Project report

Jiahao (jiahao4), Zeyu Liao(zeyu9)

Our project focuses on implementing a hand gesture recognition model and its applications, based on the paper "Deep learning for hand gesture recognition on skeletal data" [1].

**Motivation and impact:**

Human action recognition is a challenging but essential task in computer vision, with a wide range of potential applications, such as health care and military applications. Within the field, we focus on hand gesture recognition this specific area. Deep neural networks have shown promising results in this field and combining them with computational photography techniques can lead to exciting new possibilities. We chose this topic, to learn and implement natural human gesture recognition so that we can utilize it in the online e-commerce industry. We hope to learn how computational photography techniques that we learned in CS445 can create values combined with deep neural networks.

**Approach:**

Our project can be divided into three main parts: pre-processing data, training the model, and developing applications. We experimented with two datasets mentioned in the paper and selected the one with better performance. The paper employs two models, Kera and PyTorch, and we implemented PyTorch based on the paper's methodology. While we referenced parts of the paper's code, we did not copy it entirely. For the application, we began by developing an easier version of the hand gesture recognition, which focused on still image recognition. Subsequently, we expanded our work to include hand gesture recognition based on video. After this we implemented gesture recognition based on leap motion.

**Results:**

Our model achieves about same accuracy on the training and test sets compare with the paper’s result (Figure 1).

Unfortunately, our real word hand gesture detection does not work very well. Our results indicate that our implementation of hand gesture is not very accurate. Our model will give several prediction results. Results include the correct result, but the correct result is not the most common one (Figure 3). For the overall performance, we have include video demos in the Video section.

**Implementation details:**

In this project, we use Jupyter Notebook, Python, Unity, and use following packages: sklearn, glob, numpy, pickle, ndimage, train\_test\_split, itertools, torch, sklearn.utils, time, math, PIL, torchvision.transforms, and tensorboardX. For the dataset, we pre-processed DHG-14/28 Dataset and the SHREC’17 Track Dataset, but we end up with using SHREC’17 Track Dataset for the model [2] (two .pckl file in the Github repository). For the part that we referenced the original paper’s code, we have label them in the code. We implement a neural network model for gesture detection. The input to the model is a tensor with data type 'tensor' and shape (batch\_size, duration, n\_channels), where each batch consists of hand skeletons for a specific duration of time. Each hand skeleton will have 22 joints, with 3 channels. To extract features from the input data, we will first process each channel separately. We will use 1D convolutions to process each channel, and the neural network will consist of three convolutional layers and pooling layers. The output of each convolutional layer will be concatenated into a single output, which will be used as input for the next layer. Finally, the three outputs will be concatenated into one output.

The neural network architecture for this model can be summarized as follows:

* Input layer with shape (batch\_size, duration, n\_channels)
* Three 1D convolutional layers with padding, followed by a max pooling layer
* Three output layers for each channel, with a concatenation layer at the end

By using 1D convolutions and pooling layers, we can extract meaningful features from the hand skeleton data. The concatenation layer at the end allows us to combine the information from each channel into a single output, which can be used to predict the gesture being performed.

**Challenge / innovation:**

Implement the paper itself is a very challenge task.

One innovation is we use SHREC’17 Track Dataset, instead of DHG dataset. Because we have comparted the result of two dataset and SHREC’17 Track Dataset gives better results.

Another innovation is that the original paper’s model only supports 2D/3D skeleton images, we make it can detect normal images. We have two version of gesture recognition. The simple version will take images, save the image, and we need to manually input images to the model (demo video link1). The other version is based on the video frame, it will automatically detect the gesture in the frame and display result on the window (demo video link2).

**Demo Video:**

Link1: https://youtu.be/9D4GuIGvNoo

Link2: https://youtu.be/YBGKM17EmQ8

Link3:

Table

Description automatically generated

Figure 1: Model accuracy

Text, letter

Description automatically generatedText, letter

Description automatically generated

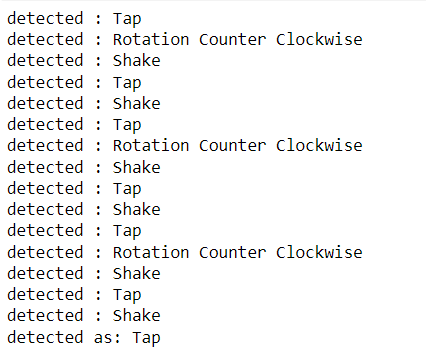
Figure 2: Evaluation of the model

Figure 3: Real world image detection result on Rotation Counterclockwise

(Detected as indicate the most common detect result)

**Citation:**

**[1]** Devineau, Guillaume, Fabien Moutarde, Wang Xi, and Jie Yang. "Deep learning for hand gesture recognition on skeletal data." In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 106-113. IEEE, 2018.

**[2]** Dynamic Hand Gesture Recognition using Skeleton-based Features ,Quentin De Smedt, Hazem Wannous and Jean-Philippe Vandeborre, 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Download from http://www-rech.telecom-lille.fr/DHGdataset/ and unzip into ./415-finalproject/dataset\_dhg1428