Hand action recognition in rehabilitation exercise method using R(2+1)D deep learning network and interactive object information

Nguyen Sinh Huy[1*], Le Thi Thu Hong[1], Nguyen Hoang Bach[1], Nguyen Chi Thanh[1], Doan Quang Tu[1], Truong Van Minh[2], Vu Hai[2]

[1]Institute of Information Technology/Academy of Military Science and Technology;
[2]School of Electronics and Electrical Engineering (SEEE)/Ha Noi University of Science and Technology;
[*]Corresponding author: huyns76@gmail.com
Received 08 Sep 2022; Revised 30 Nov 2022; Accepted 15 Dec 2022; Published 30 Dec 2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.CSCE6.2022.77-91

ABSTRACT

Hand action recognition in rehabilitation exercises is to automatically recognize what exercises the patient has done. This is an important step in an AI system to assist doctors in handling, monitoring and assessing the patient's rehabilitation. The expected system uses videos obtained from the patient's body-worn camera to recognize hand action automatically. In this paper, we propose a model to recognize the patient's hand action in rehabilitation exercises, which is a combination of the results of a deep learning network recognizing actions on Video RGB, R(2+1)D, and a main interactive object in the exercise detection algorithm. The proposed model is implemented, trained, and tested on a dataset of rehabilitation exercises collected from wearable cameras of patients. The experimental results show that the accuracy in exercise recognition is practicable, averaging 88.43% on the test data independent of the training data. The action recognition results of the proposed method outperform the results of a single R(2+1)D network. Furthermore, better results show a reduced rate of confusion between exercises with similar hand gestures. They also prove that the combination of interactive object information and action recognition improves accuracy significantly.

Keywords: Hand action recognition; Rehabilitation exercises; Object detection and tracking; R(2+1)D.

1. INTRODUCTION

Physical rehabilitation exercises aim to restore the body's functions and toward the improvement in life quality for patients who have a lower level of physical activity and cognitive health worries. A rehabilitation program offers a board of activities, including controlling muscle, gaiting (walking) and balancing, improving limb movement, reducing weakness, addressing pain and other complications, and so on. In this study, the rehabilitation focuses on physical exercises that are designed to manage the functional hand or upper extremity of patients who undergo clinical treatments for catastrophic disease, disc herniation, trauma, or accidental fractures. The main objectives are to take advantage of artificial intelligence (AI) to help GPs handle, monitor and assess the patient's rehabilitation. The final goal tends to support the patients conventionally performing their physical therapy at home. In a usual clinical setting, patients follow exercises given by technical doctors, which play an essential role in rehabilitation therapy. However, it is challenging to quantify scores because technical doctors usually observe and assess with their naked eyes and experiences. In the absence of clinical assistant tools, evaluation performances of the rehabilitation therapy are time-consuming and prevent patients from deploying the rehabilitation routines in their usual environment or accommodations. To address these issues, in this study, we deploy a wearable first-person camera and other wearable sensors, such as accelerometers and

gyroscopes, to monitor the uses of functional hands in physical rehabilitation therapy (exercises). Patients are required to wear two cameras on their forehead and chest. The cameras capture all their hand movements during the exercises and record sequences regardless of duration. A patient participates in four of the most basic climb rehabilitation exercises. Each exercise is repeated at a different frequency. Figure 1 illustrates the four rehabilitation exercises.

- Exercise 1 - practicing with the ball: pick up round plastic balls with hands and put them into the right holes.

- Exercise 2 - practicing with water bottles: hold a water bottle and pour water into a cup placed on the table.

- Exercise 3 - practicing with wooden blocks: pick up wooden cubes with hands and try to put them into the right holes.

- Exercise 4 - practicing with cylindrical blocks: pick up the cylindrical blocks with hands and put them into the right holes.

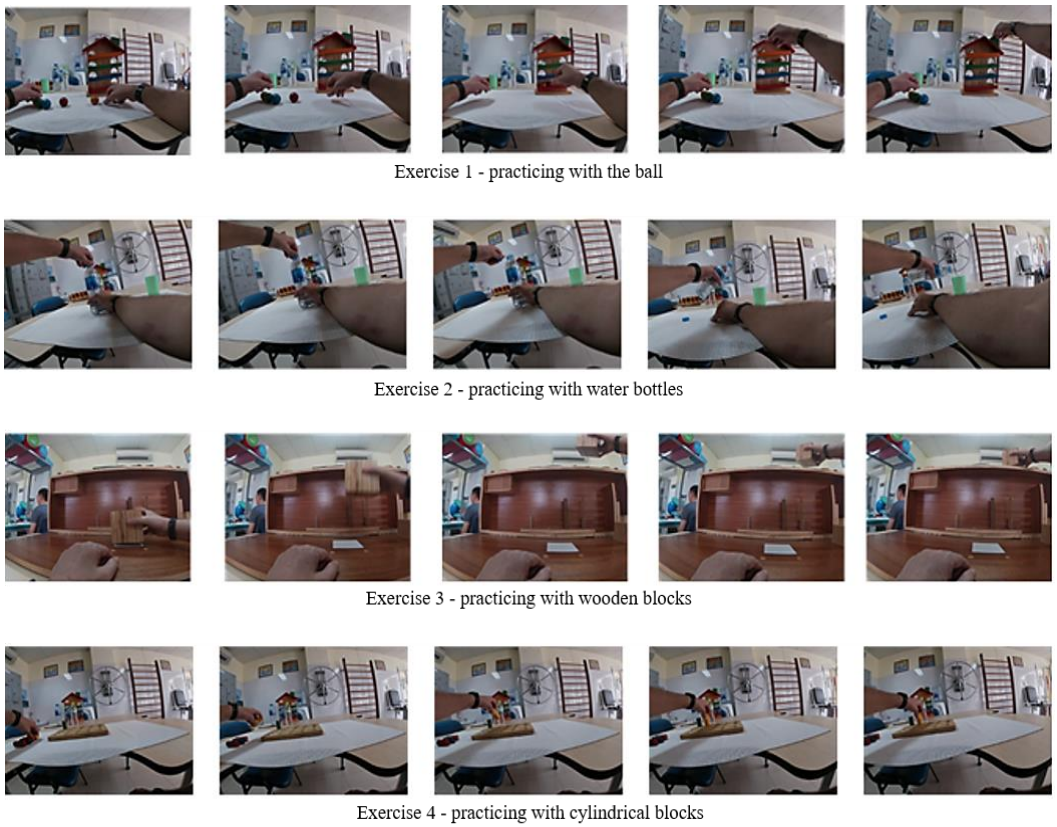

Figure 1. *Examples of rehabilitation exercises.*

The automatic recognization of what rehabilitation exercises patients have done, their ability to practice these exercises and their recovery levels will help doctors and nurses to provide the most appropriate treatment plan for them. Wearable cameras will record exactly what is in front of the patients. Camera movement is guided by the wearer's activity and attention. Interacted objects tend to appear in the center of the frame. Hand occlusion is minimized. Hands and exercise objects are the most important indicators for

recognizing the patients' exercises. However, recognizing a patient's exercise from the first-person video is more difficult than recognizing the action from the third-person video because the patient's pose cannot be estimated when they are wearing the camera. Moreover, the sharp change in the viewpoint makes any kind of tracking method infeasible in implementation, so it is difficult to apply third person action recognition algorithms.

The importance of egocentric cues for the first-person action recognition problem has attracted much attention in academic research. In the last few years, several features based on egocentric cues, including gaze, the motion of hands and head, and hand pose, have been suggested for first person action recognition [1-4]. Object centric approaches introducing methods to capture changing appearance of objects in the egocentric video have been proposed [5, 6]. However, the features are manually tuned in these instances, and they are performed reasonably well only for limited, targeted datasets. There have been no studies in the direction of extracting egocentric features for action recognition on egocentric videos of rehabilitation exercises. Hence, in this paper, we propose a method to recognize the patient's hand action in the egocentric video of rehabilitation exercises. The proposed method is based on the observation that the rehabilitation exercises of patients are characterized by the patient's hand gestures and interactive objects. Table 1 shows a list of exercises and corresponding types of interactive objects in the exercises. Based on this observation, we propose a rehabilitation exercises recognition method which is a combination of R(2+1)D [7], RGB video-based action recognition deep learning network and interactive object type detection algorithm.

Table 1. *List of exercises and corresponding exercise objects.*

	Exercise	**Interactive object**
1	Exercise 1	Ball
2	Exercise 2	Water bottle
3	Exercise 3	Wooden cube
4	Exercise 4	cylindrical block

The remaining of the paper is organized as follows. Section II describes the proposed method for a rehabilitation exercise recognition. Section III presents experimental results and discussions. Section IV concludes the proposed method and suggests improvements for future research.

2. PROPOSED METHOD

2.1. Overview of the proposed method

In this study, we propose a model to recognize patient's rehabilitation exercises in videos obtained from the patient's body-worn camera. In the proposed model, a R(2+1)D deep learning network for RGB video-based action recognition is used to recognize the hand action. The results of the R(2+1)D network are then combined with the results of identifying the main interactive objects in the exercise to accurately determine the exercise that the patient performs. The Pseudo code of the proposed method is presented in figure 2.

An overview of the proposed model is depicted in figure 3. The model includes the main components as follows:

- R(2+1)D network for hand action recognition on RGB videos.
- Module for determining the type of interactive object in the exercise.
- Module for combining hand activity recognition results and interactive object type to define exercises

Algorithm 1 Pseudo code of rehabilitation exercises recognition

$\{f_t\}$ ← Sequence of consecutive frames
$\{OBJ\_bbox\}$ ← Bounding boxs of objects in a frame
$\{Hand\_bbox\}$ ← Bounding box of patient's hand in a frame
$\{Prob\_Recognize\}$ ← Probability of exercises performed in a video
$\{Score\}$ ← Relative position evaluation score between the hand and the object
Input $\{f_t\}$ to *R(2+1)D model* get $\{Prob\_Recognize\}$
for each frame in $\{f_t\}$ **do**
    Input f_t to Object detector get $\{OBJ\_bbox\}$
end for
Input $\{f_t\}$ *to* Hand tracker get $\{Hand\_bbox\}$
Comparing hand position and detecting objects in the frames, caculating $\{Score\}$ by E.q. (4)
for each Exercise
    Caculate scores of the Exercises by E.q.(5)
end for
Select exercise performed in the sequence of frames with max score

Figure 2. *Pseudo code of rehabilitation exercises recognition.*

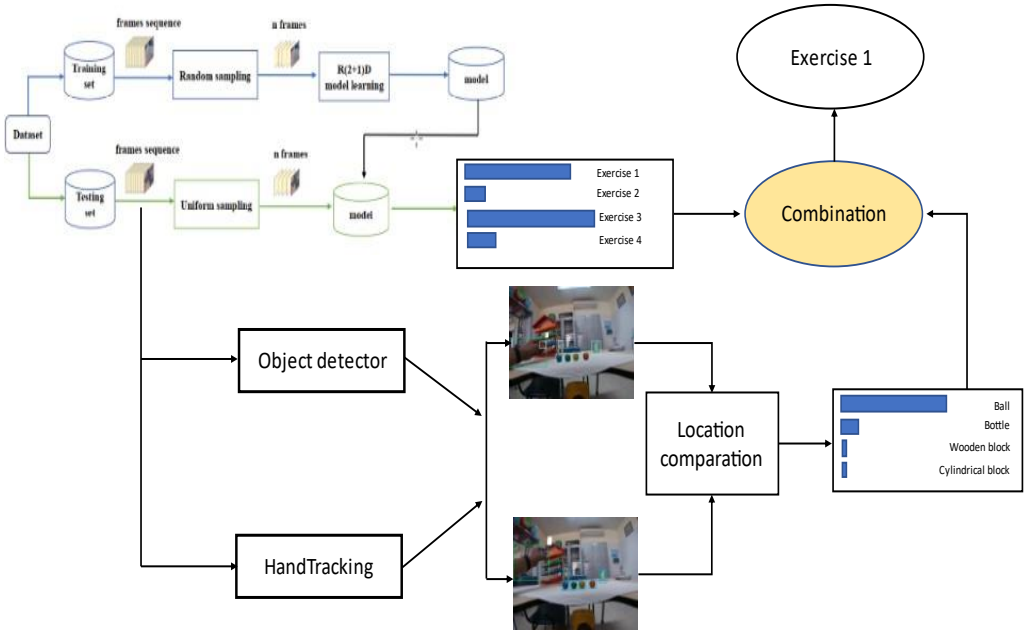

Figure 3. *Rehabilitation exercises recognition model.*

2.2. R(2+1)D network for hand action recognition.

Deep learning models have achieved many successes in image processing and action

recognition problems. In this study, we propose to use R(2+1)D deep learning network to recognize patient hand action in a rehabilitation exercise video. The R(2+1) D convolutional neural network is a deep learning network for action recognition which implements R(2+1) convolutions inspired by the 3D ResNet architecture [8]. The use of (2+1)D convolutions compared to conventional 3D convolutions reduces computational complexity, avoids overfitting, and gives more nonlinear points allowing better modelling of functional relations. The R(2+1)D network architecture is shown in figure 4. The R(2 + 1)D network performs a separation of time and space dimensions, replacing the 3D convolution filter of size (t × d × d) with a block (2 + 1)D consisting of a 2D spatial convolution filter of size (1 × d × d) and a 1D time filter of size (t × 1 × 1) (figure 5).

Layer name	Output size	18-layer	34-layer
conv1	L x 56 x 56	1 x 7 x 7, 45 3 x 1 x 1, 64	
conv2_x	L x 56 x 56	[1 x 3 x 3, 64 3 x 1 x 1, 64 3 x 1 x 1, 64 3 x 1 x 1, 64] x 2	[1 x 3 x 3, 64 3 x 1 x 1, 64 1 x 3 x 3, 64 3 x 1 x 1, 64] x 3
conv3_x	$\frac{L}{2}$ x 28 x 28	[1 x 3 x 3, 128 3 x 1 x 1, 128 1 x 3 x 3, 128 3 x 1 x 1, 128] x 2	[1 x 3 x 3, 128 3 x 1 x 1, 128 1 x 3 x 3, 128 3 x 1 x 1, 128] x 4
conv4_x	$\frac{L}{4}$ x 14 x 14	[1 x 3 x 3, 256 3 x 1 x 1, 256 1 x 3 x 3, 256 3 x 1 x 1, 256] x 2	[1 x 3 x 3, 256 3 x 1 x 1, 256 1 x 3 x 3, 256 3 x 1 x 1, 256] x 6
conv_5x	$\frac{L}{8}$ x 7 x 7	[1 x 3 x 3, 512 3 x 1 x 1, 512 1 x 3 x 3, 512 3 x 1 x 1, 512] x 2	[1 x 3 x 3, 512 3 x 1 x 1, 512 1 x 3 x 3, 512 3 x 1 x 1, 512] x 3
	1 x 1 x 1	Pooling, fully connected, softmax	

Figure 4. *R(2+1)D network architecture.*

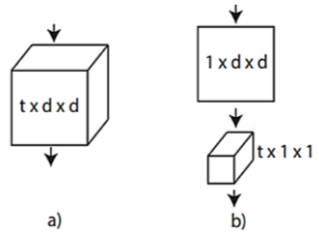

Figure 5. *a) 3D convolution filter and b) (2+1)D convolution filter.*

The framework for the patient's hand action in the rehabilitation exercise based on the R(2+1)D network is presented in Figure 6. The framework includes the following steps:

- Step 1: Collecting rehabilitation exercise video data;
- Step 2: Labeling and dividing data into a training set and test set;

- Step 3: Preprocessing data and training model with training dataset;
- Step 4: Evaluating the accuracy of the model with the test set.

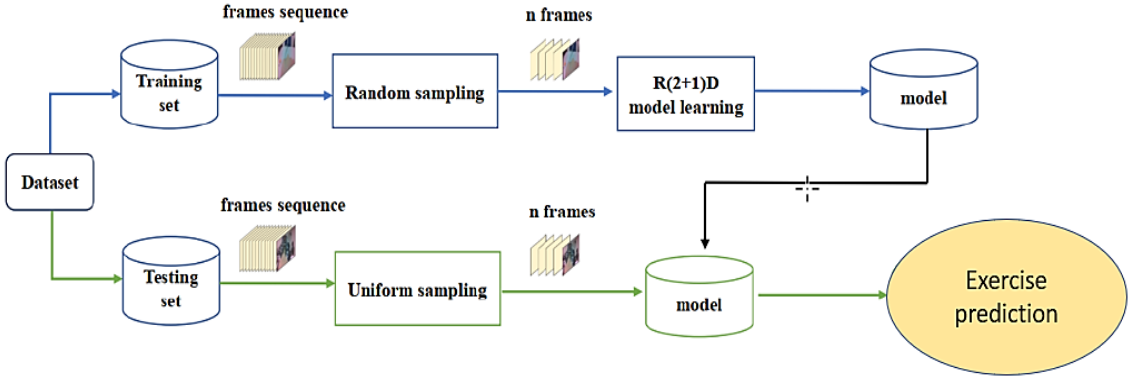

Figure 6. *Hand action recognition framework using R(2+1)D network.*

Data preprocessing method

Because there are differences in each activity and in the duration of the patient's activities in the exercises, the duration of each exercise video varies from patient to patient. Frames are densely captured in the video, but the content does not change much, so we propose using the segment-based sampling method introduced in [9]. This method is an all-video and sparse type of sampling. This method has the advantage of eliminating the duration limitation because of sampling over the entire video. It helps to incorporate the video's long-range timing information into model training. The method is suitable for collecting rehabilitation exercise video data, which overcomes the disadvantages of different times of each exercise segment. The sampling process is as follows:

Step 1: Dividing segments

The exercise video is made of many consecutive frames, so we partition each video into a set of frames at 30 fps (30 frames per second). All frames of the video are divided into equal intervals (figure 7)

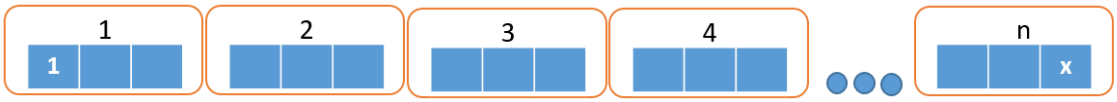

Figure 7. *Dividing segments.*
x is the total number of frames of the video; n is the number of segments we want to get.

Step 2: Selecting frames from segments

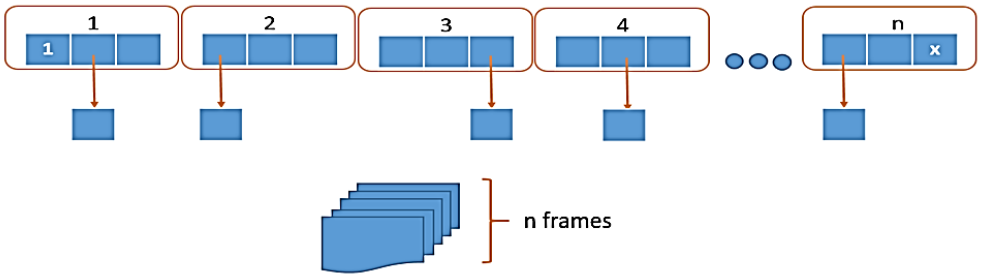

Figure 8. *Random sampling.*

- Trainning data: Randomize one frame in each segment to form a sequence of n frames. This helps the training data to be more diverse because after each time the model is trained, it can learn different features (figure 8)

- Testing data: Take a frame in the center of each segment to evaluate results (figure 9)

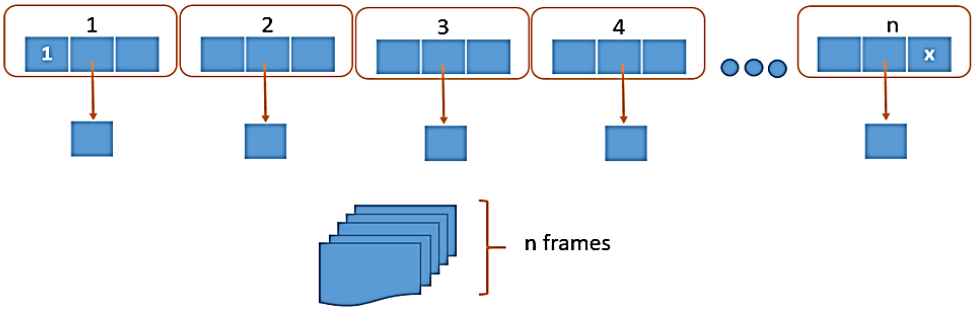

Figure 9. *Sampling at the center of each segment.*

The number of frames taken in each video is the power of 2 to fit the recognition model, and exercise videos dataset with the duration of the exercise video is from 1.5 s - 3.5 s, equivalent to 45 -105 frames. The consecutive frames of a video do not make a big difference in content, so we use n = 16 and resize the frames to 112 x 112 to fit the training process with the R(2+1)D model.

2.3. Determining the type of interactive object in the exercise

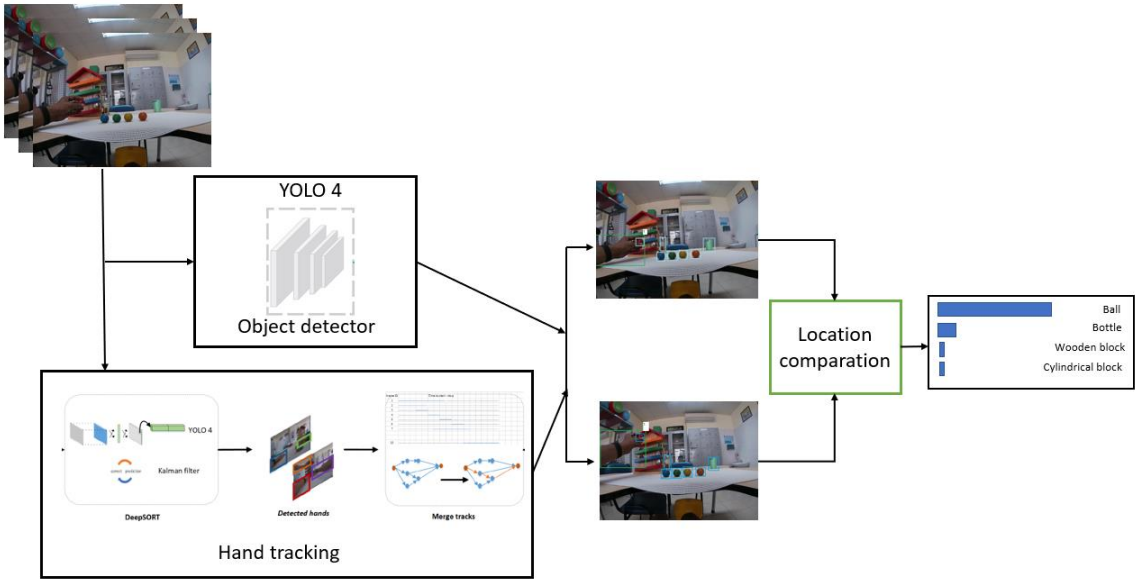

Figure 10. *Method of determining the type of interactive object.*

Figure 10 describes the method for determining the type of interactive object in the exercise. The method includes the following steps:

- Step 1: Detecting objects on frames;

- Step 2: Identifying the patient's hand on the frames;

- Step 3: Comparing hand position and detecting objects in the frames to determine the type of interactive object.

Consecutive frames in the exercise video are fed into the object detection network to detect objects (object type, bounding box of object) in the frames. In the meantime, these consecutive frames are also passed through the hand tracker to identify patient's hand on each frame. Finally, through an algorithm that compares the relative positions of hands and objects across all frame sequences, the type of interactive object will be determined.

Object detector

We propose to use Yolov4 [10] object detection network to detect objects on exercise frames. Yolo is a lightweight object detection model, which has the advantages of fast speed, low cost of computations and small number of model parameters. We used 2700 images of rehabilitation exercises labeled with object detection (object type, object bounding box) to train the Yolov4. Labeled objects include the following 4 types: ball, water bottle, cube and cylinder. These are the types of interactive objects in the exercises.

Patient's hand tracking in consecutive frames

We use the hand tracker proposed in [11] to track the patient's hand on the consecutive frame sequence of the exercise video. This is a two-step tracker to detect and locate the patient's hand per frame. First step, a DeepSORT model is used to perform hand tracking task. The second step, we use the Merge-Track algorithm to correct the misidentification of the hand bounding boxes from the results of the first step.

Compare locations and determine the interactive object type

The interactive object in the exercise is defined as an object whose distance to the hand varies at least across frames, and it has the largest ratio of intersecting areas to the patient's hand. Therefore, we propose an algorithm to determine the type of interactive object as follows:

- For every i-th frame in a sequence of n consecutive frames, calculate the score to evaluate the position between the hand and each object on the frame, according to the formula:

$$Score[k,j,i] = \frac{Inter(OBJ\_bbox_{k,j,i}, Hand\_bbox_i)}{OBJ\_bbox_{k,j}} \quad (1)$$

Where: $OBJ\_bbox_{k,j}$ - Bounding box of k-th object of class j. $Hand\_bbox_i$ - Bounding box of patient's hand on the i-th frame; *Inter (O, H)* is the intersection between O and H

- Calculate the relative position evaluation score between the hand and the j-th object on the i-th frame:

$$Score[j,i] = \max_{k \in Object\_j} \{Score[k,j,i]\} \quad (2)$$

Where $= 1 \div 4$: is the j-th object class, k is the k-th object of the j-th object class. If Yolo does not detect any object of the j-th object class in the frame, then $Score[j,i] = 0$.

- Calculate the relative position evaluation score between the hand and the j-th object class in a sequence of n consecutive frames:

$$Score[j] = \sum_{i=1}^{n} Score[j,i] \quad (3)$$

- Normalize the position evaluation score to the interval [0,1]:

$$Score[j] = \frac{Score[j]}{\sum_{j=1}^{4} score[j]} \quad (4)$$

The output of the comparison algorithm is the relative position evaluation score vector between the feature class and the handset $\{Score[j], j = 1 \div 4\}$. Where the higher the evaluation score, the higher probability that the object class is the type of interactive object.

2.4. Combining hand action recognition results and interactive object type to identify the exercise

In the rehabilitation exercises video dataset, there are a number of similar exercises i.e. the hand gestures are very similar in these exercises. The hand action recognition network is very easy to mispredict these exercises. On the other hand, from studying the exercise video data, we know that each rehabilitation exercise is characterized by a type of interactive object. Therefore, we suggest incorporating information about the type of interaction object to accurately determine the exercise that the patient did in the video:

- The output of the action recognition network is a probability vector that predicts exercises performed in a sequence of frames: $\{Prob\_Recognize[j], j = 1 \div 4\}$.

- The output of the model determines the type of interaction object is a vector that evaluates the possibility that the object class can be an interactive object type of exercise: $\{Score[j], j = 1 \div 4\}$.

- Calculate scores of the exercises:

$$Score\_exercise[j] = Prob\_Recognize[j] \times Score[j] \quad (5)$$

- The exercise performed in the sequence of frames is exercise j_0:

$$j_0 = \underset{j=1\div4}{\operatorname{argmax}}\{Score\_exercise[j]\} \quad (6)$$

3. EXPERIMENTAL RESULTS AND DISCUSTION

3.1. Dataset

We use the RebHand dataset collected from 10 patients at the rehabilitation room of Hanoi Medical University Hospital. Participating patients are asked to wear two cameras on their forehead and chest and two accelerometers in both hands during the exercise. The camera records all the patient's hand movements during the exercises. Recorded videos are divided into exercise videos and labeled. A total of 431 exercise videos of 10 patients. Length of each video from 2-5s. Table 2 shows the statistics of the number of exercise videos of 10 patients.

Table 2. *The number of exercise videos of 10 patients.*

	Patient	Exercise 1	Exercise 2	Exercise 3	Exercise 4	Train	Test
1	Patient 1	10	32	11	12		X
2	Patient 2	26	8	17	9	X	
3	Patient 3	23	6		9	X	
4	Patient 4	9	4			X	
5	Patient 5	15	7		13		X

6	Patient 6				6	X	
7	Patient 7	8	4		9	X	
8	Patient 8	7	6		8	X	
9	Patient 9	16	23	47	30		X
10	Patient 10	10	6	22	18	X	
	Total	**124**	**96**	**97**	**114**		

The exercise videos of the "RehabHand" dataset are divided into 2 sets: training set and test set. The training set consists of data from 7 patients. The test set includes the data of the remaining 3 patients.

3.2. Implementation and evaluation metric

The proposed models are implemented using Python and Pytorch Tensorflow backend. All algorithms and models are programmed/trained on a PC with a GeForce GTX 1080 Ti GPU. The action recognition network is updated via Adam optimizer, the learning rate of Adam is set to 0.0001. The model is trained 30 epochs and the model generated at the epoch with a max accuracy value on the validation set is the final model.

We used classification accuracy and confusion matrix to evaluate the proposed recognition methods.

3.3. Evaluating the accuracy of the network R(2+1)

- *Training stage*

We implemented the R(2+1)D network and trained the network using the "RehabHand" dataset for exercise recognition with training data consisting of exercise recording videos of 7 patients. The dataset is divided into 8:2 ratio for training and validation. The model is trained with the following parameters: Batch_size = 16, Input_size = 112x112, epoch = 30.

Figure 11 illustrates the average accuracy of the model during training, and table 3 presents the result of the model's accuracy for each exercise. The table and figure show that the average accuracy value of the model is practicable at 86.3%. Being 69%, the accuracy of the experiment for Exercise 3 is much lower than the ones for another exercise. It is because that experiment 3 is mistaken as Experiment 1, up to 31% (figure 12). In fact, these two exercises have the same space and implementation method, and they have a quite similar scene and hand gestures. The results of the remaining exercises are very high with exercises 1 and exercises 2 at 94%, exercise 4 at 93%.

Table 3. *Accuracy of the model in the training stage*

	Exercise	**Accuracy (%)**
1	Exercise 1	94
2	Exercise 2	94
3	Exercise 3	69
4	Exercise 4	93
	Average	*86.3*

- *Accuracy on the test dataset*

After training the model R(2+1)D, the best parameter of the model is saved. Next, we evaluate the accuracy of the model on the test set of 216 exercise videos of 3 patients,

independent with the training data. The accuracy and the confusion matrix with 4 classes of exercises according to the number of videos in the test set are shown in Table 4 and Figure 13, respectively. The accuracy of exercise recognition on the test set is 86.11%. This result is quite similar to the training result. Exercises 2 and Exercises 4 have good recognition results. However, Exercises 1 and Exercises 3 have lower accuracies because there are mutual mistakes between the two exercises.

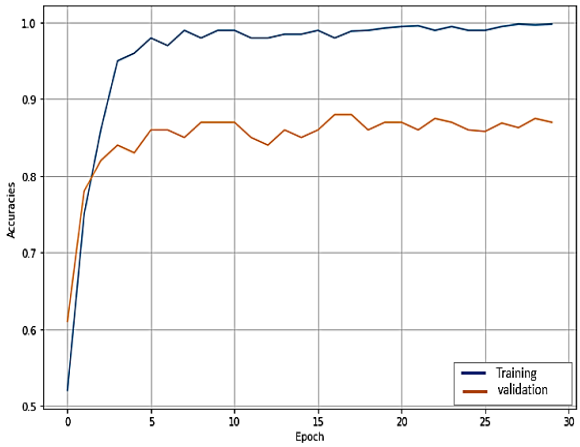

Figure 11. *Accuracy of the model during training.*

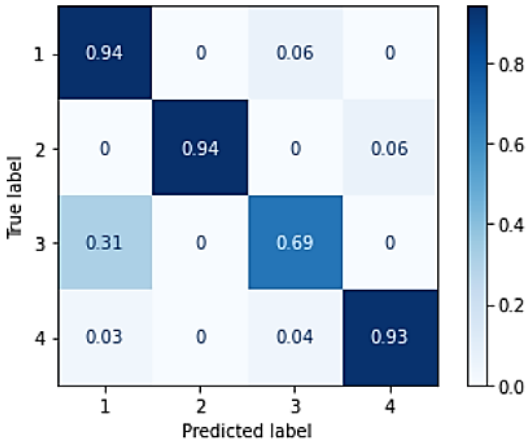

Figure 12. *Confusion matrix of the network R(2+1)D on the training set.*

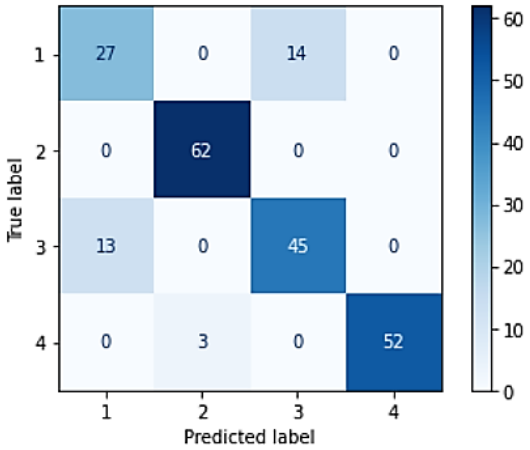

Figure 13. *Confusion matrix of the network R(2+1)D on the test set.*

Table 4. *Recognition accuracy on test set.*

	Exercises	Videos	Number of videos correctly recognized	Accuracy (%)
1	Exercise 1	41	27	65,85
2	Exercise 2	62	62	100
3	Exercise 3	58	45	77,59
4	Exercise 4	55	52	94,54
	Average			*86,11*

3.4. The accuracy of determining interactive object type

We perform this test to evaluate the accuracy of determining the interactive object type method with the test set. The test method is as follows: consecutive frames in the

exercise video are fed through an object detection network to identify the objects (object type and object bounding box) in each frame. At the same time, these consecutive frames are also passed through the hand tracker to identify the position of the patient's hand on each frame. The results of the object detector and the hand tracker are used to calculate the score vector that evaluates the object classes $\{Score[j], j = 1 \div 4\}$. The type of interactive object in the video is the j_0 th object: $j_0 = \underset{j=1\div4}{\text{argmax}} \{Score[j]\}$. The accuracies of determining the interactive object type of the exercises are shown in table 5. This table shows that the accuracy of the method for determining the type of interaction in the exercises is high with the average is 80.09%. The highest is the water bottle object with an accuracy of 93.55%. The lowest is a cylindrical object with an accuracy of 76.36%. It is because the large water bottle object in the frame is not obscured much, so the object detector is easy to detect correctly. Whereas the cylindrical object is quite small and obscured by the hand, so object detection is difficult to detect this type of object.

Table 5. *Accuracy of interactive object type determination.*

	Object	Videos	Correct determination	Accuracy (%)
1	Ball	41	33	80,49
2	Water bottle	62	58	93,55
3	Wooden cube	58	40	68,96
4	cylindrical block	55	42	76,36
	Average			*80,09*

3.5. Accuracy of the proposed combined exercise recognition method

In this experiment, we implement the proposed rehabilitation exercise recognition model and evaluate the model on the test set of 3 patients. The steps to predict the exercise according to our method are as follows: The frame sequence is sampled and fed into the R(2+1)D network to get the probability prediction vector. At the same time, the frame sequence is passed through the interactive object type determination algorithm to calculate the object type evaluation score vector. The two output vectors are then combined to determine the exercise on the video according to the method presented in section 2.4.

Table 6 shows the accuracy for each exercise, and figure 14 illustrates the confusion matrix. The average exercise recognition accuracy is 88.43%. Exercises 2 and Exercises 4 have fairly good recognition results while Exercises 1 and Exercises 3 have lower results. There is still confusion between exercise 1 and exercise 3, but the number of confusions is less than the result of exercise recognition of the network R(2+1)D.

Table 6. *The accuracy of exercise recognition on the test set of the proposed method*

	Exercises	Videos	Number of videos correctly recognized	Accuracy (%)
1	Exercise 1	41	33	80,49
2	Exercise 2	62	61	98,39
3	Exercise 3	58	47	81,03
4	Exercise 4	55	50	90,91
	Average			*88,43*

Figure 15 illustrates a chart comparing the accuracy of the exercise recognition of the proposed method and the R(2+1) D network. This figure shows that the problem recognition accuracy of the proposed method is generally greater than the accuracy of the R(2+1)D network. The proposed method improves the average accuracy from 86.11% to 88.43%. Especially, for the exercise 1, the proposed method has a remarkable increase of 14.64% in the accuracy of recognition from 65.85% to 80.49. This is because the algorithm determines the type of interactive object as "Ball" quite well, and the information is added for the recognition of "Exercise 1" more accurately. Hence, it helps to reduce the number of mutual mistakes with “Exercise 3” where the interactive object is “wooden cube”.

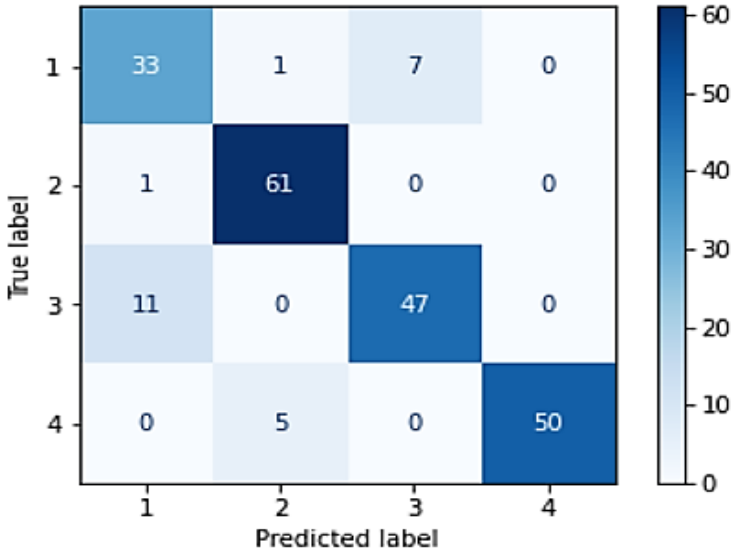

Figure 14. *Confusion matrix of the proposed method on the test set.*

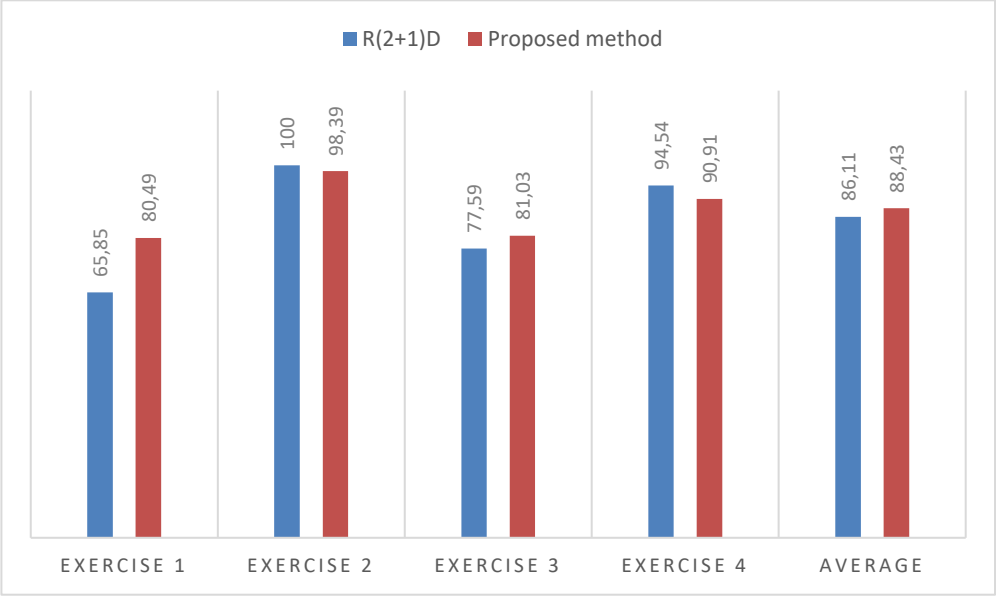

Figure 15. *Recognition accuracy of four exercises using the proposed method versus R(2+1)D network.*

4. CONCLUSIONS

The paper proposes a method of recognizing hand action in rehabilitation exercises, i.e., automatically recognizing the rehabilitation exercise of patients from the egocentric videos obtained from wearable cameras. The proposed method combines the results of the deep learning network R(2+1)D for RGB video-based action recognition and the algorithm to determine the type of interactive objects in the exercise, thereby giving the exercise recognition results with high accuracy. The proposed method is implemented, trained, and tested on the RehabHand dataset collected from patients at the rehabilitation room of Hanoi Medical University Hospital. The experimental results show that the accuracy of the exercise recognition is high and practicable and superior to the recognition results of the R(2+1)D network. It illustrates that the proposed method can reduce the rate of confusion between the exercises having similar hand gestures. It also proves that the algorithm for determining the type of interactive objects in the exercises is good to produce good results. In the future, we will continue to perform experiments with other action recognition networks to improve the accuracy and speed of the recognition model.

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

REFERENCES

[1]. Fathi, A., Farhadi, A. and Rehg, J.M. *"Understanding egocentric activities"*. In 2011 international conference on computer vision (pp. 407-414). IEEE, (2011).

[2]. Fathi, A., Li, Y. and Rehg, J. M. *"Learning to recognize daily actions using gaze"*. In European Conference on Computer Vision (pp. 314-327). Springer, Berlin, Heidelberg, (2012).

[3]. Fathi, A., Ren, X. and Rehg, J. M. *"Learning to recognize objects in egocentric activities"*. In CVPR 2011 (pp. 3281-3288). IEEE, (2011).

[4]. Li, Y., Ye, Z. and Rehg, J.M. *"Delving into egocentric actions"*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 287-295), (2015).

[5]. McCandless, T. and Grauman, K. *"Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition"*. In BMVC (Vol. 2, p. 3), (2013).

[6]. Pirsiavash, H. and Ramanan, D. *"Detecting activities of daily living in first-person camera views"*. In 2012 IEEE conference on computer vision and pattern recognition (pp. 2847-2854). IEEE, (2012).

[7]. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M. *"A closer look at spatiotemporal convolutions for action recognition"*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6450-6459), (2018).

[8]. Hara, K., Kataoka, H. and Satoh, Y. *"Can spatiotemporal 3D CNNS retrace the history of 2D CNNS and ImageNet?"* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6546-6555), (2018).

[9]. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L. *"Temporal segment networks for action recognition in videos"*. IEEE transactions on pattern analysis and machine intelligence, 41(11), pp.2740-2755, (2018).

[10].Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M. *"Yolov4: Optimal speed and accuracy of object detection"*. arXiv preprint arXiv:2004.10934, (2020).

[11].Sinh Huy Nguyen, Hoang Bach Nguyen, Thi Thu Hong Le, Chi Thanh Nguyen, Van Loi Nguyen, Hai Vu, *"Hand Tracking and Identifying in the Egocentric Video Using a Graph-Based Algorithm,"* In Proceeding of the 2022 International Conference on Communications and Electronics (ICCE 2022).

TÓM TẮT

Phương pháp nhận biết hoạt động tay trong bài tập phục hồi chức năng sử dụng mạng học sâu nhận dạng hoạt động và thông tin xác định đối tượng tương tác

Nhận dạng hoạt động của tay trong các bài tập phục hồi chức năng chính là tự động nhận biết bệnh nhân đã tập những bài tập PHCN nào, đây là bước quan trọng trong hệ thống AI hỗ trợ hỗ trợ bác sỹ đánh giá khả năng tập và phục hồi của bệnh nhân trong các bài tập phục hồi chức năng. Hệ thống này sử dụng các video thu được từ camera đeo trên người bệnh nhân để tự động nhận biết và đánh giá khả năng tập PHCN của bệnh nhân. Trong bài báo này chúng tôi đề xuất một mô hình nhận biết hoạt động của tay bệnh nhân trong các bài tập phục hồi chức năng. Mô hình này là sự kết hợp của kết quả mạng học sâu nhận biết hoạt động trên Video RGB R(2+1)D và thuật toán phát hiện đối tượng tương tác chính trong các bài tập, từ đó cho ra kết quả nhận biết bài tập của bệnh nhân với độ chính xác cao. Mô hình đề xuất được cài đặt, huấn luyện và thử nghiệm trên bộ dữ liệu về các bài tập phục hồi chức năng thu thập từ camera đeo của các bệnh nhân tham gia bài tập. Kết quả thực nghiệm cho thấy độ chính xác trong nhận dạng bài tập khá cao, trung bình đạt 88,43% trên các dữ liệu thử nghiệm độc lập với dữ liệu huấn luyện. Kết quả nhận dạng hoạt động của phương pháp đề xuất vượt trội so với kết quả nhận dạng của của mạng nhận dạng hoạt động R(2+1)D và làm giảm tỉ lệ nhầm lẫn giữa các bài tập có cử chỉ tay gần giống nhau. Sự hợp kết quả thuật toán xác định đối tượng tương tác trong bài tập đã làm cải thiện đáng kể độ chính xác của mô hình nhận dạng hoạt động.

Từ khóa: Nhận dạng hoạt động; Bài tập phục hồi chức năng; Theo dõi và phát hiện đối tượng; R(2+1)D.