Support Learning Vovinam Exercises based on Computer Vision

Final Year Project Final Report

A 4th Year Student Name

Pham Son Tung

Thai Thanh Do

Pham Hong Giang

Instructor

Dr. Phan Duy Hung

FPT UNIVERSITY

Bachelor of Computer Science

Hoa Lac campus - FPT University

2 December 2022

ACKNOWLEDGEMENT

We would like to thank our instructor, Dr. Phan Duy Hung for his patience and time, and for instructing and advising us enthusiastically.

We would like to thank all at my University, FPT, for giving us the best environment to study and grow over the years.

We would like to thank our classmates in CS1404, for letting us meet amazing people and learn a lot from them.

We always remember our family's encouragement and support. Thanks to them, we have the will, the energy and the confidence to pursue our goals.

DECLARATION

We declare that the work in this dissertation titled "Support Learning Vovinam Exercises based on Computer Vision" has been carried out by our research with FPT University Computer Science Department. It has not been previously submitted, in part or whole, to any university of institution for any degree, diploma, or other qualification.

Signed:_____Tùng __Đô _Giang_____

Date:___08/12/2022_____

Pham Son Tung, Thai Thanh Do, Pham Hong Giang and full qualifications

FPT University

ABSTRACT

Computer vision has many applications which has attracted many researchers, especially with the problems of recognizing actions, postures, movements. This Thesis offers a method to support students to perform correct postures during martial arts practice. We have collected and labeled data about the movements of a traditional Vietnamese martial art called Vovinam.

In the original paper of ST-GCN, before input into the model, we need to transform videos to the sequence of keypoint positions by frame to be handled in the next phase; it seems to be that normally the transform phase didn't reach effective performance. Therefore, the purpose of this thesis is to improve the ST-GCN model in terms of input. Firstly, we use a sequence of recently released techniques to extract the skeleton and its key points from the input video. Then, the sequence of keypoint positions by frame will be inputted into the deep learning architecture based on the ST-GCN model and the output will be the determined action. On our dataset, adding the input processing stage to the recognition model has yielded much better results than applying the original model. The final accuracy is 99.23%, showing that the model has the potential to be applied in practice.

Keywords : Computer Vision, ST-GCN, Martial Art, Vovinam

Table of Contents

ACKNOWLEDGEMENT **2**
Table of Contents **5**
List of tables **6**
List of figures **7**
LIST OF ABBREVIATIONS AND ACRONYMS **8**
1. INTRODUCTION **9**
1.1. Overview 9
1.2. Related works 11
1.3. Contribution 14
1.4. Outline 15
2. METHODOLOGY **16**
2.1. Overview pipeline 16
2.2. YoloX 16
2.3. Byte-Track 18
2.4. HRNet-pose 19
2.5 ST-GCN for Skeleton Based Action Recognition 20
2.5.1 Processing Keypoints *20*
2.5.2 ST-GCN Overview *21*
2.5.3 Implementing ST-GCN *22*
3. EXPERIMENTS RESULT AND CONCLUSION **25**
3.1. Data Collection 25
3.2. Experiments 27
3.3. Result and analysis 29
3.4. Inference 31
3.5. Conclusion and Future Works 32
References **34**

List of tables

Table 1: Data statistic 25

Table 2: YoloX parameter 27

Table 3: Byte-track parameters 28

Table 4: The parameters of ST-GCN model 29

Table 5: Results of two cases with and without processing keypoints 29

List of figures

Figure 1: Vovinam popularity spans various ages and countries 09

Figure 2: Example of human body skeleton with keypoints (single and multiple target) 11

Figure 3: Collect standard data of master's skeleton pose from the Kinect camera for grading Vietnamese traditional martial arts movements 14

Figure 4: Block diagram for classifying martial art action 16

Figure 5: Difference between YOLOv3 head and the proposed decoupled head 17

Figure 6: Byte-track associates every detection box 18

Figure 7: Illustrating the architecture of the proposed HRNet 20

Figure 8: A skeleton sequence's spatial-temporal graph 22

Figure 9: Scooping movement 25

Figure 10: Double punch movement 26

Figure 11: The meanLoss 30

Figure 12: The accuracy 30

Figure 13: Some examples of Vovinam movement recognition 31

Figure 14: Predictive process of the model on long video 32

List of abbreviations and acronyms

Abbreviations	Meaning
AI	Artificial Intelligence
YOLO	You Only Look Once
ST-GCN	Spatial Temporal Graph Convolutional Networks
FPN	Feature Pyramid Network
HRNet	High-Resolution Net
CNN	Convolution Neural network
GCN	Graph Convolution Networks
BN	Batch-Normalization
AFT	Affine Transformation
IoU	Intersection over Union

1. INTRODUCTION

1.1. Overview

People have long been undertaking physical activity to maintain and improve health. This is considered as an indispensable part of life. In multiple forms of physical activities that people use today, martial arts is one of the most effective and popular ways. In Vietnam, the national traditional martial art form of the country - Vovinam, originated in 1938 [1]. On the 80th founding day anniversary, Vietnam Vovinam Federation confirmed there were more than 2.5 million practicing the martial art in 70 countries and territories. The strong growth of the art form can be seen through the establishment of the world federation, as well as continental federations in Asia, Europe and Africa.

Figure 1. Vovinam's popularity spans various ages and countries.

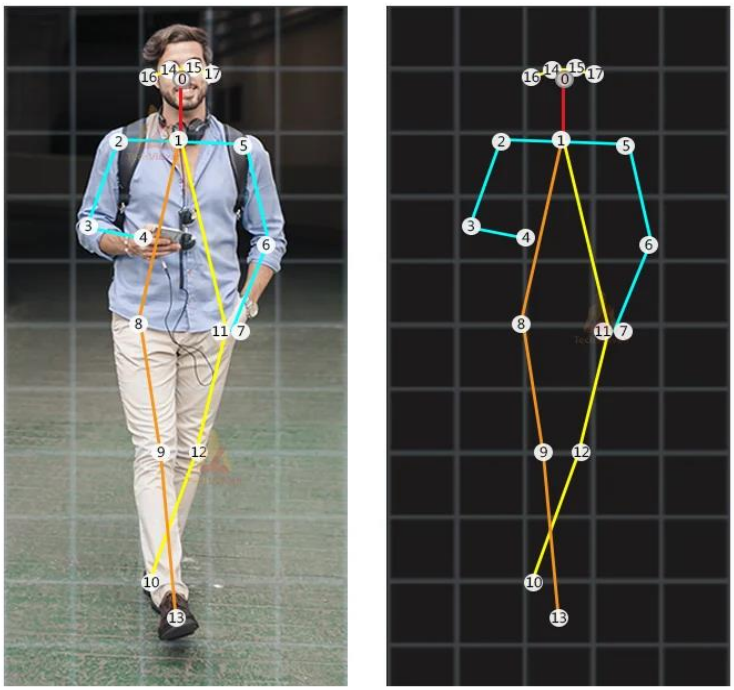
From the research of Dung et al [2], the writer claims that the strengths and technique will increase parallelly to practice time. Since the instructor could not always supervise and guide the trainee, it requires the learner to spend most of the time practicing by themselves in order to master the required skills. Along with a tremendous amount of time on self-training, learners could also utilize the application of computer vision as a useful tool to support self-training. This can later create an immensely positive effect in the training process. The support of self-training can help improve the self-assessment technique and as a result, the apprentice can achieve better results. Controlling the techniques could also ensure that trainees do not encounter any unwanted injuries or train in inappropriate ways.

The goal of this thesis is to develop a deep learning architecture based on the ST-GCN model with an additional processing stage for the input. Recognizing the challenge as a video of multiple people practicing vovinam, we used object detection to determine the human limits of each frame. Object detection is a fundamental and frequently employed approach in computer vision. It is used to identify frame components from which deeper issues can be expanded. One-stage techniques and two-stage approaches are the two primary categories of object detection. YoloX is a one-stage approach with the benefit of quick inference that was employed in this thesis. After outputting from YoloX, the frames are then tracked to build bounding boxes sequences that are linked throughout time. One of the tasks in the object detection problem is object tracking, which involves the problem of following one or more moving objects across time in a video. it involves creating a special ID for each initial detection, following each object as it moves along through video, and preserving the ID allocated to that object. From the tracking bounding boxes, we use a recent method to obtain the position coordinates of each keypoints according to the previous bounding boxes tracking. The task of simultaneously detecting objects and locating keypoints is known as keypoint detection. Keypoints might be pixel positions or locations in space. In this thesis, we use the HRNet-pose model to generate a sequence of keypoints that are skeleton keypoints extracted during the video duration. Finally, we put the sequence

of keypoints positions by frame containing important information regarding space as the position of the skeleton and time as the position it moves over time into the ST-GCN model which has a few small improvements and then takes action.

1.2. Related works

The application of action recognition has been widely researched in recent years. Especially, the rise of new research on human body skeleton dynamics contributes an important part for recognizing human actions. A variety of applications were found by learning attributes from the skeleton to recognize human actions in temporal data.



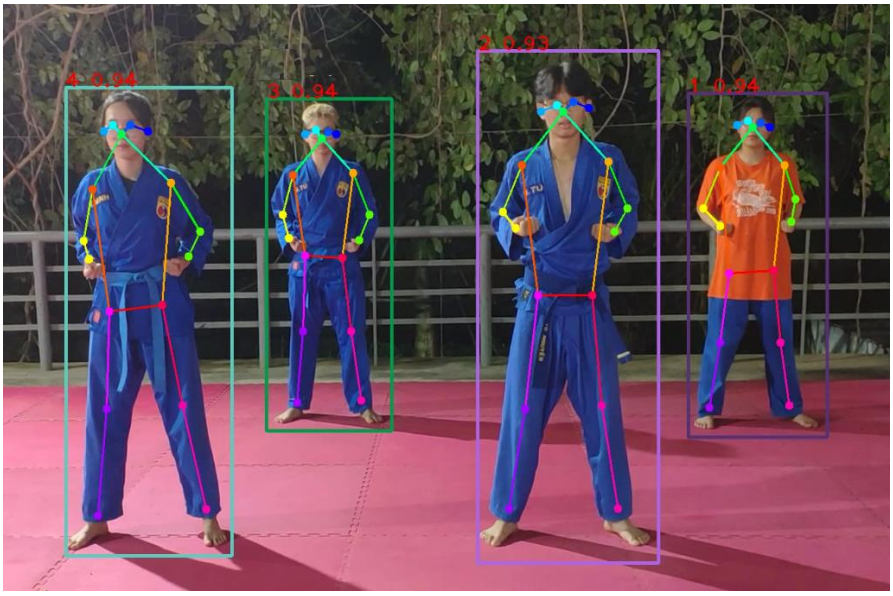


Figure 2. Example of human body skeleton with keypoints (single and multiple target)

In the field of evaluating the effectiveness and performance of human action recognition, some of the studies can be mentioned such as: tracking body movements by Malaguti et al [3], Classification of activities by Atvar et al [4], Detection of fallen persons by Solbach et al [5], Cheating Detection System by Samir et al [6], Fighting Detection by Pan et al [7]. To collect data on physical activities, studies using RGBD cameras [8], Accelerometers [9], Kinect [10]. Existing methods often manually extract features from human motion data and then use basic machine learning algorithms such as SVM, k-nearest, iso-forest… Malaguti et al [3] introduce a brand-new real-time tracking system that can enhance distributed camera networks' body posture estimation techniques. Solbach et al [5] employed machine learning methods to detect humans with falls using CNN-based human pose estimators in conjunction with stereo data to rebuild the human posture in 3D and estimate the ground plane in 3D. Samir et al [6] propose an exam cheating detection system, the suggested method makes use of single-user and multiple-user tracking techniques called Human Pose Estimation. The technology continually validates the circumstances of a student's head position and hand movement conditions during the exam to determine whether or not they are cheating based on video evidence. In order to identify the behavior posture based on the optical detection approach, Pan et al [7] add pedestrian key nodes. The response graph of

the picture is first created in the convolution structure, and the key nodes of the pedestrian body are acquired through the extreme of the response graph. Second, by linking the important nodes, the limb of a pedestrian is shown. Finally, anomalous behavior and the posture of the pedestrian are determined using the optical flow of the key nodes.

Related to the research on the system of training, grading or support in learning martial arts can include some recent studies. In [11], Pang et al. review martial arts applications of computer vision. These applications, which expand the reach of martial arts and provide more scientific instruction and technical analysis, include action recognition, stance estimation, intelligence judgment, etc. According to an analysis of the previous research, the investigation in this paper serves as inspiration for subsequent research. Based on convolutional neural networks (CNN), Cai et al. studies the human behavior detection algorithm and its interactive key technology in the martial art field [12]. The detection impact is enhanced by the important frame extraction, joint point optimization, and improved bottom-up technique. At the same time, a new neural network is created that consists of a style generating network and a separable multi branch network. In order to create the style AI image, Generative Adversarial Networks (GAN) first precisely identify the human skeleton and judge the behavior. This study lays the groundwork for future research in the area of AI art. For grading movements of Vietnamese martial arts, Tuong Thanh et al in [13] proposes the implementation of data analysis of the depth of scoring Kinect users' camera movements in grading Vietnamese traditional martial arts movements. It is an evaluation of traditional Vietnamese martial arts, aiding in the growth and preservation of ethnic elites.

Figure 3. Collect standard data of master's skeleton pose from the Kinect camera for grading Vietnamese traditional martial arts movements.

1.3. Contribution

To support martial arts training, this article focuses on the issue of evaluating whether Vovinam movements are correct or not to help users raise the movement level as well as adjust the right movements according to the martial arts exercise being practiced. In this effort, we approached the evaluation of temporal movements using ST-GCN [14], a model of dynamic skeletons known as ST-GCNs, which automatically learns not only spatial but also temporal patterns from the data. Our goal is to develop an end-to-end action recognition model that leverages the ST-GCN model with several input modifications and some recent computer vision techniques to deal with the complex moves of vovinam martial arts. With the new dataset we have built and shared with the community, we would further introduce our work in the following sections. This study suggests an advanced pipeline, which were considerately chosen by the authors. Each of the blocks has been precisely examined by the writers to generate the final process graph. Following the pipeline, we have performed auto-generator data from video to keypoints sequence and processed them appropriately for

material arts movements. Later on, the data would be put into the ST-GCN model to deal with the problem of classification of martial arts poses.

1.4. Outline

In thesis, we address the problem of Vovinam movements recognition, specifically:

Section 1 gives a gentle introduction about the problem, motivation and related works about Vovinam movements recognition.

Section 2 demonstrates our approach which is a whole new pipeline created by us. We use the STGCN model with an additional processing stage for the input. In each stage, we will explain and clarify all the models and methods implemented in that stage.

Section 3 gives an overview about the data set and how we collect and manage the data set. Experimental steps proceed to select parameters on each block of the pipeline. Then showed the evaluation method and some experiment results between before and after our fine-tuning effort for the input of the ST-GCN model.

Section 4 concludes the thesis and then makes some future work in the subject.

The final section is the list of all reference works helping to create this thesis.

2. METHODOLOGY

2.1. Overview pipeline

In this thesis, we propose the pipeline (Fig.4) to classify martial art action as follows:

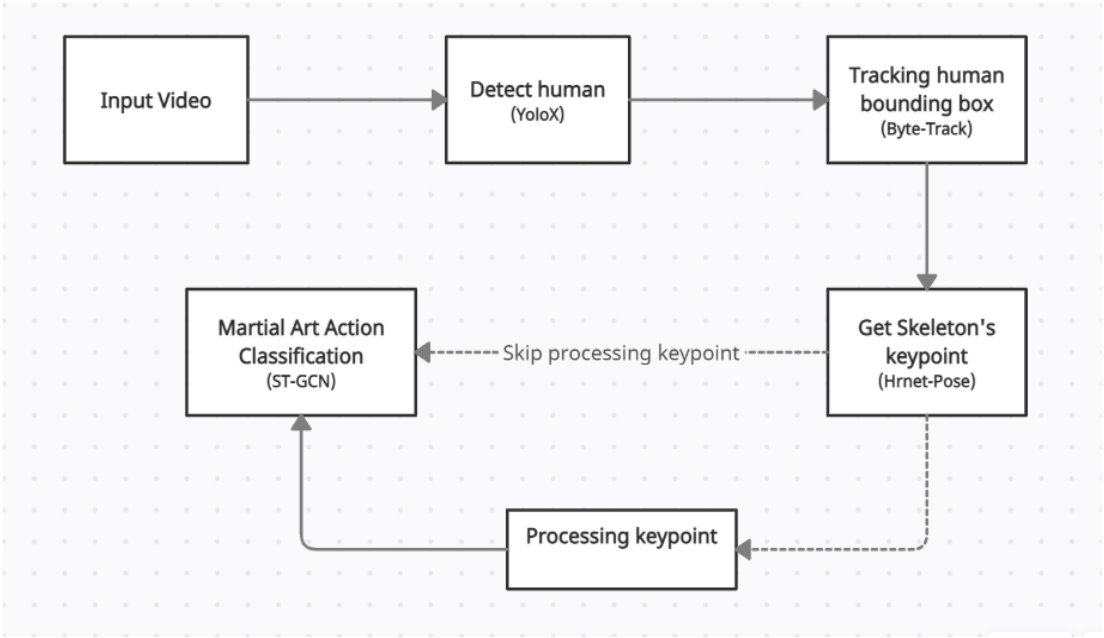


Fig. 4. Block diagram for classifying martial art action

As shown in Figure 4, pipeline with the following steps: Human detection (YoloX), tracking the bounding boxes of humans (Byte-Track), and obtaining the skeleton's keypoint (HRNet-Pose). Our proposal has the main purpose of helping the process of extracting keypoints from video automatically. We then preprocess those keypoints and feed them into the Spatial Temporal Graph Convolution Network model for classification martial art action.

More information on each block and approach is provided below.

2.2. YoloX

In the initial phase of the pipeline for this study, we employ **YoloX**[15]. With the input from each frame of the video to determine the bounding-boxes for each person in the frames. We chose YoloX for this model because it improved object detection by emphasizing anchor-free detectors [16] and had superior detector performance than the Yolo-series [17, 18].

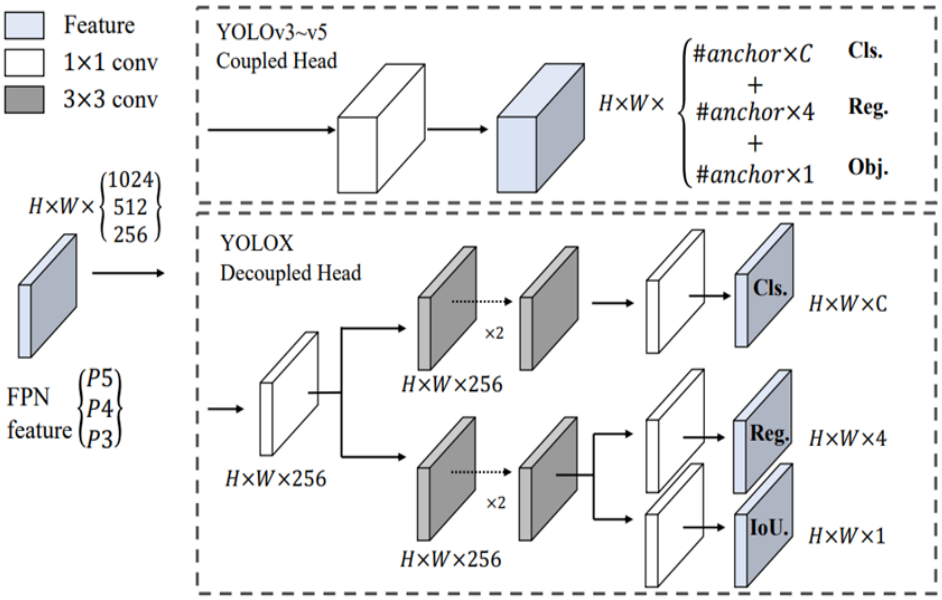


Fig. 5. YOLOv3 head and YoloX decoupled head [15]

YoloX initially implements a 1x1 conv layer for each level of FPN features in order to decrease the feature channel to 256 and afterwards inserts two parallel branches with two 3x3 conv layers for the classification and regression tasks. Regression branch is expanded with the IoU branch. Additionally, the anchor-free method significantly reduces the number of design parameters that require heuristic adjustment as well as the number of tricks required for effective performance, simplifying the detector's training and decoding phases. It is very simple to convert YOLO to an anchor-free mode. By predicting four variables straight, namely two offsets in terms of the grid's top left corner and the height and width of the anticipated box the approach reduces the forecasts for each location from three to one. The center of each object is then designated as the positive sample, and a scale range is built.

And It is also recommended that Yolox be used with the byte-track method [19], which will be introduced in the next part of the model, for better performance. We use a pretrained model for the Yolox method according to [15].

2.3. Byte-Track

In the next phase of the pipeline, after determining the bounding-boxes for each person in the frames, **Byte-track** [19] will be used to match the bounding-boxes of each person in the frames.

Now we'll look at how the byte-track works. We begin by matching the box (which is detected to have a high score) to the tracklets using motion or appearance similarity. To forecast in the new frame the tracklets placement, Byte-Track employs the Kalman filter [21]. The similarity may be calculated using the IoU or Re-ID feature distance between the predicted box and the detected box.

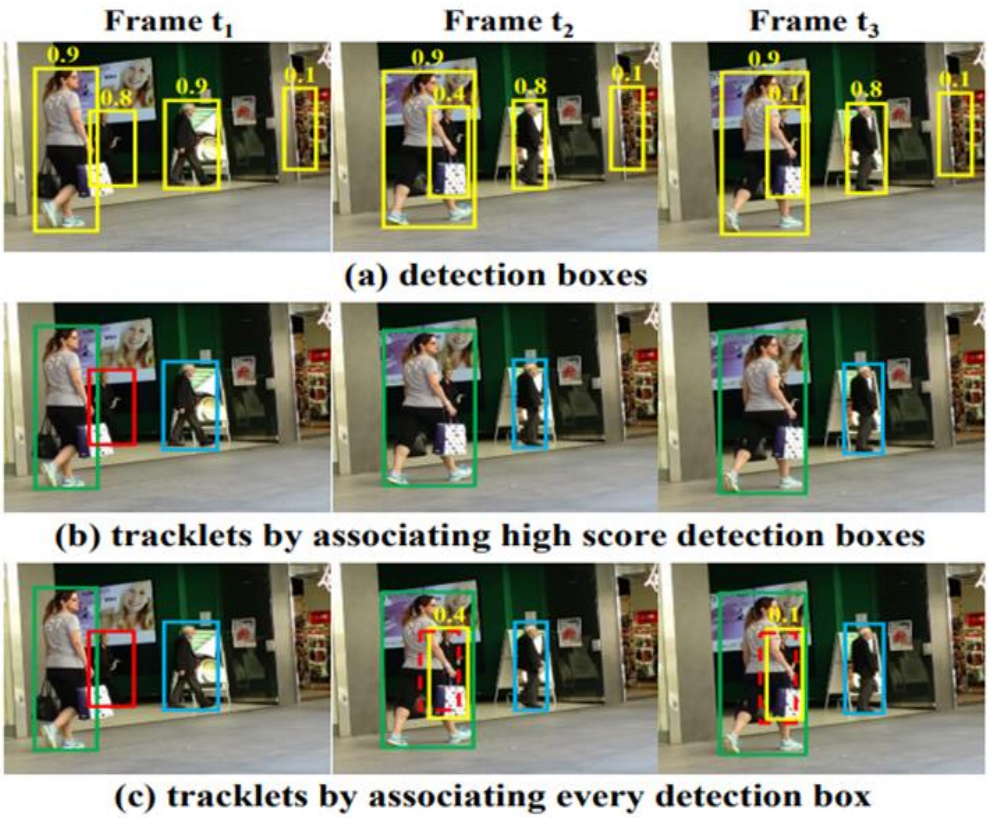


Fig. 6. Byte-track associates every detection box [16]

Figure 6 is an example of how each detection box in byte-track is associated. Figure (a) displays all of the detection boxes and their scores, whereas Figure (b) displays the tracklets produced by earlier approaches that correlate detection boxes with scores greater than a certain threshold, in this case, 0.5. In (c), which displays the tracklets produced, the same

box color denotes the same identity. The Kalman Filter's anticipated box of the earlier tracklets is shown by the dashed boxes. Based on the large IoU, the two low score detection boxes are accurately matched to the prior tracklets.

The majority of earlier methods locate identities by finding association detection boxes with scores over a threshold, then discarding the rest of the objects with low detection scores, resulting in significant actual object loss and fractured trajectories. To deal with that problem, associating nearly every detection box with Byte-track allows for tracking, as opposed to just tracking the high-scoring ones.

2.4. HRNet-pose

After matching the bounding-boxes for each person in the frames, **HRNet** [20] is used to define keypoints. The HRNet method uses input as input bounding-boxes and outputs a sequence of keypoints positions for each frame of each person.

HighResolution Net (HRNet), a cutting-edge design, is able to preserve high resolution representations throughout the entire process. Starting with a high-resolution subnetwork as the initial stage, they gradually add high-to-low resolution subnetworks to construct further stages, and then connect the multi-resolution subnetworks simultaneously. By continuously transferring data across the parallel, multi-resolution subnetworks, they carry out repeated multi-scale fusions. Over the high-resolution representations that are produced by our network, they estimate the keypoints.

In comparison to other frequently used pose estimation networks, the network has two advantages. In contrast with most previous systems that were linked in a series, with this technique, high-to-low resolution subnetworks are linked in parallel. By maintaining the high resolution rather than recovering it through a low-to-high process, the technique is able to provide a heatmap that may be more spatially precise. The majority of extant fusion systems combine low-level and high-level representations. However, in this method use recurrent multiscale fusions, high-resolution and low-resolution will be augmented representations if they have the same depth and level, and vice versa, leading in high-

resolution representations with abundant pose estimate. As a result, the anticipated heatmap may be more accurate.

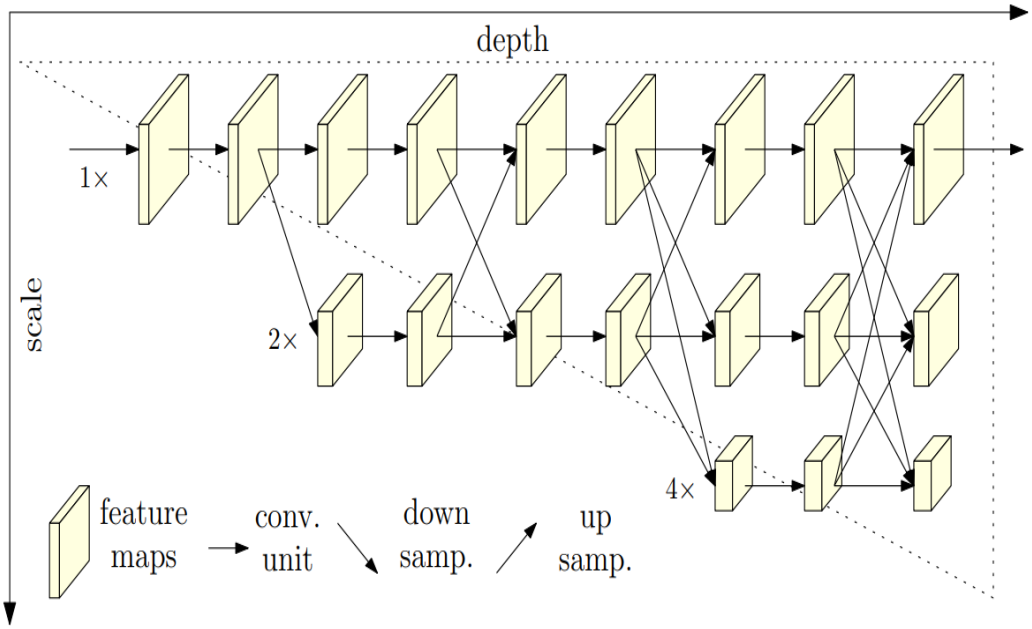


Fig. 7. Illustrating the architecture of the proposed HRNet [16]

The planned HRNet's design is seen in Figure 7. It comprises of concurrent high-to-low resolution subnetworks that frequently exchange information with other subnetworks with different resolutions (multi-scale fusion). The network's depth and the scale of the activation maps, respectively, can be determined from the horizontal and vertical directions.

2.5 ST-GCN for Skeleton Based Action Recognition

2.5.1 Processing Keypoints

In the next phase, before we input them into the **ST-GCN** model, we propose two different processing keypoints techniques to enhance the input:

- Technique 1: In the training process, we keep the number of frames constant for each action. If the video has an extra frame, it will be cut; or else if the video is missing some part, it will be processed to fill in the missing frames by interpolation as follows: Put all the

action's frames in a list with a predetermined number of frames for each action. With the first and last frames appended to the top and bottom of the list, the remaining frames of the action will be evenly placed in the positions in the list and retain the original order between frames. Finally, we go through the list in order to find the empty frames and fill them with the value of the previous frame.

- Technique 2: We use data augmentation techniques such as shifting to increase the diversity and volume of the data, allowing the model to perform better in learning.

After processing the above technique, we process the keypoints sequences of each processed frame into the ST-GCN model to train and predict the martial art movements.

2.5.2 ST-GCN Overview

Deep neural networks are a relatively new approach to automatically capturing the patterns stored in the joint's spatial configuration and also their temporal dynamics. Nevertheless, as previously said, the skeletons take the form of graphs rather than 2D or 3D grids, making it challenging to apply tried-and-true models like convolutional networks. The use of GCNs to represent dynamic graphs spanning large-scale datasets, like as human skeletal sequences, has not yet been examined. By extending GCNs to a spatial-temporal graph model known as ST-GCN for Skeleton-Based Action Recognition.

In Fig.8, body joints are shown by blue dots. The organic connections in human anatomy are used to establish the connection edges between body joints. Edges will link successive frames, linking joints in the same position. The position of joints are sent to the ST-GCN as inputs. Edges are classified into two types: spatial edges that adhere to the inherent joint connectedness and temporal edges linking the same joints across successive time steps. On top of it, information may be merged in not just the spatial but also the temporal dimensions due to the construction of many layers of spatial temporal graph convolution.

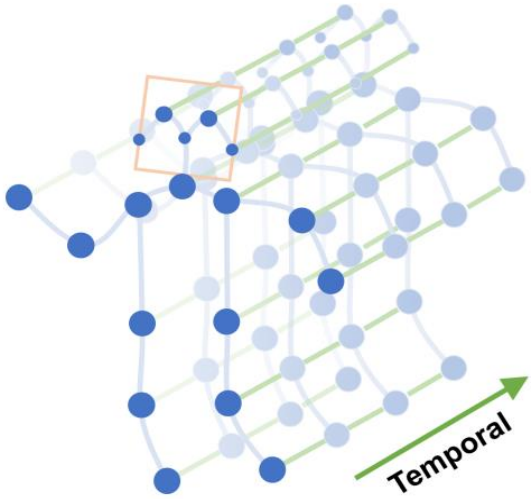


Fig. 8. A skeleton sequence's spatial-temporal graph [14]

The data is typically a sequence of frames, with each frame including a set of joint coordinates. Given a sequence of body joint coordinates in 2D or 3D, The joints served as the graph's nodes, while the inherent connectivities between human body structures and time served as the network's edges. As a result, The vector of joint coordinates in the graph nodes is the input of the ST-GCN. Just like in image-based CNNs, the input consists of pixel intensity vectors that are analyzed on a 2D picture grid. The approach employs numerous ST-GCNs layers to the input data to generate higher-level activation mappings on the graph. The basic SoftMax classifier will then assign it to the appropriate action category. Backpropagation is used to train the entire model from beginning to end.

2.5.3 Implementing ST-GCN

Graph-based convolution is more difficult to implement than 2D or 3D convolution. In this section, we go over how to build ST-GCN. An adjacency matrix A represents intra-body connections of joints inside a single frame, while an identity matrix I represents self connections. The following formula may be used to implement ST-GCN utilizing the first partitioning strategy in a single frame.

$$f_{out} = \Lambda^{-\frac{1}{2}}(A + I)\Lambda^{\frac{-1}{2}}f_{in}W, \qquad (1)$$

In which $\Lambda^{ii} = \sum j(A^{ij} + I^{ij})$.The weight vectors of numerous output channels are layered to generate the weight matrix W in this case. In practice, we can express the input activation map as a tensor of (C, V, T) dimensions in the spatial temporal situations. The graph convolution is realized by executing a regular $2D$ convolution and multiplying the resulting tensor by the normalized adjacency matrix $\Lambda^{-\frac{1}{2}}(A + I)\Lambda^{\frac{-1}{2}}$ on the second dimension.

This implementation is used for partitioning algorithms with various subsets, such as distance partitioning and spatial configuration partitioning. However, the adjacency matrix has now been disassembled into numerous matrices Aj, where $A + I = \Sigma_j A_j$. In the distance partitioning strategy, for example, $A_0 = I$ and $A_1 = A$. Eq. 1 is converted.

$$f_{out} = \sum_j \Lambda_j^{-\frac{1}{2}} A_j \Lambda_j^{-\frac{1}{2}} f_{in} W_j, \qquad (2)$$

In addition, $\Lambda^{ii}_j = \Sigma_k(A^{ik}_j) + \alpha$, $\alpha = 0.001$ here to avoid having empty rows in Aj.

The learnable edge significance weighting is simple to implement. The learnable weight matrix M is paired with each adjacency matrix. Furthermore, replace the matrices $A + I$ in Eq. 1 and A_j in A_j in Eq. 2 with $(A + I) \otimes M$ and Aj $\otimes$ M, respectively. Here, signifies the element-wise product of two matrices. The mask M is set up as an all-one matrix.

Training and Network Architecture. Because the ST-GCN distributes weights between nodes, it's crucial to maintain uniformity in the input data scale across joints. To normalize data, first pass input skeletons to a BN layer in our studies. Nine layers of ST-GCN units make up ST-GCN. Each of the first three layers has 64 output channels. The output channels for the next three layers total 128. The final three levels have 256 output channels. There are 9 temporal kernel sizes in these levels. Each ST-GCN unit employs the Resnet technology. To avoid overfitting, we randomly drop out the features with a 0.5 probability after each ST-GCN unit. As a pooling layer, the fourth and seventh temporal convolution layers' strides are also set to 2. Global pooling was then applied to the resultant tensor to create 256-dimensional feature vectors for each sequence. They are then submitted to a SoftMax classifier. Stochastic gradient descent with a learning rate of 0.01 is used to train the models. After every ten epochs, the learning rate was reduced by 0.1. When training on

the Kinetics dataset, two types of augmentation were used to replace dropout layers in order to avoid overfitting. To begin, apply random AFT to the skeletal sequences of all frames to simulate camera movement. To create an AFT from the first to the last frame, two random two-factor combinations of defined angle, scaling, and translation parameters were chosen as options. To provide the appearance of a smooth view point movement during playback, this transformation is interpolated for intermediate frames. Random movement is the name given to this enhancement. In addition, just a small portion of the original skeleton sequences were utilized for training, while all frames were used for testing. The network can handle input sequences of any length due to global pooling at the top.

3. EXPERIMENTS RESULT AND CONCLUSION

3.1. Data Collection

We gathered data by recording martial arts exercises of university students. We have selected the 9 most basic movements of the practice. This is a core subject in the program so recording is quite easy. The data is then labeled with martial arts moves. The data results are listed in Table 1.

Table 1. Data statistics

Type	Value
Number people	10
Number class	9
Total number action	778

The classes include standing still, defense and 7 basic martial arts movements of Vovinam. The following is an illustration of two actions (Fig. 9, Fig. 10).

Fig. 9. Scooping movement (đấm múc)

Fig. 10. Double punch movement (song đấm)

The procedure and process for collecting data are separated into two primary phases as follows:

Stage 1: We collect sufficient fundamental data. Afterward, employ them to train the model and evaluate its efficiency. The parameters may then be calculated and adjusted from there. The data set was divided into 2 turns with 6 students from the Vovinam club at FPT University as the main object. Turn 1 is 2 people and turn 2 is 4 people simultaneously recording. We kept a record of 16 repetitions for each motion. Data in this phase 1 includes 5 basic martial arts movements in Vovinam.

Stage 2: After data collection and complete training of the model which shows positive results. With the help of four other FPT University students who are also members of the Vovinam club, we continue to collect more data. We added 16 reps to each of the previously performed movements, and we added 32 reps to each of the four new moves. In order to improve the model's performance and demonstrate its ability to detect a variety of movements, it is intended to evaluate if the model can learn additional information as well as to add new movement classes. From there, future proofing with our proposed pipeline can be developed into a product that helps learners to learn vovinam, support self-training through our model-integrated application software.

To summarize, the data set that we contribute towards the key movements that are important to students who are learning and practicing Vovinam. With the defensive posture being the

ready position of all the main moves, and the data set records basic martial arts moves that any student can learn and practice through. Therefore, this set of training data for the Spatial Temporal Graph Convolution model can be considered as a basic introductory video tutorial for students to learn and practice Vovinam.

3.2. Experiments

We tested the problem on GPU-3090, Cuda 11.3, Ubuntu 22.04, Python 3.7. The steps that we setup environment to experiment and import data through each block of the pipeline are described as follows:

Step 1: Install Anaconda

Anaconda for Ubuntu is a distribution of the Python programming languages for scientific computing, that aims to simplify package management and deployment. Using Anaconda, we can easily install a library on a virtual environment.

Step 2: Setup Environment and Install Package

Open Anaconda Prompt, create a virtual environment with the command line:

conda create –name capstone python=3.7

After that we must activate this environment to install the required package.

In the Anaconda prompt we continue with the command line:

conda activate capstone

Continue we Install Pytorch along with CUDA 11.3 with the command line:

conda install pytorch pytorch-cuda=11.3 -c pytorch -c nvidia

The above step will take some time depending on the internet connection speed. Then proceed to install the package libraries that our pipeline uses with the command line:

pip install -r requiments.txt

Step 3: Import video data into the pipeline's model with the configuration parameters

We divided the data into the train and test set in an 80/20 ratio. Each class we take the length of the sequence is 100 frames to put into YoloX with the following parameter settings:

Table 2. YoloX parameters

Parameter	Value
Input Shape	640,640
Score Threshold	0.1
NMS_Threshold	0.7

With YoloX input of (640,640), Score Threshout used to control objects with confidence score below 0.1 will be discarded. NMS Threshold is Non Maximum Suppression to remove the redundant bounding boxes of humans in the image.

Next, we input the human bounding boxes into the byte track to track the bounding boxes of each person with the following parameters:

Table 3. Byte-track parameters

Parameter	Value
Track threshold	0.7
Track buffer	30
Match threshold	0.8

In table 3, we have Track threshold which is the threshold level to divide the bounding boxes into 2 sets, the set with high confidence score and the set low confidence score. The track buffer is the maximum number of frames that the Track can stay offline (Track is a sequence of bounding boxes that are matched against an object, Track offline is when the object that the track points to is not present). The Track will be erased if the Offline Track's frame count exceeds the track buffer. The level at which two bounding boxes match is known as the match threshold.

We then use the architecture pose_hrnet_w32 pretrained on the COCO dataset to feed the image sequence of human bounding boxes forward the HRNet-Pose model with the input (384,288).

After this step we get the sequence of keypoints of each person, enter the keypoint sequence processing stage with many selections and edits, then finally we have selected 2 processes. Using them to process helps improve the quality of keypoints to help the ST-GCN model training process in the following section achieve better results.

For evaluating a system's performance, this work uses Top-1 accuracy. The configuration for the ST-GCN is described in Table 4.

Table 4. The parameters of ST-GCN model

Parameter	Value
base_lr	0.1
batch_size	32
num epoch	100
optimizer	SGD
weight_decay	0.0001

On the Kinetics and NTU-RGB+D datasets, we fine-tuned the pretrained model ST-GCN. Then train the model with a learning rate of 0.1, batch_size of 32 and number of epochs of 100. We adopt the SGD optimizer with the same hyper-parameters and also use a weight decay of 0.1.

3.3. Result and analysis

We have tested two cases as suggested in the pipeline and the results are listed in Table 5 as well as detailed in Fig. 11 and Fig. 12.

Table 5. Results of two cases with and without processing keypoints

Model	Top1-Accuracy	Mean loss
Without processing keypoints	94.62%	0.065
With processing keypoints	99.23%	0.047

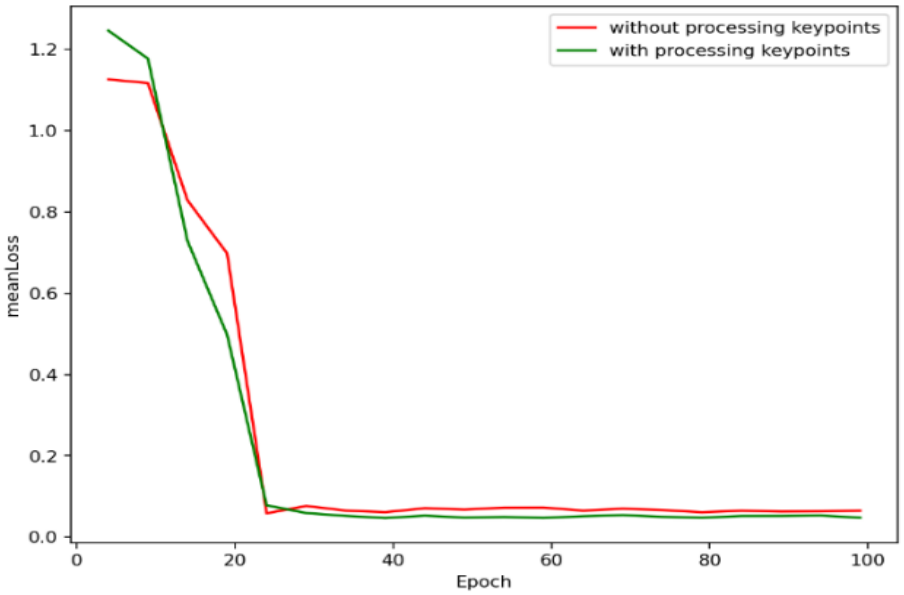


Fig. 11. The meanLoss.

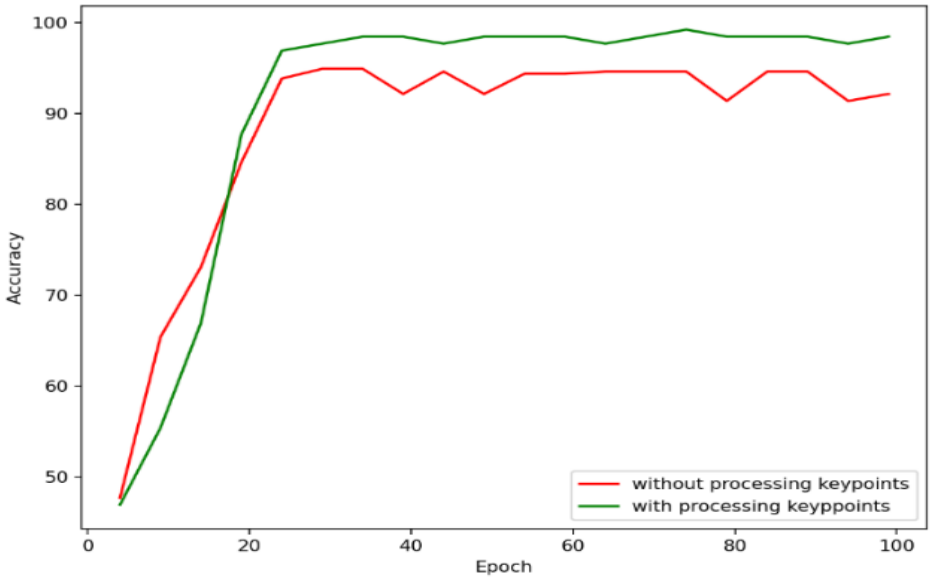
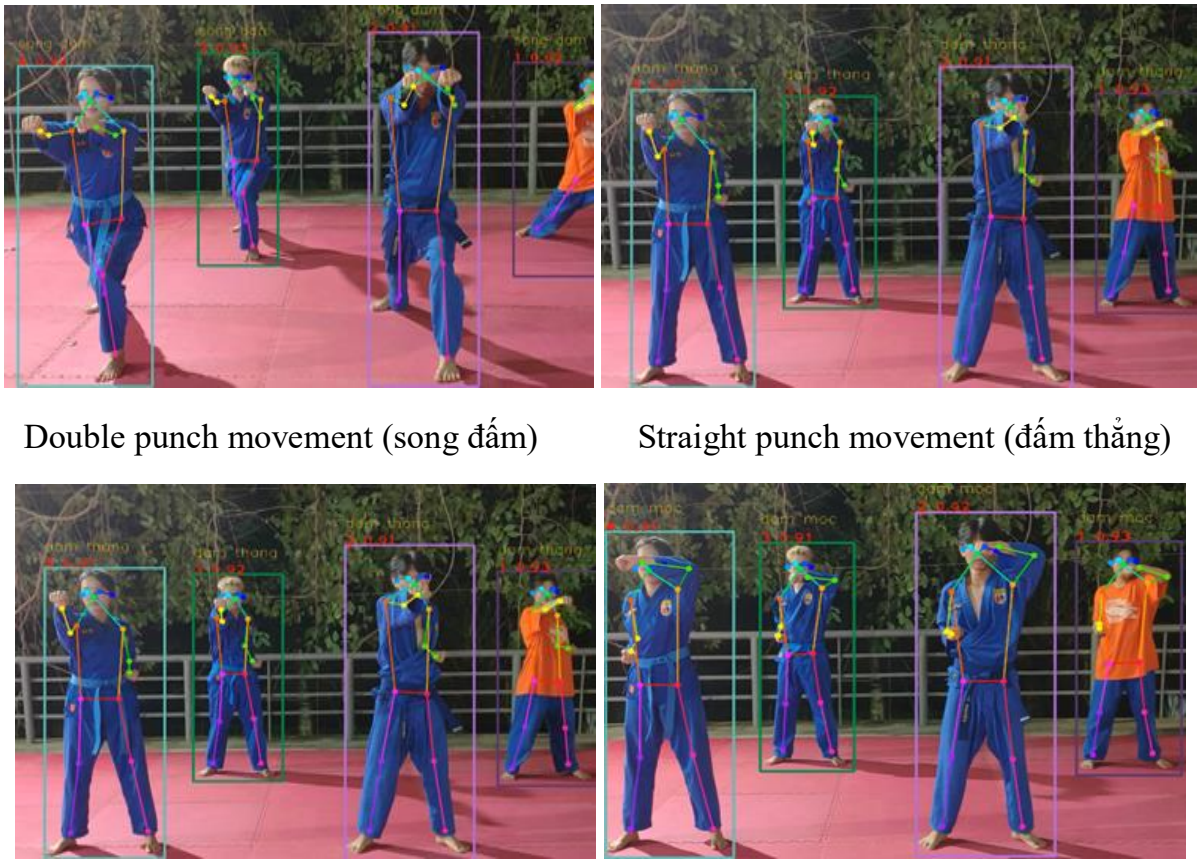


Fig. 12. The accuracy

The above results show that the proposed keypoints processing has significantly improved the accuracy from 94.62 to 99.23%. The improved results are relatively stable after the 25th epoch, which also proves that the efficiency of the keypoints processing is stable. Fig. 13 are a few examples of Vovinam movement recognition:



Double punch movement (song đấm) Straight punch movement (đấm thẳng)

Straight punch movement (đấm thẳng) Hook punch movement (đấm móc)

Fig. 13. Some examples of Vovinam movement recognition

3.4. Inference

Carrying out the implementation on a long video including a large number of motions, we use the number of frames to predict a martial art move containing 100 frames that slide across the video. To be specific, the result of the move will also be the result of the above 100 frames-prediction model. To make our projection become more apparent, we provide an example: The result displayed on the screen at the exact 100th frame is the outcome of the prediction of frame 1 to frame 100, this happens after we have put our proposed classify model into the pipeline.

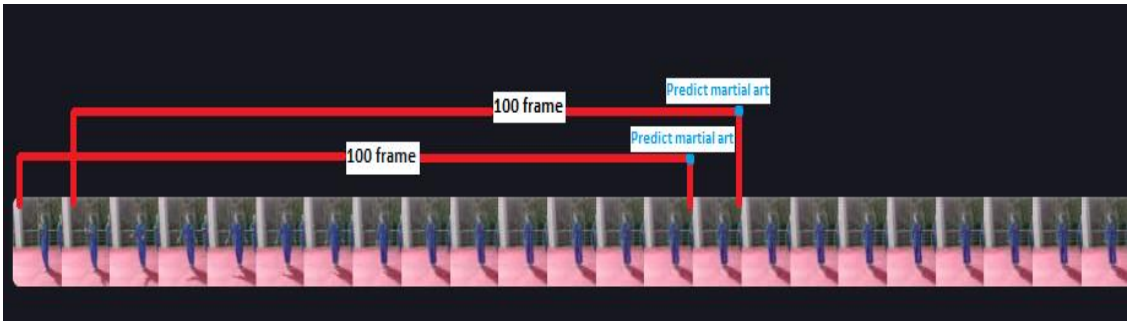


Fig. 14. Predictive process of the model on long video

The results, which can be easily visually observed, show that our model is correct in classifying the movement throughout the video. This would act as a firm verification that the model can be feasibly developed into a useful product for martial artists, who are likely to utilize the outcome.

3.5. Conclusion and Future Works

This study proposed a pipeline for identifying martial arts movements in Vovinam, a Vietnamese martial art. Each phase's appropriate approaches are thoroughly considered and selected from the most recent methods. The data was gathered and labeled, including nine classes separated into three categories: standing still, defense, and basic martial arts movement. A new processing phase for keypoints is added to enhance input for the ST-GCN model. Based on the collected dataset, we fine-tuned the loss function and parameters accordingly to achieve more accurate results as well as increase performance. Outstanding results when handling keypoints with 99.23% accuracy.

In future work, we aim to collect more martial arts lesson data to develop this method in an educational setting. Currently, this thesis mainly focuses on some movements from the dataset that we collected from recording videos of students belonging to FPT University's vovinam club. The idea is to look at models that work well with complex movements in martial arts. From there the model can be further developed on a complete database of all

lessons, movements. The development of the application into a mobile application product in order to provide the easiest support for learners is also worth paying attention to.

References

1. Vietnam's vovinam takes on the world, https://vietnamnet.vn/en/vietnams-vovinam-takes-on-the-world-E217255.html (Accessed 10/10/2022)
2. Dung, V.V., Vy, B.T.K, My, P.T.: The role of Vovinam in the life of Vietnamese people. In European Journal of Physical Education and Sport Science, ISSN: 2501 - 1235 (2016)
3. Malaguti, A., Carraro, M., Guidolin, M., Tagliapietra, L., Menegatti, E., Ghidoni, S.: Real-time Tracking-by-Detection of Human Motion in RGB-D Camera Networks. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 3198-3204, doi: 10.1109/SMC.2019.8914539 (2019)
4. Atvar, A., Cinbiş, N. İ.: Classification of human poses and orientations with deep learning. In: Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, doi: 10.1109/SIU.2018.8404498 (2018)
5. Solbach, M. D., Tsotsos, J. K.: Vision-Based Fallen Person Detection for the Elderly. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 1433-1442, doi: 10.1109/ICCVW.2017.170 (2017)
6. Samir, M. A., Maged, Y., Atia, A.: Exam Cheating Detection System with Multiple-Human Pose Estimation. In: Proceedings of the IEEE International Conference on Computing (ICOCO), pp. 236-240, doi: 10.1109/ICOCO53166.2021.9673534 (2021)
7. Pan, H., Yin, J., Ku, H., Liu, C., Feng, F., Zheng, J., Luo, S.: Fighting Detection Based on Pedestrian Pose Estimation. In: Proceedings of the 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1-5, doi: 10.1109/CISP-BMEI.2018.8633057 (2018)
8. Das, S., et al.: Quantitative measurement of motor symptoms in Parkinson's disease: A study with full-body motion capture data. In: Proceedings of the Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 6789–6792 (2011)
9. Um, T.T., Babakeshizadeh, V., Kulic, D.: Exercise motion classification from large-scale wearable sensor data using convolutional neural networks. In: Proceedings of the IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), pp. 2385–2390 (2017)
10. Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., Badia, S.B.I.: Learning to assess the quality of stroke rehabilitation exercises. In: Proceedings of the 24th Int. Conf. Intell. User Interface, pp. 218–228 (2019)

11. Pang, Y., Wang, Q., Zhang, C., Wang, M., Wang, Y.: Analysis of Computer Vision Applied in Martial Arts. In: Proceedings of the 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 191-196 (2022)
12. Cai, Z., Yang, Y., Lin, L.: Human action recognition and art interaction based on convolutional neural network. In: Proceedings of the Chinese Automation Congress (CAC), pp. 6112-6116 (2020)
13. Thanh, N.T., Tuyen, N.D., Dung, L., Cong, P.T.: Implementation of technical data analysis of skeleton extracted from camera kinect in grading movements of Vietnamese martial arts. In: Proceedings of the International Conference on Advanced Technologies for Communications (ATC), pp. 241-244 (2017)
14. Yan, S., Xiong, Y., Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In CoRR abs/1801.07455 (2018)
15. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding YOLO Series in 2021. In CoRR abs/2107.08430 (2021)
16. Xiaosong, Z., Fang, W., Chang, L., Rongrong, J., Qixiang, Ye.: Freeanchor: Learning to match anchors for visual object detection. In NeurIPS (2019)
17. Joseph, R., Ali, F.: Yolo9000: Better, faster, stronger. In CVPR (2017)
18. Joseph, R., Ali, F.: Yolov3: An incremental improvement. In arXiv preprint arXiv:1804.02767 (2018)
19. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In book: Computer Vision – ECCV 2022 (pp.1-21).
20. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5686-696, doi: 10.1109/CVPR.2019.00584 (2019)