Hand detection and segmentation in first person images using Mask R-CNN

Sinh-Huy Nguyen[1,4], Thi-Thu-Hong Le[1], Hoang-Bach Nguyen[1], Chi-Thanh Nguyen[1], Thi-Quynh-Tho Chu[2], Thi-Thanh-Huyen Nguyen[2], Hai Vu[3,4]

[1] Institute of Information Technology - AMST, Hanoi, Vietnam
[2] Department of Rehabilitation, Hanoi Medical University Hospital, Hanoi, Vietnam
[3] School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam
[4] International Research Institute MICA, Hanoi University of Science and Technology, Hanoi, Vietnam

Correspondence: Hai Vu, hai.vu@hust.edu.vn
Communication: received 03 September 2021, revised 11 October 2021, accepted 14 October 2021
DOI: 10.32913/mic-ict-research.v2022.n1.995

***Abstract*: In this work, we propose a technique to automatically detect and segment hands-on first-person images of patients in upper limb rehabilitation exercises. The aim is to automate the assessment of the patient's recovery process through rehabilitation exercises. The proposed technique includes the following steps: 1) setting up a wearable camera system and collecting upper extremity rehabilitation exercise data. After filtering, labeling the left and right hand, as well as segmenting the image area of the patient's hand, a dataset consisting of 3700 images with the name RehabHand was built. This dataset is used to train hand detection and segmentation models on first-person images. 2) conducted a survey of automatic hand detection and segmentation models using Mask-RCNN network architecture with different backbones. From the experimental architectures, the Mask-RCNN architecture with the Res2Net backbone was selected for all three tasks: hand detection, left-right hand identification, and hand segmentation. The proposed model has achieved the highest performance in the tests. To overcome the limitation on the amount of training data, we propose to use the transfer learning method along with data enhancement techniques to improve the accuracy of the model. The results of the detection of objects on the test dataset for the left hand is AP = 92.3%, the right hand AP = 91.1%. The segmentation result on the test dataset for the left hand is AP = 88.8%, the right hand being AP = 87%. These results suggest that it is possible to automatically quantify the patient's ability to use their hands during upper extremity rehabilitation.**

Keywords: *First person image, hand segmentation, transfer learning, mask R-CNN.*

I. INTRODUCTION

First Person Vision (FPV) research has received much attention in the research community recently [11]. This field of study is also known as Egocentric vision to emphasize the sensor-centered human element. FPV captures the image directly in front of the camera wearer instead of observing the entire scene like traditional surveillance cameras. Therefore, images obtained from FPV have some key advantages, such as the field of view, especially, hands and objects are in the center of the image. Image data does not include personal information (such as the face, other photo data related to biometric identification) of the person performing the operation, so it does not violate privacy and does not contain personal information in the acquired image data. Currently, wearables cameras such as GoPro Hero 9, Spectacles 2, Microsoft SenseCam used are becoming more and more popular due to the reduced cost of sensors. This has opened up a number of new application directions such as health care for the elderly, people with dementia, or in the field of sports, human-machine interaction, virtual reality, and augmentation reality. In this study, we present some initial research results on the application of FPV in the problem of upper extremity rehabilitation (especially for patients after musculoskeletal surgery, patients recovering after stroke). It is possible to use first-person imaging to evaluate the rehabilitation of patients because the field of view of hands and objects interacting during an exercise is located in the center of the image. We research and test advanced deep learning networks (Mask-RCNN) to solve such tasks as 1) detecting hand regions in images; 2) determine left or right hand; 3) hand segmentation. The results of the first task can be used as input for other problems such as tracking, hand-pose estimation. The second task was to provide information on which hand performed the exercise to compare recovery outcomes between the weak (rehabilitation) and normal hands of the same patient. The third task, hand segmentation, is used in problems that directly assess the recovery process or locate the interaction of hands and objects. Therefore, the study results can

be applied to build automated systems to quantitatively assess the recovery process for patients, helping doctors reduce costs and time and accurately estimate the process of patients' recovery for appropriate treatment.

The two main contributions of the paper include: 1) performing data collection and building a dataset for hand detection and segmentation named **RehabHand**. To our knowledge, this is the first dataset in Vietnam of patient hand images in a series of rehabilitation exercises from a first-person vision and the corresponding hand segment images with ground truth. 2) explore automatic hand segmentation and detection models using Mask-RCNN architecture [5] with different backbones. Based on the evaluation results, we recommend the Mask-RCNN model with the highest performance Res2Net [7] backbone. In the evaluation process, it is proposed to use the transfer learning method, along with some data augmentation techniques to improve the accuracy of the model.

The remainder of this paper is organized as follows: Part II describes relevant studies on FPV or egocentric vision, particularly those involving hand and object partitioning. Part III presents the construction of the dataset **RehabHand**, which is the exercises of patients with upper limb rehabilitation; The proposed method is used to detect and segment hands on the hand segments in first-person images. Part IV presents the model test settings, evaluation metrics, and test results. Finally, section V provides conclusions and directions for future work.

II. RELATED WORKS

Today, research in first-person vision focuses on recognizing actions and objects on images, especially the interaction between hands and objects. A recent survey by Bandini and colleagues [11] has comprehensively and abundantly listed related works and research directions until 2018, such as hand detection, hand segmentation; hand pose estimate; hand-object interaction activity recognition. For example, in order to recognize activities by the sensor wearer, in 2009, Ren et al. [12] laid the first foundation for the study of characterizing the sensor wearer's operation/manipulation through recognition of the object in front of him. There, the authors collected a database of 42 objects that often appear in hand operations in activities of daily living. In the proposed method, these authors used the traditional feature of SIFT and SVM classifier to detect objects. The results achieved an accuracy of only 24%. Then, Castro et al. [13] used convolutional neural networks (CNNs) to learn contextual factors combined with activity time factors to recognize activities by the sensor wearer. The authors compare the performance of Neurons networks with traditional classification methods such as k-NN, Random Forest. The proposed convolutional neural network-based technique achieves higher accuracy.

Regarding first-person databases, there have been several datasets published in the research community recently. Bambach et al. [1] introduced a new hand dataset and a deep learning model for hand detection and segmentation. Their dataset, Ego-Hands, has pixel-level annotations for hands, with two participants in each video interacting with each other. Urooj et al. [2] presented two datasets EgoYouTube – which included first-person videos containing hands in natural environments and HandOverFace with hand-held first-person images along with the appearance of arm-like areas on the image. They fine-tune RefineNet for hand segmentation and uses the hand segment result to recognize the camera wearer's hand gestures. From these surveys, it can be seen that data collected from wearable cameras and related studies published in the community about daily activities is quite common.

However, the problem of hand detection and segmentation in first-person images for rehabilitation applications has not received much attention. During these exercises, the patient wears a camera on the head (forehead) or chest. This camera captures images from a first-person perspective (sensor bearer) to capture patient hand and interactive objects data during exercises. To assess the recovery process, currently, doctors rely on visual observation of the exercise process to assess the recovery ability of the patients' hands. Therefore, once the patient's hand image is segmented during upper extremity rehabilitation exercises, it is feasible to analyze and evaluate their recovery during treatment. Likitlersuang, Jirapat et al. [4] developed a first-person video dataset, "ANS SCI" containing the hands of individuals with spinal cord injury during daily activities at home. This work proposed a deep learning model for hand detection and segmentation based on the Fast R-CNN architecture.

To the best of our knowledge, there are currently no studies on auto-detection, and hand segmentation on the first-person camera recorded images of patients during rehabilitation exercises in Vietnam. This study shows that the application of artificial intelligence can quantitatively calculate the recovery capacity of patients during rehabilitation exercises. The results will help doctors diagnose and make treatment plans for patients, especially given the data and exercises suitable to the characteristics of medical facilities in Vietnam.

III. PROPOSE METHOD

1. RehabHand dataset

The **RehabHand** dataset was collected from rehabilitation exercises of patients at Hanoi Medical University

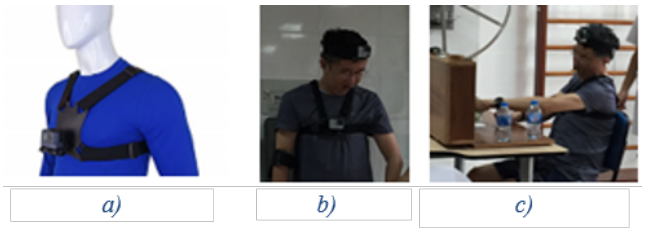


Figure 1. Set up the rehabilitation data acquisition device. a) prototype of the device; b) the patient wears the device; c) actual pictures of patients practicing exercises

Hospital. This dataset consists of frames from the first-person video captured by cameras worn by the patient on the forehead and the chest (Fig. 1). The videos are recorded with a 1080p resolution at 30 frames per second. Data were collected using GoPro Hero4 camera, San Mateo, California, USA.

The camera recorded the exercise of 15 patients performing four upper extremity rehabilitation exercises. Each patient performed each exercise five times. The content of exercises related to grasping objects in different positions is shown in Fig. 3 (next page), as follows:
Exercise 1 (Fig.3a) - practicing with the ball, pick up the round plastic balls with the hand, try to put them in the right holes;
Exercise 2 (Fig.3b) - practice with water bottles, hold a water bottle, pour water into a cup placed on the table;
Exercise 3 (Fig.3c) - practicing with the Cuboid: pick up wooden cubes with the hand, try to put them in the right holes;
Exercise 4 (Fig.3d) - practicing with cylindrical blocks, pick up the cylindrical blocks with hands, try to put them in the right holes;

From the videos obtained (videos mounted on the patient's head, chest), we decomposed into videos of different exercises, 200 video segments in total. In each of these segments, we randomly select image frames to annotate. After filtering and normalizing the data, 3700 RGB images were obtained to prepare for labeling. This selection both ensures randomness and fully represents the images of each patient and exercise. The image extracted from the patient's exercise video is saved and named the **RehabHand** dataset. The images are 1920 x 1440 pixels, with a color depth of 24 bits, and are saved as .png files. We perform annotation of the images, and each pixel in the image will correspond to a label assigned to the left-hand or right-hand, or background region. Fig. 2 illustrates the labeling results of some of the images in this dataset.

In it, the labels of the left and right hands are defined. This makes sense in practice. When detecting and partitioning the hand, it is necessary to determine which hand it is (left, right). Patients are usually weak and need one-handed rehabilitation, while the other hand can be used to evaluate and compare the recovery process (weak hand compared to normal hand).

Figure 2. Hand annotated results (left hand - green; right hand - blue)

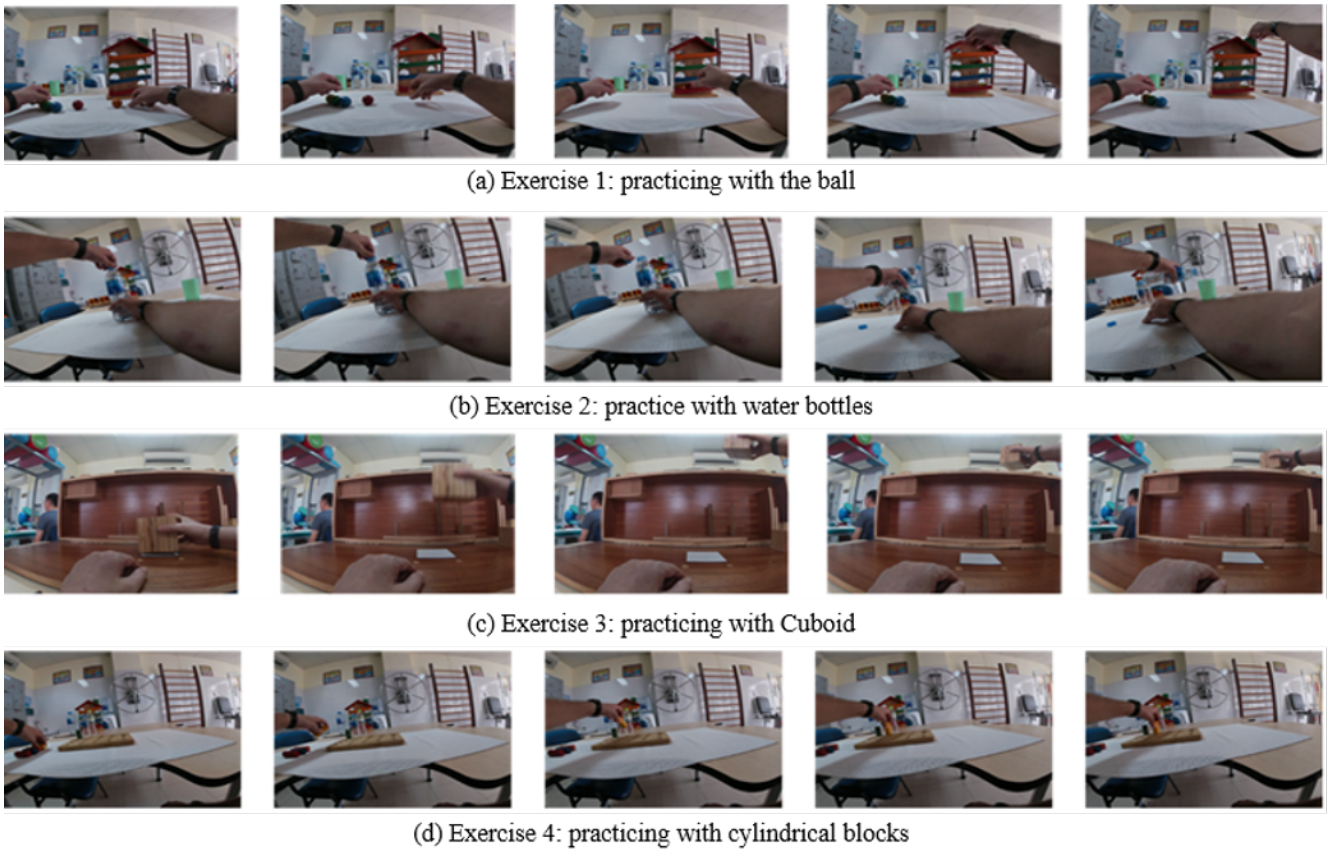
The polygon labeling for the left and right-hand areas is done manually through the VGG Image Annotator (VIA) labeling tool. Labeling results are saved in a .json format file, with the main information being: the name of the label, the filename of the labeled image, the image size, the coordinate information of the labeled areas, the coordinates of the labels. The corresponding point pair (x,y), all pairs of points (x,y) will form a polygon area surrounding the labeled arm, the id of the label (the value "7" is the label assigned to the left hand and "8" is the label assigned to the right hand). This link shares the data: https://drive.google.com/file/d/1hdeGgkTYiHvEHVEptoU8uvCbnQWjjzs/view?usp=sharing.

2. Building a Mask R-CNN network model for hand detection and segmentation

There are many ways to detect and segment hands from first-person images, such as convolutional neural networks that can detect and localize like R-CNNs. Meanwhile, single-stage networks with faster computation time only localize objects of interest such as YOLO, SSD. On the other hand, the approach with two-stage network models such as Mask R-CNN [5] has been introduced because the segmentation result needs to be more accurate than the computational time requirement.

Also, aiming for a technique that can solve all three requirements: detection of hand areas in images; left/right-hand indication; hand segmentation, we propose to build Mask R-CNN model according to the Transfer Learning method, along with evaluating the model's performance with different backbones. As a result, we selected Res2Net [7] as the backbone. Besides, we also propose some data augmentation techniques to improve the accuracy of the model.

Fig. 4 depicts the general diagram of the proposed method, including the following main parts: 1) Data Augmentation; 2) Transfer Learning with Mask R-CNN network



(a) Exercise 1: practicing with the ball

(b) Exercise 2: practice with water bottles

(c) Exercise 3: practicing with Cuboid

(d) Exercise 4: practicing with cylindrical blocks

Figure 3. Illustration of exercises in the RehabHand dataset

for hand detection and hand segmentation; 3) Save the trained model parameters; 4) Predict the hand segment on the test image. The model's outputs include circling the arms according to coordinates and length and width, hand label (left, right), and the resulting hand segmentation in the image (pixel level). These results will be input in related problems such as pose estimate, activity recognition, and measurement of movement-related features in rehabilitation exercises.

a) Data Augmentation

One of the difficulties when building automatic hand detection and segmentation models was the limited data collected for training. Additionally, because the first-person camera images are worn on the head by the patient, they can vary significantly with different environment and head movement conditions. The data augmentation step will expand the image data to cover all differences due to different environmental conditions. With data augmentation techniques, one can increase the amount of training data, thus reducing the overfitting of models. The methods of augmentation used in our work are:

- Shifting by a scale factor between -0.0625 and 0.0625: the image is shifted horizontally and vertically, with the scaling factor being a random value in the range [-0.0625 to 0.0625]. Once the random value is determined, the displacement = [tx, ty] is calculated by the scaling factor * height or width of the corresponding image.
- Increase the brightness of the image (Intensity) with a scale factor in the range [0.1, 0.3]: The contrast of the image changes with the scaling factor being a random value in the range [0.1, 0.3];
- Shift RGB color channel by random value in the range [-10, 10]: change the value of each color channel R, G, B at pixels with random increase/decrease value in the range [-10, 10];
- Shift the HSV color channel by random value: similar to each RGB color channel, the values at each pixel of the H and V channels are increased/decreased by a random value in the range [-20, 20], and for S it is in the range [-30, 30].
- We also perform image blur using Gaussian filter with mean=0, sigma=1.5, and median filter in window area w=3.

Fig. 5 illustrates some results of the applied data augmentation technique.

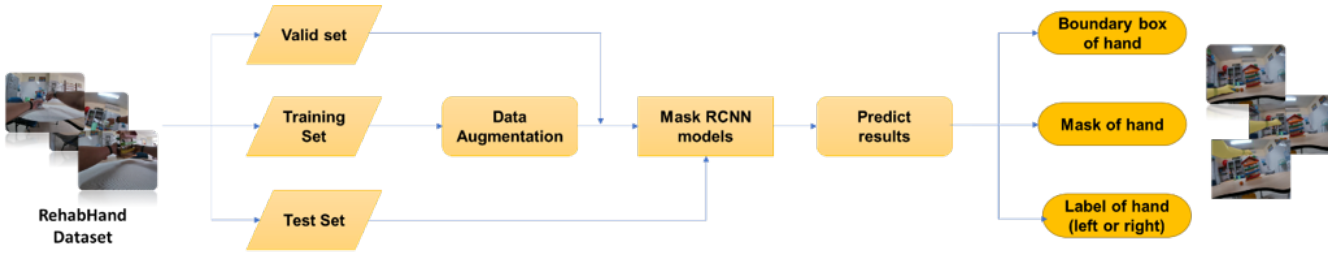


Figure 4. Diagram of the proposed method

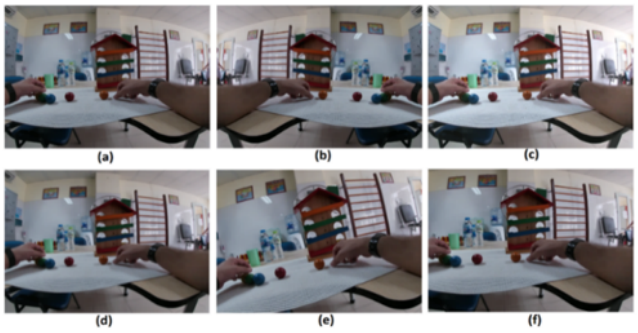


Figure 5. Some examples of data augmentation results

b) The architecture of hand detection and segmentation model

The first-person image segmentation model we use is the Mask R-CNN model. The general architecture of the Mask R-CNN network [4] is illustrated in Fig. 6. Mask-RCNN is essentially an extension of Faster R-CNN [18]. First, Faster R-CNN uses a convolutional neural network to extract feature maps from the input image. Then, the extracted image features will be passed through the RPN (Region Proposal Network) network, returning the borders around the object that can be left or right hand. The ROI pooling layer is used to bring the part of the image that was inside the bounding box in the previous step to the same size. After being brought to the same size, the image parts will be passed through a Fully Connected layer to classify objects. In addition, instead of returning the object's label and bounding box like Faster R-CNN, Mask R-CNN adds an FCN branch to show the segments (masks) of those objects. Compared with Faster R-CNN, Mask R-CNN has an improvement that replaces the RoI Pooling block with RoIAlign block, which increases the accuracy of the model more. It can be seen that Mask R-CNN is a neural network operating in two stages. Stage 1 performs feature extraction from the input image by CNN and uses RPN to give the bounding boxes of the object; Stage 2 performs the classification of the bounding boxes from phase 1, scaling the bounding boxes and creating a mask (for R-CNN Mask). The CNN network for feature extraction of the input image is called the backbone of the model, which is the image layering network that leaves out the last fully connected layer. In addition, Mask R-CNN also uses an additional network of FPN (Feature Pyramid Network) with asymmetric structure with the backbone to obtain features from many different image ratios. In this study, we evaluate the performance of the model with different backbones such as: ResNet[14], ResNeXt[15], HRNet[16], Res2Net[7], RegNet[17]. Detailed results are presented in section IV.4. After evaluation, we choose Res2Net [7] as the backbone to build the Mask R-CNN model.

The rest of our model reuses the entire architecture that the author of the Mask R-CNN model has given and uses a transfer learning method that uses parameters from the model pre-trained on a segmented dataset of objects in the wild COCO [8] as the initialization parameter for the model.

c) Model training

We use transfer learning when training hand detection and segmentation models using pre-trained model parameters on the COCO natural object segmentation dataset [8] as the initialization parameter for the model. The loss functions that are used for the models in this study are:
- L_{cls}: sum of L_{cls1} and L_{cls2}, which is the loss function of the classification. L_{cls1} is the loss calculated in the RPN network and L_{cls2} is calculated in the object detection network and generates the mask of the model objects.
- L_{bbox}: sum of L_{bbox1} and L_{bbox2}, is the loss of predicting object's position and size. L_{bbox1} is the loss calculated in the RPN network and L_{bbox2} is calculated in the object detection network of the model.
- L_{mask}: loss for predicting the object's mask, calculated by the cross-entropy function between the predicted mask and the actual mask of the object.
The final loss function used when training the Mask R-CNN network is: $L = L_{cls} + L_{bbox} + L_{mask}$
To train the image's hand detection and segmentation model, we use the Adam optimization function with momentum=0.9, initial learning rate λ=0.001, weight_decay=0.0001, batch_size=4. The weight of the net-

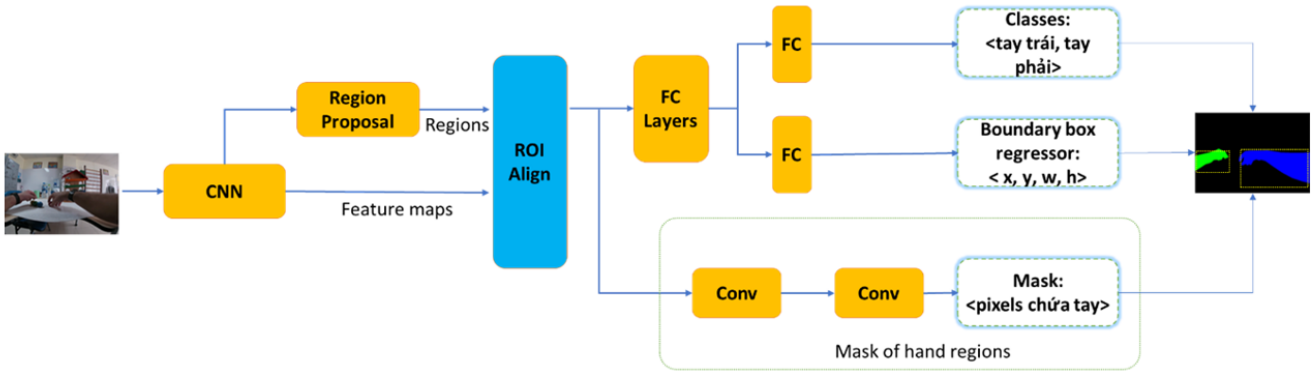


Figure 6. Mask R-CNN Network Architecture

work is pre-trained on the COCO dataset as the initialization parameter for the model. The models are then retrained on the **RehabHand** dataset. The results of the training process are illustrated in Fig. 7. It can be seen that the training converges after 24 epochs. The loss function value decreased from 1.17082 to 0.05016, with the component losses reduced as follows: loss_cls decreased from 0.36345 to 0.00448, loss_bbox 0.13667 decreased from 0.00741, loss_mask decreased from 0.64049 to 0.03752. Our models and algorithms are programmed and trained using the Python programming language and Pytorch library, Tensorflow backend on PC with GeForce GTX 1080 Ti GPU card.

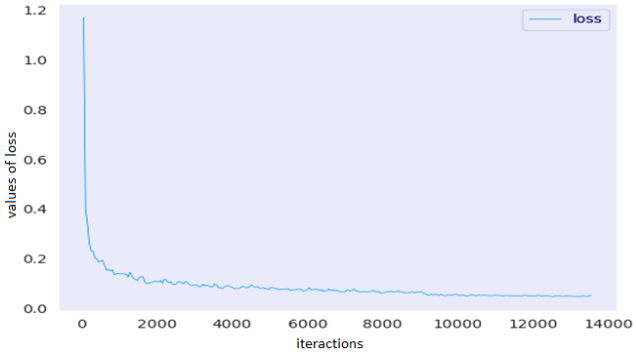


Figure 7. The value of loss function during training

IV. EXPERIEMENT AND RESULTS

1. Datasets

The **RehabHand** dataset was used with 3,700 images of 1924 x 1440 pixels that were fully and accurately labeled with a json file. We have divided this dataset into three separate subsets for training (train set), validation (valid set), and testing (test set) with the ratio of 6:2:2 respectively. There, the number of the train set images is 2220, the number of valid set images is 740, and the number of test set images is 740. Corresponding to the three image sets, there will be 3 json files labeled for the images in each set.

2. Evaluation metrics

In this work, we used AP (Average Precision) [5] as a measure to evaluate the model. The AP is an arithmetic metric that computes the inclusion of both precision and recall. AP is computed by calculating the average interpolation of the precision values over the recall values in [0,1]. We use the interpolation method at 101 recall points which is the method the authors of the dataset for segmentation of the COCO dataset proposed. We also use the AP metric proposed by these authors to evaluate the detection and segmentation models on the COCO dataset, which includes: $AP^{IoU=0.5}$ is the AP value at the threshold of IoU of 0.5, $AP^{IoU=0.75}$ is the AP value at the threshold of IoU of 0.75, and AP is the mean AP value for the threshold of IoU from 0.5 to 0.95 with each increment of 0.05

3. Performance evaluation on CNN pre-trained encoders

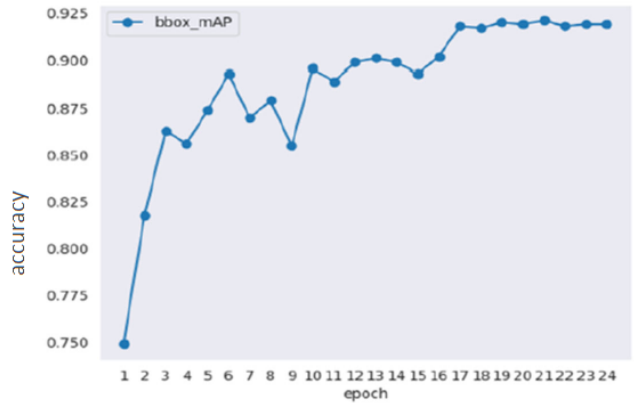
This section reports the image's hand detection and segmentation results using the Mask R-CNN model with different backbones, including ResNet, ResNeXt, HRNet, Res2Net, RegNet, and compares the performance evaluation of the model with different backbones. The experiment was conducted specifically as follows: setting up and training hand detection and segmentation models using Mask R-CNN architecture with different backbones: ResNet101 + FPN, ResNet50 + FPN, Res2Net + FPN, RegNet + FPN, ResNext + FPN and HRNet + FPN, train the models with the same training method described in section III.2. Table I shows that the Mask R-CNN model with Res2Net backbone has the highest results, with AP = 92.3% for the left-hand detection and AP = 91.3% for the right-hand detection.

TABLE I
HAND DETECTION RESULTS WITH DIFFERENT BACKBONE

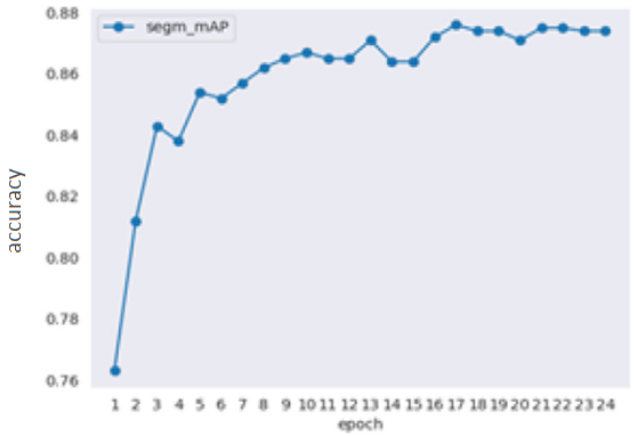
Backbone	$AP^{lefthand}(\%)$	$AP^{righthand}(\%)$	$AP^{IoU=0.50}(\%)$	$AP^{IoU=0.75}(\%)$	AP(%)
ResNet101 + FPN	90.9	89.6	97.3	**96.8**	90.2
ResNet50 + FPN	90.2	88.8	97.5	96.8	89.5
Res2Net + FPN	**92.3**	**91.1**	97.7	96.6	**91.7**
RegNet + FPN	92	90	**97.8**	96.8	91
ResNext + FPN	92.1	91	96.7	**97.8**	91.7
HRNet + FPN	91.6	90.4	97	95.8	91

TABLE II
HAND SEGMENTATION RESULTS WITH DIFFERENT BACKBONE

Backbone	$AP^{lefthand}(\%)$	$AP^{righthand}(\%)$	$AP^{IoU=0.50}(\%)$	$AP^{IoU=0.75}(\%)$	AP(%)
ResNet101 + FPN	88.5	88.9	97.3	**96.7**	87.3
ResNet50 + FPN	88.3	86	97.5	96.9	87.2
Res2Net + FPN	**88.8**	**87**	97.7	97.2	**87.9**
RegNet + FPN	88.6	86.9	**97.8**	97.3	87.7
ResNext + FPN	88.7	86.6	97.3	**97.8**	87.7
HRNet + FPN	88.8	8.62	97	97	87.6



a) Hand detection



b) Hand segmentation

Figure 8. Hand segmentation accuracy on valid set during training

The average detection results are $AP^{IoU=0.50}$ = 97.7%, $AP^{IoU=0.75}$ = 96.6%, AP = 91.7%. Similarly, Table II also shows that the Mask R-CNN model with Res2Net backbone achieved the highest results, the left-hand segment with AP = 92.3%, the right-hand segment with AP = 88.8%, the average segmentation results are $AP^{IoU=0.50}$ = 87%, $AP^{IoU=0.75}$ = 97.7%, AP = 87.9%. Both tables show Mask R-CNN model with Res2Net backbone resulted in the highest performance. Hence we chose the Mask R-CNN architecture with Res2Net backbone as the architecture for our hand detection and segmentation model.

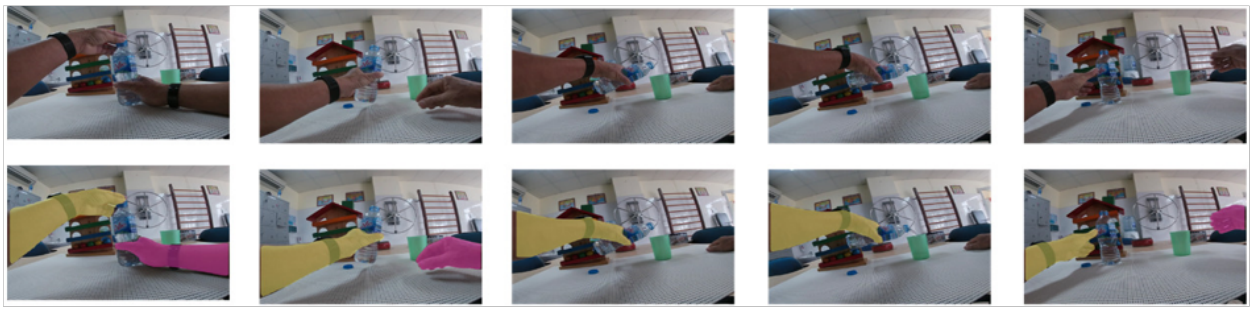
4. Results of hand detection and segmentation using Mask R-CNN architecture with Res2Net backbone

After investigating and evaluating hand detection and segmentation models using Mask R-CNN architectures with different backbones, we select Mask R-CNN with Res2Net backbone as the architecture for hand detection and segmentation model because of the highest results.

Observing the accuracy measurement values in detecting and segmenting objects on the valid set, as illustrated in Fig. 8, shows that the mean AP (mAP) value of the bounding box of the evaluated objects on the valid set increases from 0.749 to 0.919. Similarly, the mean AP value (mAP) of the segment of evaluated objects on the valid set increased from 0.763 to 0.874. Table III is the prediction results on the test data set with each class of two-handed objects.

This table shows that the left-hand detection and segmentation results are better than the right-hand. In addition,

a) Segmentation results on several frames with cylinder block exercise



b) Segmentation results on several frames with cylinder block exercise

Figure 9. Several illustrations of arm segmentation results with different exercises.
In each exercise, the top row is the original image, the bottom row is the segmentation result

TABLE III
PROPOSED MODEL EVALUATION RESULTS ON EACH CLASS

Object	Detection $AP_{bbox}(\%)$	Segmentation $AP_{segm}(\%)$
Left hand	91.1	87
Right hand	**92.3**	**88.8**

we also evaluate the accuracy of the average model on all feature classes in the image, including left-hand, right-hand, and background, and the average results on all layers are shown in Table IV.

TABLE IV
PROPOSED MODEL EVALUATION RESULTS ON ALL CLASS

Bbox/Segment	$AP^{IoU=0.50}$ (%)	$AP^{IoU=0.75}$ (%)	AP (%)
Hand Detection	97.7	96.6	91.7
Hand Segmentation	97.7	97.2	87.9

This table gives us the general results of the model: the average object detection is $AP^{IoU=0.5}$ = 97.7%, $AP^{IoU=75}$ = 96.6%, AP = 91.7% and the average object segment is $AP^{IoU=0.5}$ = 87%, $AP^{IoU=75}$ = 97.7%, AP = 87.9%. Fig. 9 visualizes some examples in rehabilitation exercises. This result shows that although the shape of the hand can change due to the posture and the way the patient performs, the resulting segmentation of the image of the hand was identified (left hand, right hand) and precisely segmented at the pixel level. Using the proposed technique, the segmentation results are almost without over/under segmentation like traditional techniques.

Discussion:

Effects of data augmentation techniques: In order to evaluate the impact of data augmentation techniques, we perform the following tests: first, install and train the data augmentation skipping model. Then install and train the model using data augmentation techniques. Finally, compare the results on the experimental dataset to assess the impact of data augmentation techniques. Table V and Table VI are comparisons of hand detection and segmentation results with or without using data augmentation techniques. These data sheets show that data augmentation contributes significantly to improving model accuracy.

TABLE V
COMPARE THE RESULTS OF THE HAND DETECTION

Data Aug.	$AP^{lefthand}$	$AP^{righthand}$	$AP^{IoU=0.50}$	$AP^{IoU=0.75}$	AP
No	91.7	90.6	96.9	96.4	90.6
Yes	**92.3**	**91.1**	**97.7**	**96.6**	**91.7**

Execution time: Mask R-CNN networks are commonly known to significantly improve computation times compared to traditional R-CNN networks, although they cannot yet reach the speeds of 1-stage networks like SSDs or YOLOs. By reducing the image resolution to 1330 x

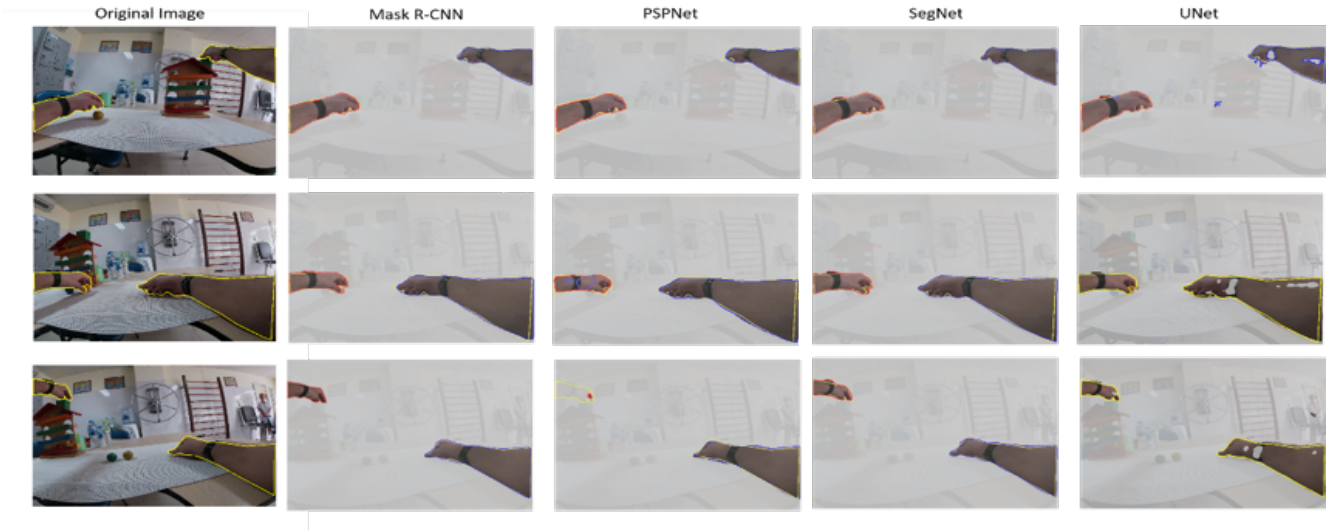


Figure 10. Illustration of some left- and right-hand segmentation results using Mask R-CNN and other comparable networks such as PSPNet, SegNet, and Unet. Column 1: Original image and segmentation results from the ground truth. Column 2: Segment results left hand (red line) and right hand (blue line) using Mask R-CNN network (alpha-blending mode from segmentation results). The yellow line is the ground-truth line. Similarly, columns 3,4,5 are the results of PSPNet, SegNet, and Unet. The color convention of the resulting segmentation contour is the same as for the R-CNN Mask result. The rows represent images at each different hand position (both left and right hand).

TABLE VI
COMPARE THE RESULTS OF THE HAND SEGMENTATION

Data Aug.	$AP^{lefthand}$	$AP^{righthand}$	$AP^{IoU=0.50}$	$AP^{IoU=0.75}$	AP
No	88.2	86.3	97.4	96.5	87.2
Yes	**88.8**	**87.0**	**97.7**	**97.2**	**87.9**

800 pixels, the calculation time can reach three fps. The image size was reduced without loss of details to achieve acceptable computation time.

Compare the results of Mask R-CNN and some popular image segmentation deep learning networks: We install three deep learning networks of object segmentation to compare with Mask R-CNN network, including SegNet[19], Unet[20], and PSPNet [21]. These models are trained and tested using the **RehabHand** dataset with the same training and testing data set as the settings with the Mask R-CNN model mentioned above. Specifically, the data set is divided into three separate subsets for training (train set), monitoring (valid set), and testing (test set) with the respective ratio of 6:2:2. Table VII shows the results of hand object segmentation with different deep learning networks. This table shows that Mask R-CNN achieved the highest accuracy with IoU for the left hand of 89.52% and the right hand of 91.87%. In addition, Fig. 10 illustrates the results of Mask R-CNN with the compared networks of SegNet, Unet and PSPNet with some image data from the **RehabHand** dataset. The illustrations show that Mask R-CNN gives the highest segmentation accuracy results compared to the compared networks. Mask R-CNN gives stable and accurate segmentation results in most cases. Meanwhile, Unet and PSPNet often suffer from under-segmentation or confusion between the left and right-hand regions. SegNet also gives good results in almost cases but often missing segmentation at the level of detail such as fingertips or arm parts.

In this paper, we have built a dataset for rehabilitation patients. These are people who are undergoing rehabilitation treatment after injury or stroke. To our knowledge, datasets similar to **RehabHand** have not been published. The patient's hand activity is different from other popular data sets such as EgoGesture [22], EgoHands [23], EGTEA [24]. These are databases collected from ordinary people with data collecting daily activities, activities in the kitchen. The performance evaluation of networks based on R-CNN architecture with databases obtained from the first-person camera can be referred to in the article [11]. The results also show that R-CNN networks have also achieved good results with data sets obtained from first-person cameras such as EgoGesture, EgoHands, and EGTEA. Although these datasets have variations in light intensity and have

TABLE VII
HAND SEGMENT WITH DIFFERENT NEURON NETWORKS

Network	**IoU**^{lefthand}**(%)**	**IoU**^{righthand}**(%)**
SegNet	80.95	83.65
Unet	75.57	79.65
PSPNet	67.58	80.02
MaskRCNN	**89.52**	**91.87**

complex backgrounds with diverse types of hand activity in daily life.

V. CONCLUSION

We have presented initial studies on artificial intelligence in the assessment of the rehabilitation process as a prerequisite for carrying out further studies. Collection and segmentation labeling, hand object detection totaling 3700 images from first-person videos of patients in upper extremity rehabilitation exercise and used this dataset for models training and pilotting. The dataset is shared within the research community. Propose a transfer learning method based on the Mask-RCNN model to detect and segment hand objects on first-person images of patients in upper limb rehabilitation exercises. After model implementation, training, and evaluation based on Mask-RCNN architecture with different recently announced backbones including ResNeXt, HRNet, Res2Net, ResNet different backbones including Resnet50, Resnet101, Res2Net, we choose Mask-RCNN architecture with Res2Net + FPN backbone as architecture due to best performance and high accuracy results in hand detection and segmentation. The left-hand detection result is AP = 92.3%, the right-hand is AP = 91.1%. The result of left-hand segmentation is AP = 88.8%, and the right-hand is AP = 87%. Although the obtained results are reliable and stable, the Mask-RCNN with backbone architecture Res2Net has to learn a rather large number of parameters and requires a lot of computational resources. The next studies are expected to research other deep learning techniques to improve the model so that the model can be deployed on systems with limited computational resources. We will also continue to study the features that can be extracted from the study results to provide quantitative results in the recovery process of patients participating in upper extremity rehabilitation exercises.

ACKNOWLEDGMENTS

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2017.315.

REFERENCES

[1] S. Bambach, S. Lee, D. Crandall, and C. Yu. Lending, “Detecting hands and recognizing activities in complex egocentric interactions”, *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[2] Urooj, Aisha, and Ali Borji, "Analysis of hand segmentation in the wild", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Likitlersuang, Jirapat, et al, "Egocentric video: a new tool for capturing hand use of individuals with spinal cord injury at home", *Journal of neuro engineering and rehabilitation*, No.16.1, p.83, 2019.

[4] He, Kaiming, et al, "Mask r-cnn", *Proceedings of the IEEE international conference on computer vision*, 2017.

[5] Lin, Tsung-Yi, et al, "Microsoft coco: Common objects in context", *European conference on computer vision. Springer, Cham*, 2014.

[6] He, Kaiming, et al, "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[7] Xie, Saining, et al, "Aggregated residual transformations for deep neural networks", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[8] Gao, Shanghua, et al, "Res2net: A new multi-scale backbone architecture", *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[9] Wang, Jingdong, et al, "Deep high-resolution representation learning for visual recognition", *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[10] Radosavovic, Ilija, et al, "Designing network design spaces", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Andrea Bandini, José Zariffa, “Analysis of the hands in egocentric vision: A survey”, *IEEE Transaction on Pattern Analysis and Machine Interaction*, 2020.

[12] X. Ren and M. Philipose, “Egocentric recognition of handled objects: Benchmark and analysis”, *Computer Vision and Pattern Recognition Workshops 2009, IEEE Computer Society Conference on. IEEE, 2009*, pp.1–8, 2009.

[13] D. Castro, et al, "Predicting Daily Activities from Egocentric Images Using Deep Learning", *Proceedings of the 2015 ACM International Symposium on Wearable Computers*,2015.

[14] He, Kaiming, et al, "Deep residual learning for image recognition", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[15] Xie, Saining, et al, "Aggregated residual transformations for deep neural networks", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[16] Wang, Jingdong, et al, "Deep high-resolution representation learning for visual recognition", *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[17] Radosavovic, Ilija, et al, "Designing network design spaces", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[18] Shaoqing Ren, et al, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, *Proceedings of NIPS*, 2015

[19] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", *IEEE transactions on pattern analysis and machine intelligence*, pp.2481-2495, 2017.

[20] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation", *International Conference on Medical image computing and computer-assisted intervention. Springer, Cham*, 2015.

[21] Zhao, Hengshuang, et al, "Pyramid scene parsing network", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[22] Yifan Zhang, Congqi Cao, Jian Cheng, Hanqing Lu, “EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition”, *IEEE Transactions on Multimedia*, Volume 20, Issue 5, May 2018.

[23] Bambach, Sven and Lee, Stefan and Crandall, David J. and Yu, Chen, “Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions”, *Proceeding of The IEEE International Conference on Com-*

puter Vision (ICCV), 2015.
[24] Alireza Fathi, Xiaofeng Ren, James M. Rehg, "Learning to Recognize Objects in Egocentric Activities", *Proceeding of Computer Vision and Pattern Recognition (CVPR)*, 2011

Sinh-Huy Nguyen was born in 1976 in Thai Binh. The author graduated from the Military Technical Academy in 2000 and received his Master's degree in 2008 also from the Military Technical Academy. He is currently a researcher and PhD student at the Institute of Information Technology, AMST, Hanoi, Vietnam. The author's research area is image processing, machine vision and recognition.
Email: huyns76@gmail.com

Thi-Thu-Hong Le was born in 1980 in Thanh Hoa. The author graduated from Hanoi University of Science and Technology in 2003 and received a Master's degree in 2010 from the Military Technical Academy. He is currently a researcher and PhD student at the Institute of Information Technology, AMST, Hanoi, Vietnam. The author's research area is image processing, machine vision and recognition.
Email: lethithuhong1302@gmail.com

Hoang-Bach Nguyen was born in 1984 in Thai Binh. The author graduated from the Military Technical Academy in 2007 and received a Master's degree in 2020 also from the Military Technical Academy. He is currently a researcher at the Institute of Information Technology, AMST, Hanoi, Vietnam. The author's research area is image processing, machine vision and recognition.
Email: nhbach2203@gmail.com

Chi-Thanh Nguyen received the Ph.D. degree in Computer Science from Nagaoka University of Technology, Japan in 2012. He is currently a researcher at the Institute of Information Technology, AMST, Hanoi, Vietnam. His research interest includes deep learning, computer vision, medical image analysis, and natural language processing.
Email: thanhnc80@gmail.com

Thi-Quynh-Tho Chu was born in 1989 in Hanoi. The author graduated as a General Doctor in 2013, graduated as a Resident Doctor of Rehabilitation in 2017 at Hanoi Medical University. Currently, the author is a Lecturer at the Department of Rehabilitation at Hanoi Medical University, a doctor at the Rehabilitation Department of Hanoi Medical University Hospital.
Email: chuthiquynhtho.hmu@gmail.com

Thi-Thanh-Huyen Nguyen was born in 1973 in Hai Phong, graduated as a General Doctor from Hanoi Medical University in 1996, received the degree of Resident Doctor, Level I Specialist in Rehabilitation in 2000, received Doctor of Medicine Level II majoring in Rehabilitation in 2019. Currently Head of Rehabilitation Department, Hanoi Medical University Hospital, lecturer in Rehabilitation Department, Hanoi Medical University. Member of Vietnam Rehabilitation Association. Member of the International Association for the Study of Pain (IASP). The author's research area is: rehabilitation in neurological, musculoskeletal, surgery, geriatric and pain treatment.
Email: huyen.rehab@gmail.com

Hai Vu received a PhD in Computer Science from Osaka University, Japan, 2009. Currently, the author is a lecturer at the Institute of Electronics and Telecommunications, and a research fellow in the Department of Vision. computer, MICA Institute of International Studies, Hanoi University of Science and Technology. The author's research areas of interest include diagnostic-assisted medical image analysis, computer vision, and identification.
Email: hai.vu@hust.edu.vn