2D Skeleton-Based Dance Genre Recognition Using CNN

Zikui Cai
UC Riverside
zcai@engr.ucr.edu

Ruiwen Zhao
UC Riverside
rzhao@engr.ucr.edu

Xiaodi Fan
UC Riverside
xfan027@engr.ucr.edu

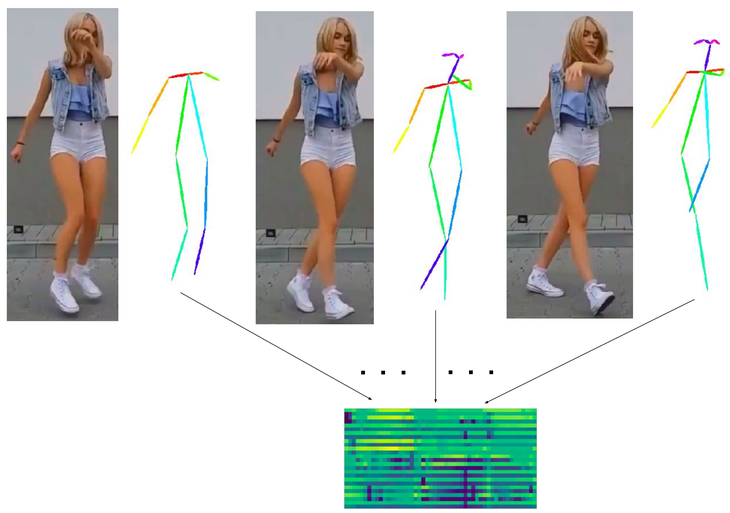
Introduction

Data mining is a process of extracting interesting yet previously unknown knowledge from a massive set of data. Due to the rapid developments in multimedia acquisition and storage technology, the amount of multimedia data available to users has increased and is increasing exponentially. There is a great potential for video-based data mining applications in many areas including security and surveillance, personal entertainment, sports which contain human activities. Being able to understand the actions of human is crucial to the efficient management of video data.

Recent research in data mining has offered many effective solutions to this problem. One of the most notable methods being utilized today is Convolutional Neural Network (CNN). It can detect articulated human pose or skeleton whose movement is a very good representation of human actions. In addition, CNN can also learn the patterns from the sequences of skeletons which are detected by CNN itself from the raw data.

Proposed Method

1. Extract 2D points using CNN
2. Represent a sequence of skeletons as a picture
3. Image Classification using CNN



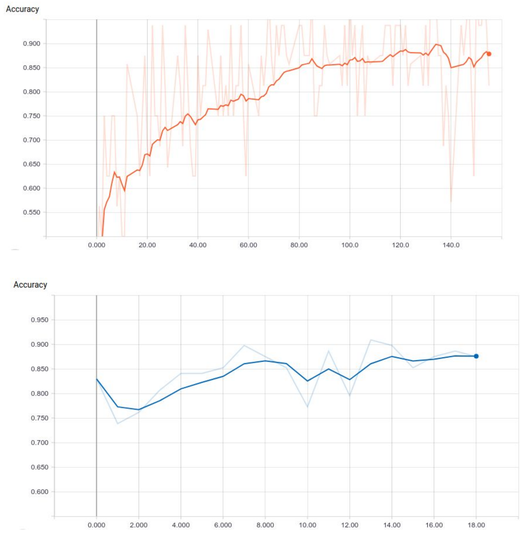
Problem Definition

To make this problem more challenging, we make our own dataset of street dance which is composed of a variety of articulated actions and movements. The videos are collected from youtube and each video is splitted into several 2-second subclips with continuous motion. There are around 100 subclips in each of the 4 categories.

shuffling breaking locking popping

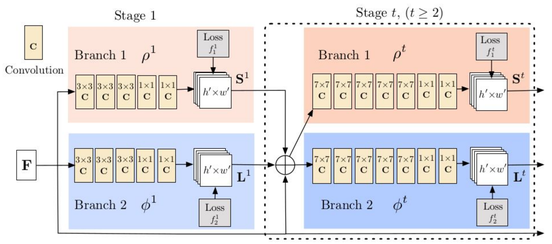
Results

In 10 epochs, training accuracy and testing accuracy reached around 87.5%.

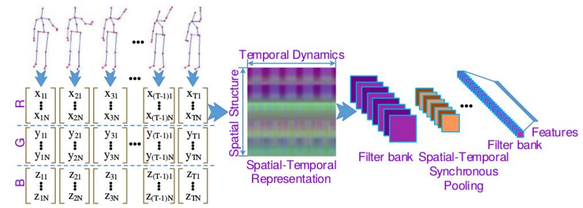


Related Work

Wei et al. introduced Convolutional Pose Machines which consist of a sequence of convolutional networks that produce increasingly refined estimates for part locations.



Yong et al. proposed a way to represent a sequence as an image; however, their work based on 3D positions with depth information.



We generalized their work to real life situation where only 2D positions can be obtained.

Conclusions

In this project, we leveraged CNN to effectively extract human skeletons and to classify images. In addition, we proposed a new way to represent the 2D skeleton points sequence and verified the validity of this approach through our experiment.

The accuracy and speed of human skeleton extractor largely determine the quality of skeleton-based human action recognition. While using the Convolutional Pose Machine to extract the keypoint, we noticed this method is quite slow because it took 5 seconds to process a normal 1080p video frame on Nvidia 1080ti. Besides that, the skeletons in quite a few frames of the video were not extracted perfectly, which influenced the accuracy of the following method.

Our future work will be on improving the accuracy and speed of the model to make it usable in real life human action recognition.

UNIVERSITY OF CALIFORNIA UCRIVERSIDE