IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

Everybody Dance Now

C Mounica Reddy[1], Pooja S R[2], Deepthi N V[3], Chandana P[4], Mr. Suhas S[5]

8th Semester, Department of Computer Science, The National Institute of Engineering, Mysore, India[1,2,3,4]

Assistant Professor, Department of Computer Science, The National Institute of Engineering, Mysore, India[5]

Abstract: This paper presents a basic strategy for "do as I do" : taken a recording of person performing dance which is given as the input to the model as source, the model creates recording of a beginner dancing with the movements which is same as the source recording .Video to video interpretation is utilized as a middle portrayal. From the source subject, we extricate the postures to pass on the movement. Then, the extracted postures are applied to train the appearance mapping to generate target subject. We anticipate two sequential frames that are likely to contain similar object or objects to obtain video results and present a different pipeline for sensible face synthesis. In spite of the fact that our strategy is very basic, it delivers shockingly convincing outcomes. This persuades us to likewise give a forensics device to identify the fake, which can recognize recordings blended by our framework from genuine information.

Keywords: Computer vision, pose detection, pose extraction, deep learning, cGAN

I. INTRODUCTION

Ever imagined dancing like Michael Jackson?" Possibly in dreams! ", might be the answer, but its indeed possible now. Here, we are going to implement a basic however successful methodology for "Do as I Do" video retargeting – naturally passing on the movement from source to target recording. Given two recordings – one of the people who can dance and the other of the individual who we teach to dance – we pass on movement between these subjects by studying a basic video-to-video interpretation. With our system, we make different recordings, empowering amateur beginners to turn and whirl like ballet dancers, perform kendo, or move as lively as pop stars.

To pass on movement between two recording subjects in a outline by-outline way, we should become familiar with a mapping between pictures of the two people. Our main objective is to look for the similar image interpretation by analysing source and target sets. Be that as it may, we don't have relating sets of pictures of both the subjects playing out the same movements in order to manage to learn this interpretation. Regardless of whether the two subjects play out a similar performance, it is still improbable to have a definite frame to frame sequence as body shapes, movement varies from person to person.

We see that keypoint-based posture stores movement impressions over time while trying to restore the identity of source as much as we can and thus acts as middle portrayal between any two subjects. In this manner, we use posture stick figures got from off-the-rack human posture indicators, for example, OpenPose, as a middle portrayal for image to image transfer. We at that point get familiar with a picture to-picture interpretation model between posture stick figures and pictures of our target individual. To pass on movement to target subject from the source subject, the posture stick figures are inserted to our trained model in order for target subject to acquire pictures same as the source subject.

II. RELATED WORK

In the course of the most recent two decades there has been broad work devoted to movement transfer. In the past, strategies were centred on making new stuff by controlling existing video film. For instance, Video Rewrite makes recordings of a subject expressing a group of words which they didn't initially articulate. This can be done by discovering the mouth position which matches with the required speech.

Computer Graphics can also be used to deal with movement transfer to play out this in 3D. Inverse kinematic solvers were used to solve the retargeting problem in animated characters.

A few methodologies depend on aligned multi-camera arrangements to 'filter' a target on-screen character and their movements are controlled in another video by fitting a 3D design of the target subject. To acquire 3D data, Cheung et al. propose an expound multi-view framework to adjust a customized kinematic model, to estimate 3D joint and provide pictures of a person executing new movements. Present day posture detection frameworks including OpenPose and DensePose enables quick posture extraction in an assortment of situations. Simultaneously, the ongoing development of

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

frame to frame interpretation models, pix2pix, CoGAN, UNIT , CycleGAN, DiscoGAN, Fell Refinement Networks, and pix2pixHD, have helped in producing high quality images.

III. IMPLEMENTATION

A. OpenPose

Openpose is a bottom-up approach in which we first detect all the key points of the person present in the frame. Body parts such as neck, left shoulder, right shoulder, hip and soo on are identified as keypoints. Any two keypoints can be combined to form a pair. For example, a connection between neck to the left shoulder and neck to the right shoulder.

In openpose architecture the input we are getting is the RGB image of 224 X 244 x 3. First we need to apply the feature extraction layer which is a VGG 19 used to extract the features from the image. Next we have to split the next part of the network into two parts and these two branches will predict two different things. First branch will predict a set of 18 confidence maps each one will tell which particular part we are trying to estimate. Second branch will give 38 different output which represent the degree of Association (to obtain the pairs).

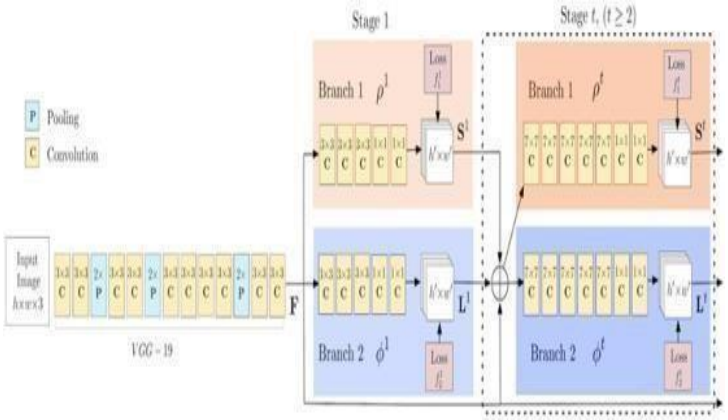

Fig. 1 Architecture of OpenPose model: Colour image is given as input whose size is h x w. We get the output as an array of matrices consisting of keypoints.

B. TensorFlow

TensorFlow is a Python-favourable open source library for numerical calculation that makes AI quicker and simpler. Architecture of TensorFlow works in three parts:

- Pre-processing the data
- Build the model
- Train and estimate the model

TensorFlow takes input as a multi-dimensional array, also called tensors. You will be able to construct a kind of flowchart of operations (also called as Graph) that you want to perform thereon input. The input goes in at one end, and then it flows through this technique of multiple operations and comes at the opposite end as output.

C. cGAN

The conditional generative adversarial network(cGAN) is type of GAN which creates a general framework for different applications. Basically GAN is used to identify whether the image obtained from the output is fake or real. The loss is minimized by training generative model by GAN. We consider two main parts of cGAN namely Discriminator and Generator. Discriminator is used as real or fake classifier. Generator is used to minimize the loss Where as cGAN is a condition on an input image to generate corresponding output image. In cGAN, the generator learns to generate fake samples with the specific condition or characteristics rather than a generic sample.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

D. Epoch

Neural network is created of neurons which are connected to each other. Weight is represented by each connection present in neural network. Weight describes the importance of this relationship of the neuron we get when multiplied with the input value.
The first phase is forward propagation. Whenever the input file is passed, the data is transformed to the form which is received from the previous layer neurons. Present layer neurons forward the results to successive layer. When the information is passed through all layers, and after all the calculations made by neuron layers ,the last layer is going to have results.

To estimate and test the loss, we have used a loss function. This measures how relevant was our predicted results when compared to standard result. After calculating the loss, we propagate the data backwards. Hence, its name: backpropagation. We adjust the weights of interlinks of neurons till we get the predicted results after training the model. This minimize the loss in the network.

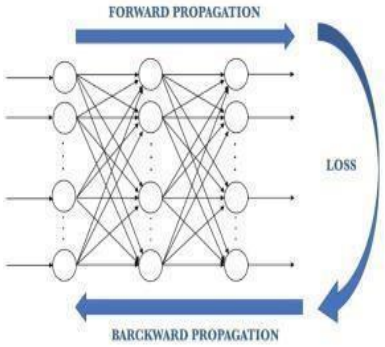

Fig. 2 This diagram shows the neurons and how the data is passed through the neural network.

IV. METHODOLOGY

Our model aims at generating a video of target person having the same body movements as the source person, provided with source person video as input. Considering desired pose as x,a pose from the f(x) domain should be converted into pose g(x) which has desired game character.The problem is divided into two parts inorder to solve this problem.

- x is obtained by estimating pose
- x is then transformed to desired game character pose by using CGAN model.

A. Extracting the Poses

Feasible locations of each keypoint in the image is predicted. Then we extract pose stick figure by connecting all possible keypoints as shown in the Fig. 3.

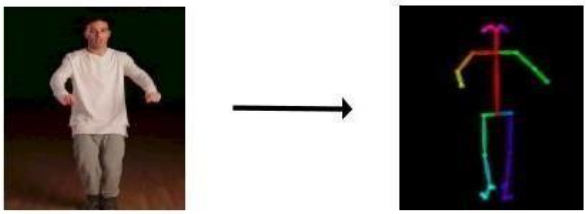

Fig. 3 Extracting the poses: From the video we are extracting the stick figures by predicting the location of each keypoint and the image.

B. Dataset for Training

We have chosen a fortnite dance video as our dataset consisting of 26000 images. Before obtaining the images, we have cropped the video so that unwanted portions are removed. Target video must be able to capture adequate rate of motion such that we obtain frames having least shakiness with high quality .For training, stick figures of target person is created using a model which is called as pose detector.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

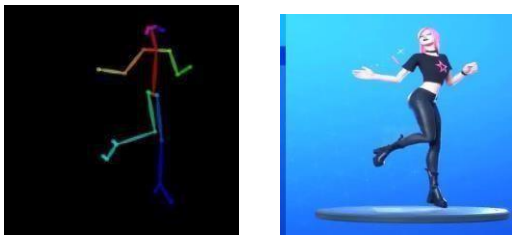

Fig. 4 Datasets: Images are extracted from a YouTube video and their corresponding stick figures are obtained using open pose. These are stored in different folders and are used for training purposes.

C. Training process

In this phase, We have trained the generator model G of cGAN such that the generated pose stick figures is made to pass through G. Training is further proceeded by Discriminator D by which the mapping G learns discriminator D which classifies the real (x, y) pairs and fake (G(x),y) pairs. we have used VGGNet pre-trained model which helps to optimize result by matching target images(y). With the generated images G(x). Target person's G(x) is generated out of G model which was trained earlier using pix2pixHD framework.

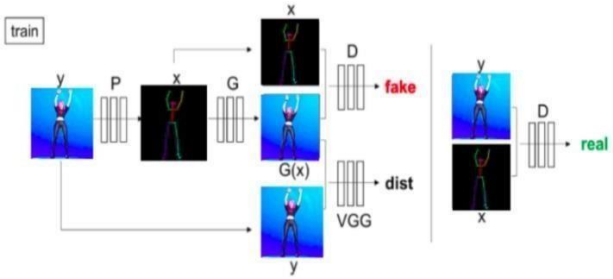

Fig. 5 Training: Image shows training of generated model G and Discriminator D to obtain G(x) of target person.

D. Testing

Testing is done at different level of abstraction. The model was introduced with new set of images and their corresponding stick figures and was tested by setting constant weights. Iterative testing is done by constantly running the model until the number of errors reduce in every iteration.

E. Transfer

In this phase, pose joints of the source subject are obtained by P called as pose detector. We use normalization process(Norm) to generate the joints for target subject. Then, we have applied the training mapping G.

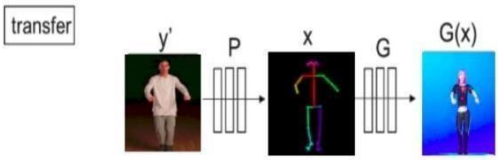

Fig. 6 Transfer phase: Normalized process is used to produce joints of target person.

V. RESULT

The real time motions of target subject for around 5 minutes at 30 frames per second was recorded. YouTube video or any other video with high resolution can be chosen as source video. The star was not printed on the t-shirt. Overlap of hand was not registered. Due to incomplete pose estimation, the generated image has a missing hand. These were some of the negative results produced. But jerk is the common problem which reduces the quality of videos. Shakiness occur whenever the speed of the input motion varies with that of the one which is being trained. We can overcome these by giving good training to the model.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

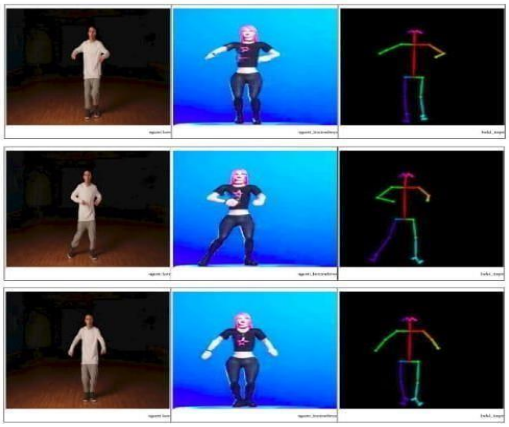

Fig 7. Results of source image, synthesized image and pose stick figures.

VI. CONCLUSION

In this paper, we presented the model for creating nearly lengthy videos of a target subject, who has the same body movements as that of the input recording of source subject dancing. In various scenarios, Openpose is the most used pose detection system for extracting the poses at high speed. The open source data of the single dancer is of the high resolution ,which is being used for training the movement relocations and to generate recordings. The face GAN setup used in the model improves both computable and subjective results of the model. Although, jittering degrades the standards of the result. Still, this method is capable of generating videos given a various input.

ACKNOWLEDGMENT

We would like to thank the Principal, **Dr. G Rohini Nagapadma** , the Head of the Department of Computer Science and Engineering, **Dr. V K Annapurna** and the faculty of The National Institute of Engineering, Mysuru, for providing immense support for the research and the resources to develop the prototype.

REFERENCES

[1] Jehee Lee and Sung Yong Shin.In advance of 26th conference on Computer graphics and interactive techniques, pages 39–48. ACM Press/Addison-Wesley Publishing Co., 1999.

[2] German KM Cheung,Takeo Kanade, Jessica Hodgins, and Simon Baker . Markerless human motion transfer. Visualization and Transmission, 2004 in 3D Data Processing, . 3DPVT 2004. Proceedings. 2nd International Symposium on, pages 373–378. IEEE, 2004.Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In CVPR, 2018.

[3] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4D Video Textures for Interactive Character Appearance. Computer Graphics Forum (Proceedings of EUROGRAPHICS), 33(2):371–380, 2014.

[4] Zhe Cao,Yaser Sheikh , Shih-En Wei, and Tomas Simon. Realtime multi-person 2D pose estimation uses part affinity fields. In CVPR, 2017.

[5] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In advance of the 24th conference on interactive techniques and Computer graphics , pages 353–360. ACM Press/Addison-Wesley Publishing Co., 1997.

[6] Wissam J Baddar, Geonmo Gu, Sangmin Lee, and Yong Man Ro. Dynamics transfer gan: Generating video by transferring arbitrary temporal dynamics from a source video to a single target image. *arXiv preprint arXiv:1712.03534*, 2017.

[7] Patrick Esser, Ekaterina Sutter, and Bjorn Ommer. A varia-¨ tional u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[8] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.