Project Proposal:
A Method for Human Video Motion Transfer

Zheyu Zhang
E-Mail: Zheyu.Zhang@campus.lmu.de
Matriculation number: 12217492

Jinhe Bi
E-Mail: Jinhe.bi@campus.lmu.de
Matriculation number: 12524569

1 Problem Definition

Human video motion transfer concerns the creation of videos in which individuals mimic one another's movements, which has attractive potential applications in movies, interactive games, virtual shopping, etc. With the development of Generative Adversarial Networks (GANs) (Goodfellow et al. (2014)) and GAN-based image-to-image translation techniques (Wang et al. (2018)), recent advances in human motion transfer have achieved great success.

But in order to generate videos for another person, many recent methods must gather new data and train new models, which is extremely time consuming and expensive. Our project idea aims to achieve: transfer a person's sporting performances to a novel (amateur) target after observing the subject perform standard movements for just a few minutes by learning a simple video-to-video translation.

2 Dataset

For training and evaluation we plan to use the part of the sporting videos from Youtube and OpenPose, which has fixed, clean background, contains a single person performing ”basic” actions and can be splitted into frames, e.g Afrobeats Dance Workout.

3 Approach

The project is in the broad scope of the topic ”Image-to-Image Translation” and ”Video Synthesis”. In order to achieve the above purpose, i.e. automatically transferring the motion from a source to a target subject, we have to extract poses from the source subject and apply the learned pose-to-appearance mapping to generate the target subject.

Combining what we have learned in class, We intend to focus on the GANs-based image-to-image generation process(Wang et al. (2018)). Based on single-frame image generation, we hope to work on large amounts of personalized and detailed high-resolution video. For a higher resolution, we can design different specialized GAN to add more details with a smaller section of the image as input. To achieve this goal, we plan to initially divide our method into two parts: pose detection and video generation. See Figure. 1 for an overview of the pipeline.

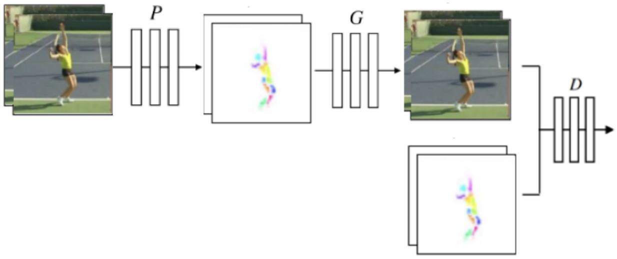

Figure 1: Pipeline

3.1 Pose Detection

In this part, to encode the pose of the input image, we can use a pre-trained pose detector(e.g. OpenPose). Then we can plot the keypoints from the estimated joint coordinates and finally draw a pose stick figure for further generation. From different frame-to-frame in different videos, the people may hold a closer of farther distance from the camera or different body proportions. So we should apply a normalization process before the GANs.

3.2 Video Generation

For an adversarial single frame generation process, which the generator synthesize images in order to fool the discriminator D, it is not suitable for video synthesis since they may cause temporal artifacts or lack some details in motion. So we want to add several specialized GAN for different parts of human body to keep a high-resolution.

4 Evaluation and Expected Results

We can compare our results to some baseline methods by using multiple target and motions. By collecting multiple types of self-making videos, we can use the nearest neighbors as a baseline method. With the L2 distance between the corresponding joints, we can retrieve the closest match in the training target, etc.

5 Hardware

CPU or 6GB GPU from CIP pool.

References

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.