Support Learning Vovinam Exercises based on Computer Vision

Pham Son Tung, Thai Thanh Do, Pham Hong Giang
Instructor Phan Duy Hung
FPT University, Hanoi, Vietnam
TABLE OF CONTENTS

1. INTRODUCTION
2. METHODOLOGY
3. EXPERIMENTS RESULT AND CONCLUSION
INTRODUCTION
Overview

- 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.

- Along with a tremendous amount of time on self-training, the application of computer vision in supporting self-training will be very effective and convenient.
Related works

• The rise of new research on human body skeleton dynamics contributes an important part for recognizing human actions.

• For grading movements of Vietnamese martial arts, Tuong Thanh et al proposes the implementation of data analysis of the depth of scoring Kinect users' camera movements in grading Vietnamese traditional martial arts movements.
Contribution

- Our goal in this thesis is to develop an end-to-end action recognition model that leverages ST-GCN kernels with several modifications and some recent computer vision techniques to deal with the complex moves of vovinam martial arts.

- This study suggests a pipeline where 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.
METHODOLOGY
Overview pipeline

Block diagram for classifying martial art action
YoloX

- We chose YoloX for this model because it improved object detection by emphasizing anchor-free detectors and had better detector performance than the Yolo-series.

- And it is also recommended that Yolox be used with the byte-track method for better performance.
Byte-Track

- Most of previous approach identify identities by identifying 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.

- Byte-track tracks nearly every detection box, not just the highest-scoring ones.
HRNet-Pose

- In contrast with most previous systems that were linked in a series, high-to-low resolution subnetworks are linked in parallel.

- In this method use recurrent multiscale fusions, high-resolution and low-resolution will be augmented representations. As a result, the anticipated heatmap may be more accurate.

Illustrating the architecture of the proposed HRNet
Processing Keypoints

- Technique 1: Interpolation missing frames.

- Technique 2: data augmentation techniques.
  - 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.
ST-GCN

- Relatively new approach to automatically capturing the patterns stored in the joint's spatial configuration and also their temporal dynamics.
- 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.
$l_{ti}(v_{t,j}) = \begin{cases} 
0 & \text{if } r_j = r_i \\
1 & \text{if } r_j < r_i \\
2 & \text{if } r_j > r_i 
\end{cases}$

1. The root node itself.
2. Centripetal group: the neighboring nodes that are closer to the gravity center of the skeleton than the root node.
3. Otherwise, the centrifugal group.

$l_{ST}(v_{q,j}) = l_{ti}(v_{t,j}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K$
Implementing ST-GCN

**GCN**

\[ f_{out} = \Lambda^{-\frac{1}{2}} (A + I) \Lambda^{-\frac{1}{2}} f_{in} W \]

**ST-GCN**

\[ f_{out} = \sum_j \Lambda_j^{-\frac{1}{2}} A_j \Lambda_j^{-\frac{1}{2}} f_{in} W_j \]

The graph convolution is implemented by performing a $1 \times \Gamma$ standard 2D convolution and multiplies the resulting tensor with the normalized adjacency matrix $\Lambda^{-\frac{1}{2}} (A + I) \Lambda^{-\frac{1}{2}}$ on the second dimension.

**Network architecture and training**
EXPERIMENTS RESULT AND CONCLUSION
Data Collection

Table 1. Data statistics

<table>
<thead>
<tr>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number people</td>
<td>10</td>
</tr>
<tr>
<td>Number class</td>
<td>9</td>
</tr>
<tr>
<td>Total number action</td>
<td>778</td>
</tr>
</tbody>
</table>

Examples of Vovinam movement recognition
Experiments

Environment

- GPU-3090, Cuda 11.3, Ubuntu 22.04, Python 3.7

Setup virtual environment

- Step 1: Install Anaconda.
- Step 2: Setup Environment and Install Package.
  
  conda create -n capstone python=3.7

  conda activate capstone

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

  pip install -r requiments.txt

- Step 3: Import video data into the pipeline’s model with the configuration parameters.
### YoloX
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Shape</td>
<td>640,640</td>
</tr>
<tr>
<td>Score Threshold</td>
<td>0.1</td>
</tr>
<tr>
<td>NMS Threshold</td>
<td>0.7</td>
</tr>
</tbody>
</table>

### ByteTrack
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track threshold</td>
<td>0.7</td>
</tr>
<tr>
<td>Track buffer</td>
<td>30</td>
</tr>
<tr>
<td>Match threshold</td>
<td>0.8</td>
</tr>
</tbody>
</table>

### ST-GCN
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>base lr</td>
<td>0.1</td>
</tr>
<tr>
<td>batch size</td>
<td>32</td>
</tr>
<tr>
<td>num epoch</td>
<td>100</td>
</tr>
<tr>
<td>optimizer</td>
<td>SGD</td>
</tr>
<tr>
<td>weight decay</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
## Result and analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Top1-Accuracy</th>
<th>Mean loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without processing keypoints</td>
<td>94.62%</td>
<td>0.065</td>
</tr>
<tr>
<td>With processing keypoints</td>
<td>99.23%</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Results of two cases with and without processing keypoints
The meanLoss
The accuracy
Examples of Vovinam movement recognition
Inference

Predictive process of the model on long video
Conclusion & 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.

- 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.