Paper Reading Report-04

Han Zhang
u7235649

Abstract

*This is my reading report for the paper titled: "***Everybody dance now***", authored by Caroline Chan et al, and published in 2019 IEEE/CVF International Conference on Computer Vision (ICCV).*

All ENGN8501 submissions will be subject to ANU's TurnitIn plagiarism check, against both the original paper, internet resources, as well as all other students' submissions. So please make sure that you must write your own reports, and declare the following originality statement:

I, Han Zhang, hereby confirm that I am the sole author of this report and that I have compiled it in my own words.

1. Problem Statement

The problem to solve in this paper is to use a simple method to perform "do as I do" motion transfer, and gives a method to distinguish whether a video is synthesized. This can be used in making film and television special effects and games, and the fake-detector can be used to solve legal and ethical issues caused by fake videos.

Early methods focused on create new content with existing video footage. Some optical flow or classical computer graphics method of motion transfer achieved the goal in 3D. Some methods rely on a calibrated multi-camera system to build 3D models of target actors and manipulate their movements. Recent research uses deep learning, but relies on more detailed input representation. There are also methods to separate motion from appearance and synthesizing videos with novel motion, but do not focus on synthesizing detailed action videos. Modern methods can generate a single image of a new human pose and introduced new structures and losses, but these methods are not designed for motion transfer. There are also motion transfer methods that learns the mapping between videos, but they are more complicated and need more resources.

This method synthesizes new actions and only uses 2D videos. This method is based on recent research in two independent directions:robust pose estimation, and realistic image-to-image translation to achieve single image generation. Based on these, the author uses pose detection as an intermediate representation, and considers time coherence, so as to obtain a realistic video. They also gave a method to distinguish composite videos.

2. Summarise the paper's main contributions

The authors claim that they have designed a simple method to implement movement transfer, based on modern pose detection systems and image-to-image conversion models, using pose detection as an intermediate representation to generate time-correlated and realistic videos, and giving the discriminating video The method of authenticity.

Although their work is largely based on the work of predecessors, they innovatively and effectively integrate and improve the methods of two independent directions, and find a reasonable intermediate representation, which uses less cost to obtain more realistic results.

3. Method and Experiment

The method is divided into three stages. In the pose detection stage, the pre-trained pose detector (OpenPose) is used to estimate the 2D joint coordinates and create the pose stick figure of the original video.

The global pose normalization stage explains the difference in body shape and position of the source and target, transforming the key points of the source person's pose to make it consistent with the target person since the body proportions of people in different videos may be different. They used a linear mapping between the nearest and farthest ankles and calculated the scale and translation accordingly.

Finally they use Generative Adversarial Networks (GAN) to learn the mapping from pose stick figures to target person image based on the adversarial single frame generation process. The generator network *G* synthesizes images to deceive the multi-scale discriminator *D*, and *D* discriminates the grand truth and fake images generated by *G*, so that they are trained together and promote each other. They added a learned model of temporal coherence to enhance the temporal coherence between adjacent frames for video synthesizing. The objective is

$$\begin{aligned}\mathcal{L}_{smooth}(G,D) = \mathbb{E}_{x,y}[logD(x_t, x_{t+1}, y_t, y_{t+1})]+ \\ \mathbb{E}[log(1 - D(x_t, x_{t+1}, G(x_t), G(x_{t+1})))].\end{aligned} \tag{1}$$

To add more details and realism to the face, they added a special GAN G_f, input a small part of the image $G(x)_F$ around the face and the corresponding pose stick figure x_F, and generate a residual $r = G_f(x_F, G(x)_F)$ and added to G. The discriminator D_f discriminates real and fake faces, similar to pix2pix target. The objective is:

$$\mathcal{L}_{face}(G_f, D_f) = \mathbb{E}_{x_F, y_F}[logD_f(x_F, y_F)] + \mathbb{E}_{x_F}[log(1 - D_f(x_F, G(x)_F + r))]. \quad (2)$$

They employ training in stages, and the full image GAN and the face GAN are trained separately. Firstly is the full objective:

$$\min_G((\max_{D_i} \sum_{k_i} \mathcal{L}_{smooth}(G, D_k)) + \lambda_{FM} \sum_{k_i} \mathcal{L}_{FM}(G, D_k) + \lambda_P(\mathcal{L}_P(G(x_{t-1}, y_{t-1}) + \mathcal{L}_P(G(x_t), y_t))) \quad (3)$$

where $\mathcal{L}_{GAN}(G, D)$ is the single image adversarial loss in the pix2pix, $\mathcal{L}_{FM}(G, D)$ is the discriminator feature-matching loss in pix2pixHD, and $\mathcal{L}_P(G(x), y)$ is the perceptual reconstruction loss which compares pretrained VGGNet features. Then is the face GAN:

$$\min_{G_f}((\max_{D_f} \mathcal{L}_{face}(G_f, D_f)) + \lambda_P \mathcal{L}_P(r + G(x)_F, y_f)) \quad (4)$$

They compared this method with other methods in video and single frame. For video, they used Mechanical Turk to evaluate the result. Compared with the PoseWarp method, the video generated by this method is more realistic no matter the face GAN module is used or not. For single frame, SIMM and LPIPS show that compared with frame-by-frame (FBF) and temporal smoothing (FBF+TS), this method (FBF+TS+Face GAN) also performed better.

They also trained a fake-detector to distinguish the authenticity of the video. They collected a data set and trained the fake-detector by generating multiple fake videos through a synthetic model. In general, the fake-detector can distinguish the authenticity of the video very well.

4. Critical Analysis

4.1. Are the paper's contributions significant?

The contributions are significant. Some of the previous methods require multi-camera systems, some require 3D modeling, some require more detailed input, and some are more complex and require more resources. This method integrates the work of predecessors well, and only needs easy-to-obtain 2D video to generate realistic video. They also proposed methods to distinguish the authenticity of videos to help solve possible legal and ethical issues.

4.2. Are the authors' main claims valid?

The main claims are valid. They integrated the work of predecessors and proposed a new method, gave mathematical derivation and network model and compared the differences with other methods through experiments, the result show that their method does perform better.

4.3. Limitation and weaknesses

The results shows that sometimes OpenPose makes mistakes in detection, and sometimes texture artifacts appear. Combining the target video with different clothes or lighting conditions, improving the pose detection module, and reducing the artifacts of high-frequency textures may help. Their pose normalization stage did not consider limb lengths or camera positions, making the generated video action different with source. This can be improved by optimizing the normalization method. Their model sometimes is difficult to infer poses that are not absent in the training data set. This can be improved by improving the data set.

4.4. Extension and future work

The authors can try other pose detection methods for better results, and optimize pose normalization methods to reduce the difference in actions. They can also expand the data set with more combinations of clothing and lighting conditions to reduce texture artifacts, and try more types of actions to enhance the adaptability to different actions.

In the future, this method can be used in making entertainment and even professional video special effects. The fake video detector can help solving legal and ethical issues.

4.5. Is the paper stimulating or inspiring ?

This paper is exciting. The previous methods are usually limited or not easy to implement. This method is easier to implement and only uses easily accessible videos. They also gave a fake video detector.

4.6. Conclusion and personal reflection

In conclusion, this paper proposes a simple method to solve the movement transfer problem, integrating the results of the predecessors in two separate directions, using 2D video as the input source to generate realistic videos. They also proposed methods for distinguishing the fake videos.

I may try to combine optical flow method to perform rough modeling to assist the pose detection algorithm to improve the robustness.

This paper tells me that finding a suitable intermediate expression may reduce the level of detail of the input data, and the rational use of GAN can enhance the performance of the algorithm.

References

[1] Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei Efros. *Everybody Dance Now*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV).