is sent through a socket to the visualization module, which responds to the command and by performing the associated operation such as rotating the image, adjusting the opacity, zooming in or out, or fusing or selecting vessels. The steps of this process are as follows:

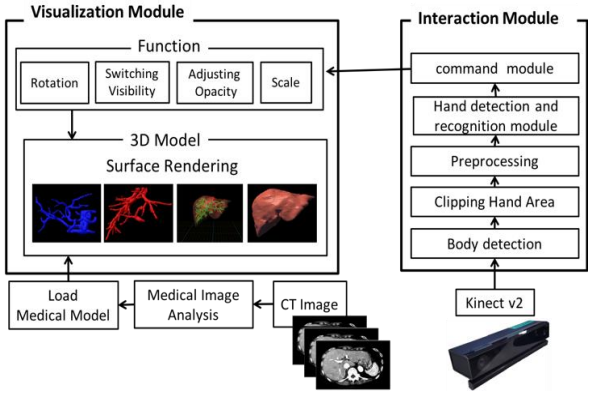


Figure 1. Diagram of our proposed system

A. Hand Image Pre-processing

The depth information and skeleton tracking provided by the Kinect are utilized to generate a depth-resolved image of the hand. First, a depth image of the user (Fig. 2(a)) is generated. Then, calibration is conducted between the color and the depth from the camera. Using the right-hand joint point as the center, a 100×100 pixels square region is extracted as the region of interest (ROI) (Fig. 2(b)). A depth image with a range from d - 30 cm to d + 5 cm, where d is the depth of the right-hand joint point. was defined as the hand image.

The segmented hand image is shown in Fig. 2(c). Then, an opening operator and a median filter are applied to remove the noise (i.e., pixels of other regions) from the hand image) (Fig. 2(d)). To reduce the computation time and focus the analysis on regions of the hand shape, the image was cropped and resized to 32×32 pixels to be used as an input to the convolutional deep neural network.

(a) (b) (c) (d)

Figure 2. The depth hand image for gesture recognition (the rad point is the right hand joint point detected by Kinect). (a) Depth image from Kinect. (b) Decision ROI of depth hand image. (c) Segmented depth hand image including noise. (d) Depth hand image excluding noise.

B. CNN Architecture

AlexNet [10] was adopted as the deep network architecture. The architecture of AlexNet is summarized in Fig. 3. It contains eight learned layers: five convolutional and three fully-connected layers. The input is hand image of 224×224 pixels. The output is 25 classes of gestures.

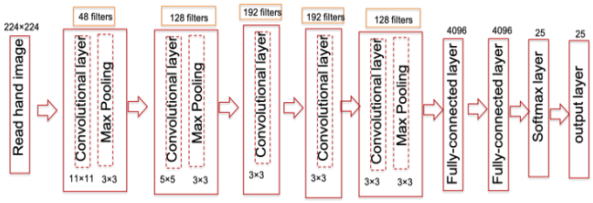
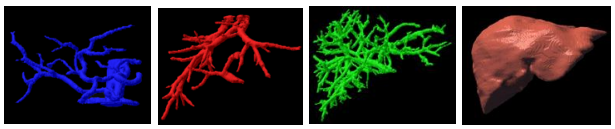


Figure 3. AlexNet for hand gesture recognition.

C. Visualization Module

In the visualization module, surface models of hepatic anatomical models, including hepatic artery, hepatic portal vein, hepatic vein and liver parenchyma (Fig. 4) are generated by converting the volume data to each component to a triangulated mesh surface using marching-cube algorithms. The volume data is segmented semi-automatically from computed tomography (CT) images under the guidance of a physician [1], [2]. Compared with traditional slice-by-slice visualization and review techniques, the surgeon can easily recognize the liver geometry and locations of vessels during the surgery from the 3D surface rendering of the anatomical model as shown in Fig. 5. For further details, Please refer [1], [2] for detailed information about CT data and segmented liver and vessel data. The system offers four visualization functions: rotation, zoom in/out, adjustment of opacity, fusion and selection of vessels.



(a) Hepatic artery (b) Hepatic portal vein (c) Hepatic vein (d) Liver

Figure 4. Visualization of liver and its vessels.

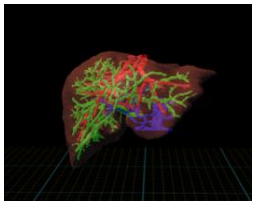


Figure 5. Visualization of fused liver and its vessel structure.

III. Experimental Results

A. Dataset

A novel multi-angle view dataset of hand gesture, named MaHG-RGBD [13], was generated. The ergonometric design of data recorder system is shown in Fig. 6.

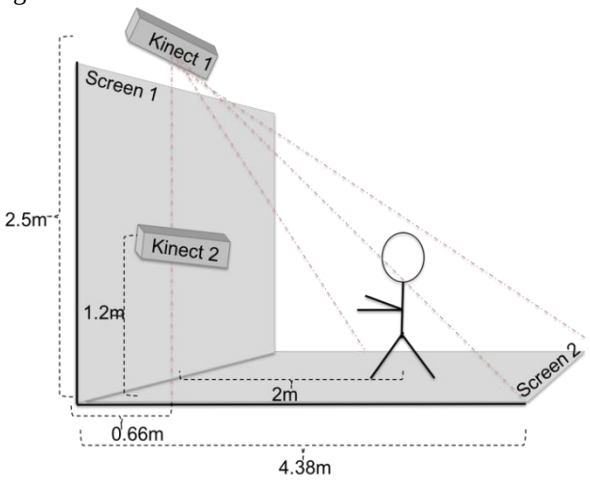


Figure 6. Illustration of the MaHG-RGBD dataset acquisition setup.

The dataset consists of 25 gestures. Since the surgical environment often has limited space and a complicated background, a Kinect sensor titled at an angle of 45 degrees (Kinect 1 in Fig. 6) was used to collect images of the same gestures. Moreover, depth images of the 25 hand gestures were recorded by the tilted Kinect sensor as shown in Fig. 7. Each class was recorded from 15 participants. 100 images were generated per class by repeating the same hand gesture with slight movements. wo modalities (depth and color) were recorded and included in the MaGH-RGBD dataset. Thus, the dataset comprised 150,000 ($2 \times 2 \times 15 \times 25 \times 100$) tuples, each consisting of the depth and color of the hand region. Each hand image was 300 × 300 pixels. For additional information about the dataset, please refer Ref. 13.

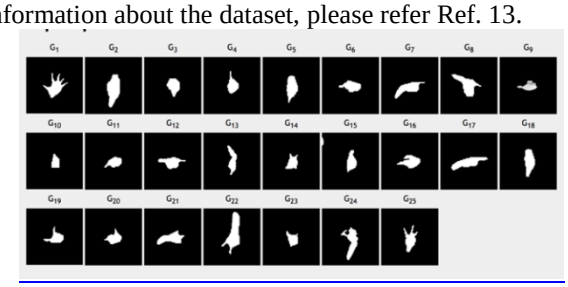


Figure 7. Typical depth images of the 25 hand gesture classes recorded by the tilted Kinect sensor in MaGH-RGBD dataset.

B. Recognition Resultst

15-fold cross-validation was used for validation. Data from 14 participants were used to generate the training data and the data from a single subject was used as testing data. Twenty percent of the training data was selected randomly as the validation set. The validation was repeated 15 times and the recognition results were verified in all cases. The average of the confusion metrices for the AlexNet-based algorithm is shown in Table I. Using the confusion matrices, the $sensitivity_j$, $precision_j$, and $F1_j$ score were calculated for each of the 25 gestures (j=1, 2, …, 25) as shown in Eqs. (1)-(3):

$$sensitivity_j = \frac{TP_j}{TP_j + FN_j} \tag{1}$$

$$precision_j = \frac{TP_j}{TP_j + FP_j} \tag{2}$$

$$F1_j = \frac{2TP_j}{2TP_j + FP_j + FN_j} \tag{3}$$

where TP_j, FN_j and FP_j are the frequencies of true positive, false negative and false positive, respectively, for gesture class j. Since the $F1$ score is an indicator of the classification accuracy of the system, the gestures were ranged in order of decreasing mean $F1$ scores as shown in Fig. 8.

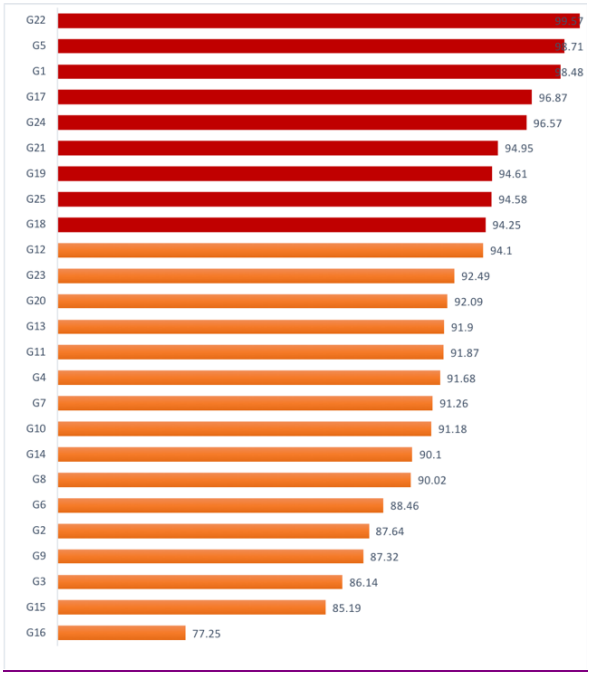


Figure 8. Ranking of gestures for AlexNet. The ranks are based on the F1 scores.

TABLE I. The Average Confusion Matrix of AlexNet for the 25 Gestures. Lager Miss Classification Errors (> 5) Are Indicated in Color and Zero Values and Not Included

Truth	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18	G19	G20	G21	G22	G23	G24	G25
G1	97.7										0.27					0.73			1.07		0.13			0.07	
G2		88.4	0.07	0.27	0.47	0.13			2.8				0.07			2.07		**5.67**						0.07	
G3			90.13	0.2	0.47			3.73	1.8		0.07							1.6					0.07	1.93	
G4			0.27	8.7 0						0.53				0.13	**10.33**				0.93				0.27		0.53
G5			0.13		99.7								0.07					0.07							
G6						86.3					0.27					**12.6**	0.7				0.07				
G7						0.13	96.4	0.07		0.53		1			0.13	0.4	1.27						0.07		
G8		0.13	9.0					87.47	1.07			1.73		0.07									0.53		
G9		5.27	6.73			0.13		0.07	85.67		1.13					0.27		0.53		0.07				0.13	
G10			0.07	0.13						89.87				2.07	7.87										

G11					0.53				1.93		90.13		0.27	0.07		5.53			0.13	1.27		0.13			
G12		4.13				0.2	0.53	2.8				91.47	0.13				0.07	0.13							0.53
G13		0.13			0.6				2.6				92.57		2.2			1.73							0.07
G14			0.07	0.2			0.13			0.07				86.53	0.07								10.87		2.07
G15				1.53				0.07		4.93			3.67	0.27	89.53										
G16		3.2				6.87	2.73			0.13	3.6					77.07				5.73			0.67		
G17						0.13	3.67				0.07	0.13				0.07	95.87				0.07				
G18		0.47			0.6								1.13					97.8							
G19	0.07					0.47			0.2		0.13					0.27			98.27	0.33	0.27				
G20						0.27					0.47								6.53	92.0			0.33	0.4	
G21							7.8		0.07							0.53					91.6				
G22											0.07											99.93			
G23				0.4								0.07		0.8	0.07				0.8	0.4	0.8		96.47		0.2
G24	0.67		2.8		0.2			0.13	0.07															95.8	0.33
G25				0.07						1.07			3.67	2.13											93.07

TABLE II. COMPARISON OF OVERALL PERFORMANCE OF GESTURE RECOGNITION METHODS.

Method	Dataset	Operation 1	Operation 2	Operation 3	Operation 4	Operation 5	Operation 6	Operation 7	Operation 8	Operation 9	Accuracy	time
HOG + SVM [8]	Old dataset [12]	Hand Open	Hand Close	Grasp	Finger Up	Finger Down	Finger Left	Finger Right	Palm Up	Palm down	87.5	0.123s
		91	91	87	89	95	86	84	82	83		
LeNet [10]	Old dataset [12]	Hand Open	Hand Close	Grasp	Finger Up	Finger Down	Finger Left	Finger Right	Palm Up	Palm down	84.3	0.0028s
		86.7	81	84.7	94.1	77.4	85	86.2	84.7	78.9		
AlexNet	New dataset [13]	G22	G5	G1	G17	G24	G21	G19	G25	G18	96.51	0.0069s
		99.57	98.71	98.48	96.87	96.57	94.95	94.61	94.58	94.25		

As well as our previous systems [8], [10], nine gestures were selected to control the visualization operations. However, the previous systems were trained using an old dataset [12] of only nine kinds of gestures. Since the recognition accuracy was low for some gestures, the average accuracy were only about 85%. In the new system, in addition to using a deeper network (AlexNet), the new MaGH-RGBD dataset was used [13] and the nine gestures with the highest recognition accuracies were used (Fig. 8) to control the visualization operations. Table II provides a comparison of the proposed method with the approaches used in pervious works [8], [9] in terms of the recognition accuracy for each gesture and the prediction time. The results show that the new version of the system using new dataset and AlexNet produces better results than the previous versions [8], [10]. The average recognition accuracy was improved to 96.51% and the response rate (or operation rate) was 30fps, which is the same as the with LeNet [9] as the frame rate of the Kinect is 30 fps. However, the processing time with the new method is slightly longer than that with LeNet.

IV. CONCLUSION

In this paper, we proposed a robust hand-gesture-recognition system based on the AlexNet and Kinect for touchless visualization of 3D medical images. Using our system, a surgeon can interact with 3D medical images in real time without touching any devices. This could effectively reduce the risk of contaminating the surgical field and infecting the patient. The recognition accuracy was evaluated for 25 different gestures using deep-learning algorithms and nine of the most robust gestures were selected to control the system. This real-time (30 fps), accurate (96.51% accuracy), and flexible Kinect-based touchless visualization system represents a significant improvement over previous version.

In the future, we plan to combine depth and color image modalities and multi-view images to increasing the robustness of the system.

ACKNOWLEDGMENT

The authors would like to thank Dr. M.Kaibori of KANSAI Medical University for providing medical images and advice on surgical support systems. The authors wish to thank Mr. Seiju Tsujinaga. This work is supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant Nos. 18H03267, 18K11454, 18K18078, 17H00754.

REFERENCES

[1] M. Kaibori, *et al.*, "Novel liver visualization and surgical simulation system," *J. Gastrointest Surg*., vol. 17, pp. 1422-1428, 2013.

[2] T. Tateyama, *et al.*, "Patient-specified 3D-visualization for liver and vascular structures and interactive surgical planning system," *Medical Imaging Technology*, vol. 31, pp. 176-188, 2013 (in Japanese).

[3] E. Nasr-Esfahani, *et al.*, "Hand gesture recognition for contactless device control in operating rooms," arXiv Prepr. arXiv1611.04138. 2016.

[4] L. Gallo, "Controller-free exploration of medical image data: experiencing the Kinect," National Research Council of Italy Institute for High Performance Computing and Networking, 2011.

[5] K. Yoshimitsu, *et al.*, "Development and initial clinical testing of "OPECT": An innovative device for fully intangible control of the intraoperative image-displaying monitor by the surgeon," *Neurosurgery*, suppl. 1, pp. 46-50. 2014.

[6] G. C. Ruppert, *et al.*, "Touchless gesture user interface for interactive image visualization in urological surgery," *World J. Urol.*, vol. 30, pp. 687-691, 2012.

[7] A. Mewes, *et al.*, "Touchless interaction with software in interventional radiology and surgery: A systematic literature review," *Int. J. CARS*, vol. 12, p. 291, 2017.

[8] J. Q. Liu, R. Fujii, T. Tateyama, Y. Iwamoto, and Y. W. Chen, "Kinect-Based gesture recognition for touchless visualization of medical images," *International Journal of Computer and Electrical Engineering*, vol. 9, no. 2, pp. 421-429, 2017.

[9] J. Q. Liu, T. Tateyama, Y. Iwamoto, and Y. W. Chen, "A preliminary study of kinect-based real-time hand gesture interaction system for touchless visualization of hepatic structure in surgery," *Medical Imaging and Information Sciences*, 2018, submitted.

[10] J. Q. Liu, T. Tateyama, Y. Iwamoto, and Y. W. Chen, "Kinect-Based real-time gesture recognition using deep convolutional

neural networks for touchless visualization of hepatic anatomical models in surgery," in *Smart Innovation, Systems and Technologies*, Springer, Cham, 2018, vol. 98.
[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2323, 1998.
[12] R. Fujii, T. Tateyama, T. Kitrungrotsakul, S. Tanaka, and Y. W. Chen, "A touchless visualization system for medical volumes based on Kinect gesture recognition," in *Innovation in Medicine and Healthcare*, Y. W. Chen, *et al*., Eds., Springer, 2016, pp. 209-215.
[13] J. Q. Liu, K. Furusawa, S. Tsujinaga, T. Tateyama, Y. Iwamoto, and Y. W. Chen, "MaHG-RGBD: A multi-angle view hand gesture RGB-D dataset for deep learning based gesture recognition and baseline evaluations," in *Proc. of IEEE ICCE*, 2019, accepted for publication.
[14] A. Krizhevsky, I. Sutsjever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Sys.*, p. 9, 2012.

Jiaqing Liu was born in 1993, received a B.E. degree in 2012 from Northeastern University, China. Now he is a graduate student in Graduate School of Information Science and Engineering, Ritsumeikan University in Japan. His research interests include image processing and analysis, virtual reality, body detection and deep learning.

Kotaro Furusawa was born in 1997, B.A. expected in Information Science 2019 at Ritsumeikan University in Japan. His research interests include image processing, body detection and deep learning.

Tomoko Tateyama received a B.E. degree, a M.E degree in 2003, a Ph.D degree from University of the Ryukyus, Okinawa Japan in 2001, 2003 and 2009, respectively. She was an assistant researcher in 2009-2012, and assistant professor in 2013-2015 at Ritsumeikan University. She is currently an assistant professor in Hiroshima institute of Technology, Hiroshima Japan. Her research interests include Medical image analysis visualization, computer anatomical model, pattern recognitions, computer graphics and vision, virtual reality, development computer aided surgery/diagnosis system. And she is the IEEE SMC and EMBC member.

Yutaro Iwamoto received the B.E. and M.E., and D.E. degree from Ritsumeikan University, Kusatsu in Japan in 2011 and 2013, and 2017, respectively.
He is currently an assistant professor at Ritsumeikan University, Kusatsu, Japan. His current research interests include medical image processing and computer vision, and deep learning.

Yenwei Chen received a B.E. degree in 1985 from Kobe Univ., Kobe, Japan, a M.E. degree in 1987, and a D.E. degree in 1990, both from Osaka University, Osaka, Japan. From 1991 to 1994, he was a research fellow with the Institute for Laser Technology, Osaka. From October 1994 to March 2004, he was an associate professor and a professor with the Department of Electrical and Electronic Engineering, University of the Ryukyus, Okinawa, Japan. He is currently a professor with the College of Information Science and Engineering, Ritsumeikan University, Kyoto, Japan. He is also a chair professor with the College of Computer Science and Technology, Zhejiang University, China. He is an associate editor of International Journal of Image and Graphics (IJIG) and an editorial board member of the International Journal of Knowledge based and Intelligent Engineering Systems. His research interests include pattern recognition, image processing and machine learning. He has published more than 200 research papers in these field.