CSCI 1430 Final Project Report:
James is Dancing Now

GANNON: Darius Atmar, Zejiang Shen, Yueyi Sun
Brown University
May 10, 2019

Abstract

Motion transfer between objects in a video stream has long been regarded as a difficult task. But Deep Learning (DL) techniques like Generative Adversarial Nets (GAN) help to eliminate these difficulties. In this project, we reproduce a DL-based method in [3] that transfers a source object's dance to another target object. We verify the model results by transferring the motion from Bruno Mars' dance to Professor James' movements. We improve the pose normalization method to achieve high quality outputs. Several other experiments are conducted based on transferring Professor James' dance to our teammates' movements.

1. Introduction

Watching your friend or professors dancing like someone else is fun! Do you ever wonder what it would be like to dance like a professional?

The increased amount of research in deep learning today has enabled creative applications in image translation and video translation tasks, the performances of which are poor using traditional tasks. In this project, we utilize the result of a novel deep learning model called *Everybody Dance Now*[3] to perform a video to video translation task. Specifically, we transfer the motion of a person in a source video, which is a dance in this case, to that in a target video. The dance in the source video is usually of a professional quality, thus hard for an unskilled dancer in the target video to perform. The final result is a fun and novel video of the target person as the output of the model. Moreover, we verify this concept on videos containing our teammates and Professor Tompkin.

2. Related Work

This approach involves two components, namely, pose estimation and image-to-image translation. Hence, we describe some background work for them.

Pose Estimation Human pose estimation is an challenging and important visual task. It requires a model to find a set of pre-defined key points for human based on the image inputs. Sometimes an estimation of 2D coordinates of the key points is sufficient, while 3D real-world coordinates are needed in some cases. And there could be multiple people in the input image. It sets higher demands for the models, which should be able to discriminate between different people. Deep-Learning based methods show promising results in this field during the past few years. DeepPose[6] is a network that employs a cascade of regression task to predict and refine each key point. Using the activation mask is another intuitive idea. Convolutional Pose Machines [8] is one such method that uses a multi-stage approach to generate multiple channel, which is a activation mask for a key point for the object in the input image, respectively. Furthermore, works like [2] can deal with multiple poses in the input in real time.

Image to Image Translation People want to generate an output image based on an input image for different purposes, e.g., style transformation, photo enhancement, face swap, and season transformation. The classic approach is to train the model to use a set of aligned image pairs. However, it is often hard or impossible to obtain the set of aligned image pairs which forces the image to image translation from supervised learning to unsupervised learning. Generative adversarial networks (GANs) have obtained impressive achievements during the recent years to realize the unsupervised image to image translation. Cycle-GAN [10] is one example to use unpaired image to image translation which can successfully realized the transformation of season, style, and animal features. GAN could generate blurring images which may correspond to information from multiple images being combined in the single output. Zhu et al. [11] addressed this issue by combining latent vector as conditional GAN. Another challenge in image to image translation is the multiple target instances. InstaGAN [4] resolved the multiple target challenging by incorporating the instance information.

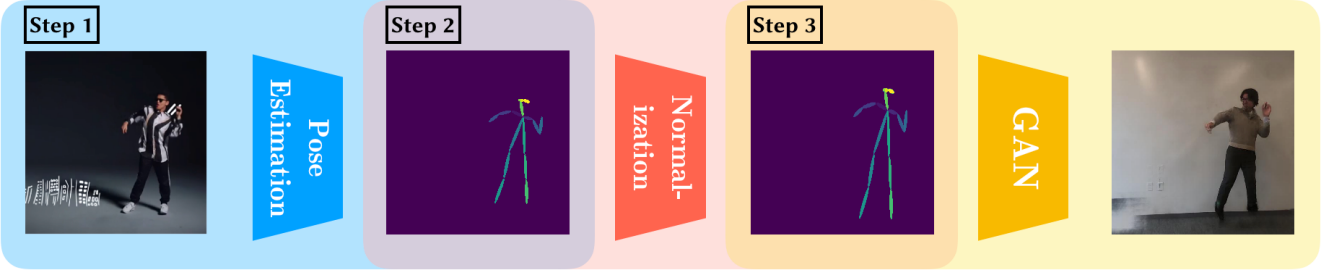

Figure 1. This figure denotes three major steps during the generation process. The steps are, pose estimation, which is generating pose from the source image, pose normalization, which is adapting the pose from the source to the target, and image generation, which is the process for generating the target's dance based on the pose.

3. Method

3.1. Overview

There are three main steps in our project: **Pose Estimation**, **Pose Normalization**, and **Dance Transformation**. The three steps are shown in Figure 1. First, we applied pose estimation on the objects in the selected frames of the source and target videos. Next, we applied pose normalization to resolve the issue of inconsistency of camera distance, body height, and physique. Last, we used pix2pixHD [7] to generate the target dance video. More specifically, we train a GAN on videos of a target object, whose input is the pose estimation mask and output is the generated image. And then we use the GAN to generate the target object's dancing image based on the pose estimation results of the source object.

3.2. Pose Estimation

The pose estimation method we used in this project is based on OpenPose[1, 2, 5, 9]. It is an open source, real-time multi-person keypoint detection library to generate 2D pose for selected frames of the source and target videos.

(a) Missing Feet (b) Unreasonably long legs (c) Wrong movements

Figure 2. Some failure cases during our experiments. *(a)*: The feet in generated video were sometimes missing. *(b)*: The generated object have unreasonably long legs. *(c)*: When the object wants to squat, the floor moves upwards but not objects moves downwards.

3.3. Pose Normalization

Pose normalization is to *center* poses in the input. Due to high variability of the movements of the source object, a direct transfer of the source to the target can be problematic if the source object deviates too much from the center position.

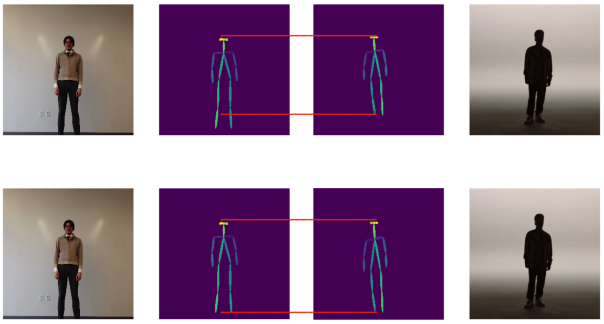

Figure 3. Normalize poses between source and target to account for height differences.

However, with the proposed pose normalization method, we still encounter significant failure cases (shown in Figure 2) in the experiment.

After investigation, we conclude the problem results from the incompatible height of the source and target object. When the source was shorter than the target, the output video resulted in unnatural shortening of the target's legs, and disappearance of the feet. When the source was taller than the target, the output video resulted in unnatural elongation of the target's legs.

The solution to this was to normalize the pose estimations between pose and target by matching the target pose's height to that of the source pose. This was done by calculating the coordinates of the head, along with the height of the pose, and rescaling/resizing the poses to match. Figure 3 denotes the effects of the adaptation method. Note: In the submitted code, however, we only include a rescaling method that requires a manually computed scaling ratio as an input.

3.4. Dance Transformation

We base our dance transformation on pix2pixHD[7]. It is a GAN model that transfers the dancing style from source object to target object. The training process is shown in Figure 4. The dataset used for training is the pair of the target pose key points' mask and the target images. There are

two parts in our model: the generator and discriminator. The input of the generator is the pose estimation and the output of the generator is the generated frame, while the input of the discriminator is the pair of (pose estimation, generated frames) or (pose estimation, real frames). The output of the discriminator equals 1 if the discriminator thinks the pair is from real frame, 0 otherwise. Our generator is trained to generate true-to life frames which can fool the discriminator. The discriminator is trained to distinguish the input pairs better.

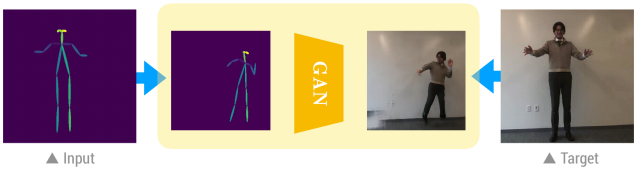

Figure 4. This figure denotes the training inputs and outputs of the GAN model.

4. Results

4.1. Transfering Bruno's Dance to James's Movements

In the first part of our project, we transferred Bruno's dance to James's movements. The generated video has high motion continuity and dance simulation. Figure 5 shows one frame from our generated video.

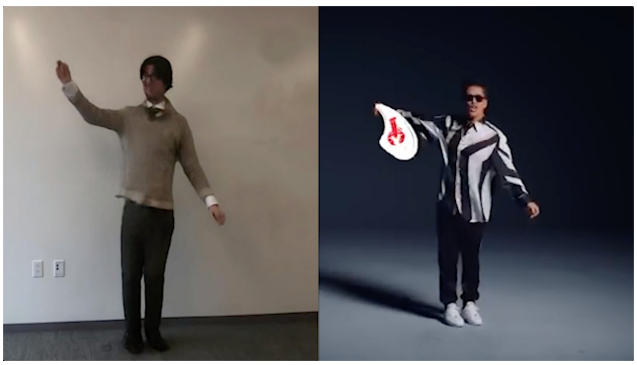

Figure 5. The Results of Transferring Bruno Mar's Dance to James' Dance. *Left*: Generated James' Dance. *Right*: Generated Bruno's Dance.

4.2. Transferring James's Dance to Teammates' Movements

In the second part of our project, we treat James as our source object and transfer his dance to our teammates. Figure 6 shows the possibilities of different qualities of generated video by the same source video. In fact, the quality of generated video mainly depends on the quality of target video. The original target video in Chan et al.'s study is formed around 20 minutes of real time footage at 120 frames per second [3]. Our teammates target video is formed around 1-2 minutes of real time footage at 120 frames per person. This is the reason why there are different qualities in the generated videos.

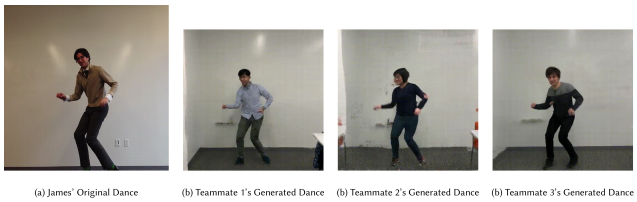

Figure 6. The Results of Transferring James' Dance to Our Teammates.

4.3. Discussion

The most challenging part in our project is the mismatched pose estimation. We improve the quality of generated video by normalizing the pose estimation in the source and target videos.

The quality of the target video is essential to generate ideal results. There are three important factors: **length of video**, **range of motion of subject**, and **frame rate**. The length of the shot and the range of motion determines the existence of the match between source pose estimation and target pose estimation. The frame rate reduces the blurring effect in each frame.

There are further improvements we can make in the future study. The pose estimation in our project can only detect body and limbs, which lead to blurring and missing hands in some generated frames. We would like to improve our pose estimation to detect hands shape to address the blurring and missing hands issue.

5. Conclusion

We use real-time Multi-Person Pose Estimation to generate the 2D pose for objects in selected frames of the source and target videos. Since the source object and target object are likely to be inconsistent with the camera in distance, height, and physique, we need to normalize pose estimation to match the source pose and the target pose. Next, we apply pix2pixHD [7] based on source pose estimation and GAN to generate a video in which target object dances in the source object's dancing style. In the first part of our experiment, we successfully transferred Bruno Mars' dancing style to James' in a relatively high quality video. In the second part of experiment, we transferred James' dancing style to all three teammates. It is shown in our presentation slides.

This project shows the possibility of imitating others' movements, and dance is just one interesting example. For each target object, we only need to train a single model well and then use this model to generate any videos we like.

This technology can be extended to animation, where characters act like professional dancers or martial artists. This concept could also be extended to non-human subjects such as animals and fictional creatures in animation, for example.

Methods that can deal with the 3D occlusions in this scenario can be considered as a possible future research. Our model achieves limited performance when the hand of the target appears in front of the chest of the target object.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 2

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 1, 2

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 1, 3

[4] S. Mo, M. Cho, and J. Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018. 1

[5] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[6] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[7] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[8] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1

[9] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

[11] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017. 1