Hand Detection in UKA Surgery Videos using Deep Convolutional Neural Network

Shadman Sakib[1,3], Belayat Hossain[1], Takafumi Hiranaka[2], Syoji Kobashi[1]
[1]*Graduate School of Engineering, University of Hyogo, 2167 Shosha, 671-2280, JAPAN*
[2]*Department of Orthopedic Surgery, Takatsuki General Hospital, 1-3-13 Kosobe-cho, Takatsuki, Osaka, 569-1192, JAPAN*
[3]*Department of Electrical and Electronic Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh*
sakibshadman15@gmail.com[1,3], belayat.edu@gmail.com[1], takafumi.hiranaka@gmail.com[2], kobashi@eng.u-hyogo.ac.jp[1]

***Abstract*—Through the advancement of surgical procedures, Unicompartmental Knee Arthroplasty (UKA) has contributed to significant clinical results over the last decade. Detection of hand is one of the significant and crucial tasks in Orthopedic Surgery (OS) especially for the purpose of tool handling. While a variety of vision-based approaches were used for the recognition of hand gestures in the field of computer vision, proper detection of the hand in OS is still a difficult challenge in the surgical working environment. This is solely due to the different camera angles, frequent camera movement from the surgeon, high light exposure, inhomogeneous illumination, etc. Therefore, for the proper detection of hand with the presence of tools from UKA surgical video images, we proposed a Deep Convolutional Neural Network (DCNN) model. The primary objective of this paper is to classify hand, nohand, and nonsurgery regions from five UKA surgery videos. The model is trained using Resnet-50 for recognition accuracy. Results demonstrate that the model achieved a classification accuracy of 96%. Moreover, the performance of the model was also evaluated based on precision, recall, and f1-score for each class.**

***Contribution*—A custom knee surgery dataset has been created for the analysis hence a part of the project has been conducted.**

Keywords—Unicompartmental Knee Arthroplasty (UKA); Orthopedic Surgery; Convolutional Neural Network (CNN); Image classification; Surgical video analysis

I. Introduction

Unicompartmental Knee Arthroplasty (UKA) also known as has become one of the promising surgical procedures used to treat femero-tibial degeneration in the unicompartment by replacing the damaged knee parts. It is a promising less invasive surgical option for unicompartmental knee osteoarthritis to relieve pain and obtain function [1]. Over the last two decades, UKA surgery has gained popularity with the developments in implant design and surgical techniques. This reduces the discomfort and has contributed to change in the lives of people suffering from serious knee arthritis. In general, UKA is required for people with medial or lateral osteoarthritis of the knee to improve knee function as well as satisfaction in improving quality of life [2]. Hand detection in UKA is important especially during the tool handling and surgery procedure recognition. With the automatic detection of hand, one can easily understand whether the surgery is taking place or not.

There are some conventional studies based on surgical workflow, phase, and activity recognition but until now, in recent literature, there are no studies specifically based on hand detection in UKA surgery. A novel method for surgical phase recognition has been proposed by Twinanda et. al [3] where they utilized convolutional neural network (CNN) architecture called EndoNet in combination with SVM and Hierarchical hidden markov model (HHMM) in laparoscopic surgery and achieved an accuracy of 92.8%.

In this work, we classify the three UKA surgical environments (hand, nohand, and nonsurgery). The primary purpose is to detect the hand in UKA surgery videos using Residual Neural Network ResNet with 50 layers (ResNet-50) [4], a Deep Convolutional Neural Networks (DCNN). Figure 1 shows the three classes of the image used in this paper.

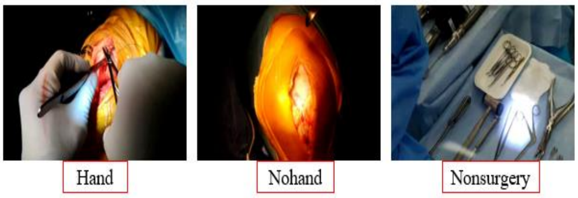

Figure 1. Sample images of three classes from UKA dataset

II. Materials and Methodology

Usually, the UKA surgery follows 27 major steps [5] in which approximately 120 categories of surgical implants are used together with the assembly instruments required during the process. Figure 2 demonstrates the overall framework of our proposed methodology used for classification.

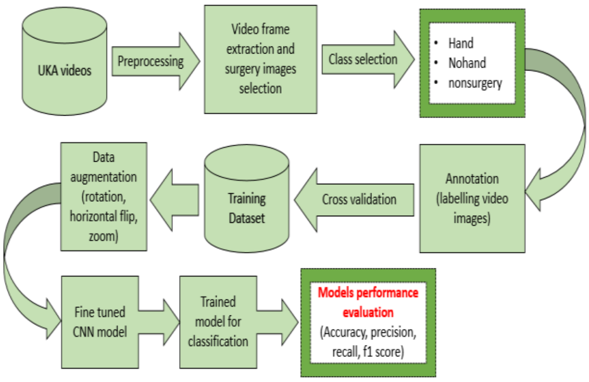
UKA videos
Preprocessing
Video frame extraction and surgery images selection
Class selection
Hand
Nohand
nonsurgery
Annotation (labelling video images)
Cross validation
Training Dataset
Data augmentation (rotation, horizontal flip, zoom)
Fine tuned CNN model
Trained model for classification
Models performance evaluation
(Accuracy, precision, recall, f1 score)

Figure 2. Workflow of the proposed methodology

A. Dataset

We evaluated our model based on our private dataset [6]. The dataset comprises of three categories named as: hand, nohand, and nonsurgery. For our experiment, we have used five cropped and labeled UKA videos. Initially, the videos were at different lengths and frames per second (fps) in mp4 format. The video parameters after cropping are shown in Table I. Three videos were utilized for training purposes

(UKA 4, UKA 5, and UKA 6) and one video (UKA 1) for testing and validation purposes.

TABLE I. UKA VIDEOS AND PARAMETERS AFTER CROPPING

Video no.	UKA Video Name	Length (minutes:seconds)	Frame Rate (fps)	Dimension
1	UKA 1	45:12	30	1280 × 720
2	UKA 4	55:47	30	1280 × 720
3	UKA 5	47:14	30	1280 × 720
4	UKA 6	36:53	30	1280 × 720
5	UKA 7	48:11	30	1280 × 720

B. *Structure of the proposed DCNN model*

For our experiment, we chose a pre-trained network ResNet-50, a DCNN for the purpose of the tool handling action detection. We reshaped all the surgery images to 224 × 224 pixels in size for the input. The pixel values of each training image are normalized within the range of [0-255]. The first layer is the convolutional layer 1 (conv2) comprises of 64 filters with 1 × 1 kernel size. Subsequently, the second and third layer of the network is the conv2, conv3 consists of 128 filters with 1 × 1 kernel size.

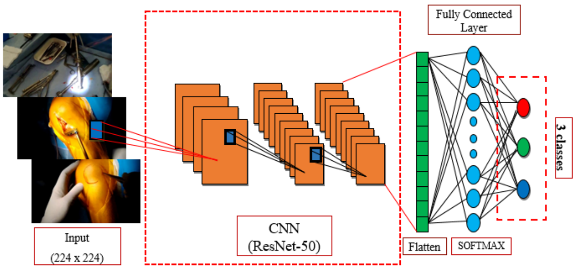

Figure 3. DCNN model for the classification of UKA surgery video images

With each convolution layer, we used rectified linear units (ReLU) as an activation function and batch normalization for improving the performance. Finally, for the final output, we used softmax activation for each of the three classes.

III. RESULTS AND DISCUSSION

In this experiment, we have utilized three UKA videos out of six videos for training our DCNN model and one UKA video for validation purposes. We alienated our training dataset into three classes (hand, nohand, and nonsurgery) for classification, each class with 1000 images. For prediction, we used the validation dataset contain the three classes as mentioned above with 2997 images in total. For the model compilation, for training the network, we used Adam as an optimizer with the initial learning rate of 1e-1 for 60 epochs, mini batch size of 4, and categorical cross-entropy as a loss function.

We evaluated the performance of our trained DCNN model. The matrices we used during the training of the dataset were accuracy, precision, recall, and f1-score. The overall classification performance is presented in the form of a confusion matrix shown in Figure 4. As seen from the slope of the confusion matrix that the "nohand" class achieved the highest classification accuracy of 99%. The lowest accuracy of 89% was observed on “nonsurgery” because it encounters heavy confusion with the areas which includes “hand” as well as “nohand”. Furthermore, from the class “hand” 98% of accuracy is perceived. In addition to the confusion matrix, as our classes are balanced, the performance of the models’ was evaluated by measuring the models overall accuracy.

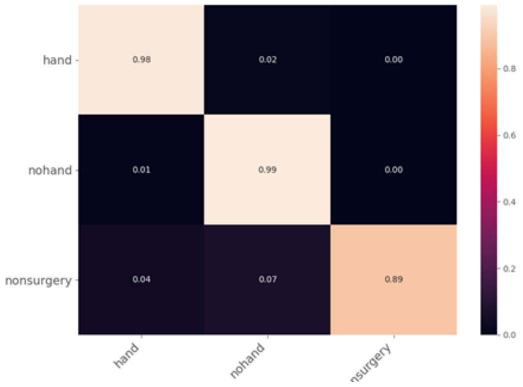

Figure 4. Normalized confusion matrix for the three classes of UKA dataset

Figure 5 represents the accuracy and loss plot of the of our DCNN model throughout the training process. The accuracy of our model on the validation dataset is obtained 96%.

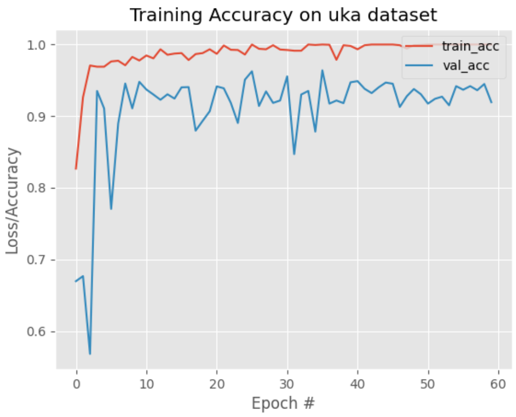

Figure 5. Accuracy plot on UKA dataset

Other matrices (precision, recall, and f1-score) that were derived from the confusion matrix are also calculated and presented in Table II. From Table II, we can say that our model is very efficient in identifying the hand, nohand, and nonsurgery. For the class nonsurgery, it demonstrated the least efficiency with a recall of 0.89.

TABLE II. PERFORMANCE MEASURES ON UKA DATASET

Class	Precision	Recall	F1-score
Hand	0.95	0.98	0.97
Nohand	0.92	0.99	0.96
Nonsurgery	1.00	0.89	0.94

Figure 6 shows the training and validation loss of our model. It can be seen from the figure that the validation loss was at its pick for the early few epochs. Afterward with the increase of the number of epochs, the loss decreases dramatically making the model more stable.

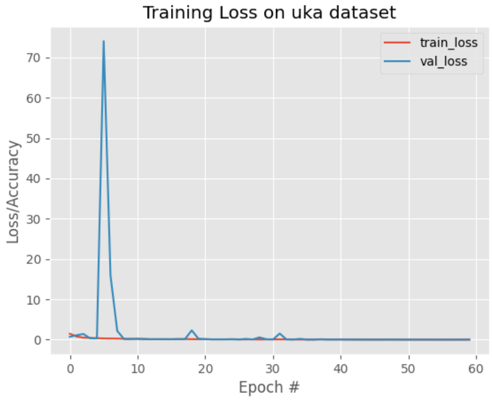

Figure 6. Loss plot on UKA dataset

Nevertheless, it is evident that ResNet-50 can perform reasonably well in the case of the UKA surgery hand detection and classification along with the other surgical environments. The ResNet-50 could classify almost all the classes available in UKA dataset. However, the model got confused whilst classifying the classes "nohand" and "nonsurgery," because of some of the identical regions between both the two classes.

IV. Conclusion

In this work, we successfully presented a method to classify different surgical environments from UKA surgery video images for the purpose of hand detection. For the experiment, five videos were analyzed and annotated with the class labels. for the classification, we applied a deep convolutional neural network such as ResNet with 50 layers on our dataset consisting of 1000 images in class. Results demonstrate that the models' performance with a classification accuracy of 96%. Moreover, the other performance measures for each class were calculated. It was found that the class "hand" has the highest f1-score of 0.97, class "nohand" holds the highest recall of 0.99 and the class "nonsurgery" with the highest precision 1.00. Our future work includes extending this approach to detect hand based on the surgical phase using deep learning. The region of hand will be segmented for a few specific surgery phase hence by preparing the dataset the model will be evaluated based on the desired algorithm.

References

[1] K. Shiwaku, A. Teramoto, S. Nuka, T. Matsumura, K. Watanabe, and T. Yamashita, "Varus kinematics at knee flexion affect clinical outcomes of unicompartmental knee arthroplasty: Intraoperative navigation-based kinematics evaluation," *Asia-Pacific J. Sport. Med. Arthrosc. Rehabil. Technol.*, vol. 20, pp. 6–11, 2020.

[2] D. S. Casper, A. N. Fleischman, P. V. Papas, J. Grossman, G. R. Scuderi, and J. H. Lonner, "Unicompartmental Knee Arthroplasty Provides Significantly Greater Improvement in Function than Total Knee Arthroplasty Despite Equivalent Satisfaction for Isolated Medial Compartment Osteoarthritis," *J. Arthroplasty*, vol. 34, no. 8, pp. 1611–1616, 2019.

[3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] S. Nishio, M. Hossain, B. Hossain, M. Nii, T. Hiranaka, and S. Kobashi, "Real-Time Orthopedic Surgery Procedure Recognition Method with Video Images from Smart Glasses Using Convolutional Neural Network," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 379–384.

[6] B. Hossain, S. Nishio, H. Takafunio, and S. Kobashi, "A deep learning approach for surgical instruments detection in orthopaedic surgery using transfer learning," in *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2020, vol. 11315, p. 113151M.