*International Journal of Advanced Culture Technology Vol.7 No.2 195-200 (2019)*
*DOI 10.17703/IJACT.2019.7.2.195*

**IJACT 19-6-24**

# Interaction art using Video Synthesis Technology

Kim Sung-Soo*, Eom Hyun-Young*, Lim Chan*

** 1,SoongSil University Global School of Media, Seoul ,Korea*
*sdasd@gmail.com*
**2SoongSil University Global School of Media, Seoul ,Korea*
*ehy0321@naver.com*
**3(Corresponding author) SoongSil University Global School of Media, Seoul ,Korea*
*chanlim@ssu.ac.kr*

***Abstract***

*Media art, which is a combination of media technology and art, is making a lot of progress in combination with AI, IoT and VR. This paper aims to meet people's needs by creating a video that simulates the dance moves of an object that users admire by using media art that features interactive interactions between users and works. The project proposed a universal image synthesis system that minimizes equipment constraints by utilizing a deep running-based Skeleton estimation system and one of the deep-running neural network structures, rather than a Kinect-based Skeleton image. The results of the experiment showed that the images implemented through the deep learning system were successful in generating the same results as the user did when they actually danced through inference and synthesis of motion that they did not actually behave.*

***Keywords:*** *Interactive Art, Deep Learning, GAN, VVVV*

## 1. Introduction

The invention of photography, telephone, movie, etc. had a direct or indirect effect on various fields. Media art, a combination of art and invented media technology, is being developed faster and more variously in combination with artificial intelligence, IoT and VR. Media art has interactive characteristics that users participate in their own works. In 2018, Cho Sung-hee and Kim Eun-jung proposed user-participated interactive art using various sensors [1], Lim Yi-joon, and Im Chan proposed participatory advertising that allows the user to manipulate the globe according to the movement of the user's hands [2]. In 2013, Kim Young-eun and others surveyed the user experience through interactive interactive art and found that interactive art using motion sensors could provide a deeper three-dimensional viewing experience than interactive art using touch screens [3]. These interactive arts can satisfy a variety of needs, beyond simply being part of media art.

In this paper, it is proposed that people who are tired of everyday life use image synthesis technology to create composite images that mimic the movements of the objects they aspire to. The composition of this paper is as follows. First, describe the studies related to the proposed system, explain the structure of the proposed system and the results of the experiment, and finally describe the conclusions of this paper based on the results of the experiment.

Manuscript Received: April 22, 2019 / Revised: May 08, 2019 / Accepted: May 20, 2019
Corresponding Author: chanlim@ssu.ac.kr
Dept. Computer Science, Soongsil Univ, Chan Lee, Dongjak-gu, Seoul, Korea

## 2. Related Works

2.1. Unsupervised Methods

The proposed system is a system that learns the position of the target images that you want to synthesize based on the images of the user you are input to create images where the target images are dancing. Because the system composes images rather than single images, it is necessary to minimize human intervention in the process of synthesizing images. Therefore, the proposed synthesis technique of this system was used. The composite images were configured to move to the vvvv environment, one of the interactive art tools, to give users several special effects.

2.2. Video synthesis system

Image synthesis, the main technology of this paper, was built on the basis of General Adaptive Networks (GAN) architecture during deep learning. The GAN network proposed by Ian Goodfellow and others in 2014 is a representative notepad learning network, and it improves performance by learning imitation data produced by the generation model until it is classified to an accuracy with which the classification model converges 50% [4].

$$g^* = \arg\min_g \max_d v(g,d). \quad \text{(equation 1)}$$

$$v(\theta^{(g)}\theta^{(d)}) = E_{x\sim p_{data}} \log d(x) + E_{x\sim p_{noise}} \log(1-d(x)). \quad \text{(equation 2)}$$

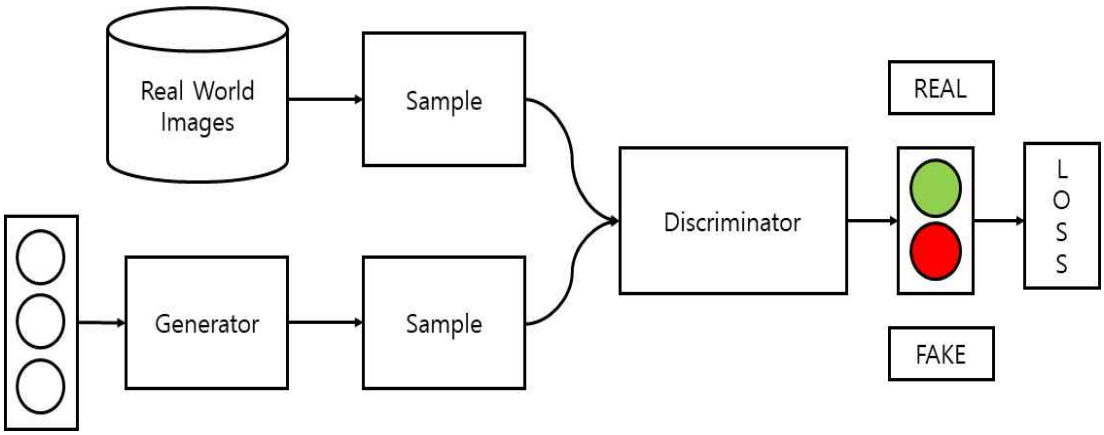



**Figure 1. Structure of GAN**

The proposed project consists of learning and synthesizing images on a frame basis of frame. In 2018, Ting-Chun Wang et al. developed a system that can synthesize up to 30 second-long 2K resolution video using GAN structure, which was a disadvantage of the existing imaging synthesis system [5]. In 2018, Caroline Chan et al. developed a system that utilizes the original deep learning to extract the shape of Skeleton according to each movement of the target in the original image and synthesize it according to the Skeleton estimate extracted from the target image [6].

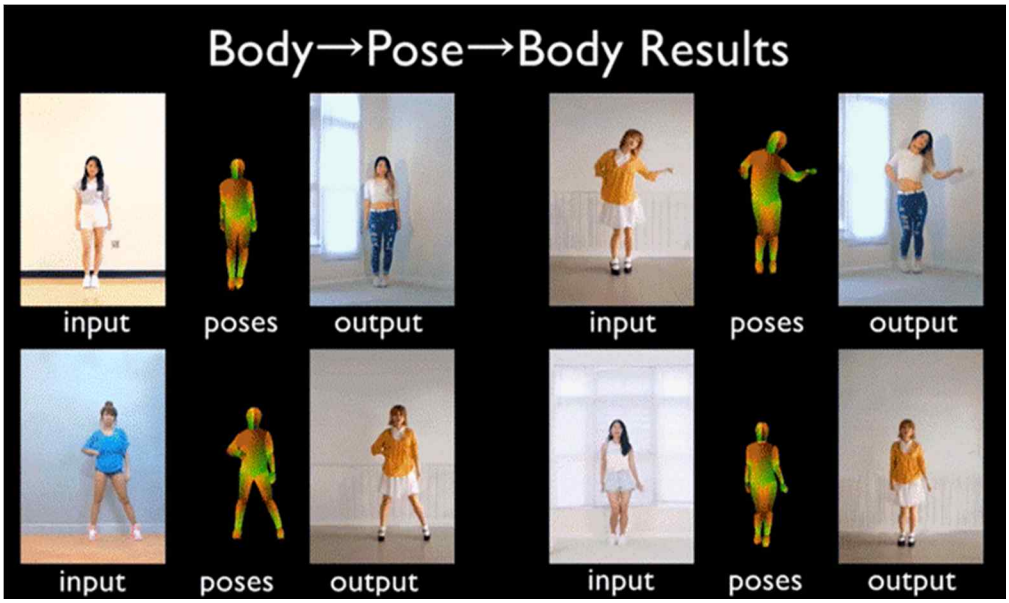



**Figure 2. vid2vid sample**

2.3. Skeleton Estimation

It is necessary to extract the skeleton information value of the original and target images in order to act on the target of the target in the original image. In 2015, Qifei Wang et al. proposed a Kinect-based pose estimation system and succeeded in achieving high accuracy for 12 postures [7]. However, Kinect-based postural estimation algorithm has the disadvantage of not being able to use non-image data taken from Kinect. In 2016, Zhe Cao et al, using a deep learning inference system based on VGG [8], which is intuitive and highly accurate, estimated the human skeleton in the image in real time using a regular RGB camera to confirm that the mAP (Equation 3) of 59.8 was shown [9].

$$\text{mAP} = mean\left(\frac{TP}{TP+FP}\right) \quad \text{(equation 3)}$$

## 3. Design and Implementation of Imaging System

Using deep learning technology, this paper proposes a system that synthesizes images and gives users special effects on synthesized images so that they can imitate the dance of the characters of the images they target when they enter the images. The process of the system proposed in this paper is shown in [Figure]. First, when the user inserts the image, the server creates a composite image of the target's pose through the Vid2Vid neural network. The purpose of the images created is to be linked back to VVVV so that users can put their desired effects in

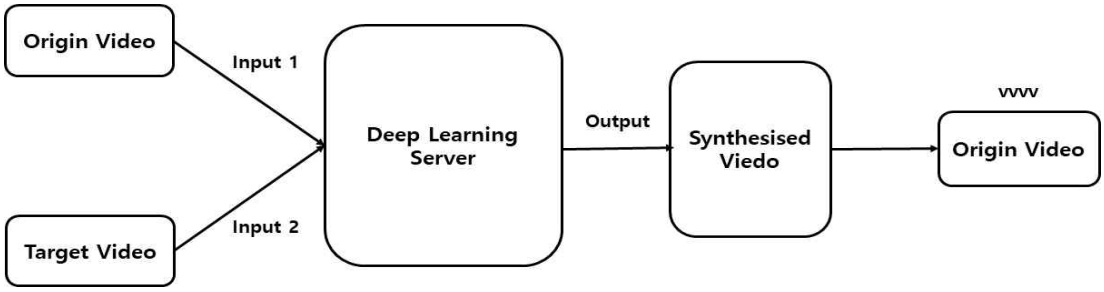



**Figure 3. System Process diagram**

3.1. Skeleton Estimation System

The Skeleton estimation system of this paper is built on the basis of python language and the deep learning framework used for system development is pythorch. Realtime Multi-Person Pose Estimation was used as a pose estimation algorithm. A set of pre-learning data for the estimation of Skeleton was based on a set of Microsoft's public data, the Coco dataset. Skeleton estimation system is developed by storing input images divided by frame and then estimating and storing the poses of pictures based on the studied neural network weights.

**Figure 4. Realtime Multi-Person 2D Human Pose Estimation**

3.2. Imaging Synthesis System

The imaging synthesis system used in this paper used nvidia's Fix2pixHD network to synthesize high-grade photography using conditional GAN and consisted of python language, pythorch base, just like the Skeleton estimation system [10]. Both the input and target images are stored in frame-by-frame image files and the position most similar to that of the target image is matched to the position of the original image to proceed with synthesis.

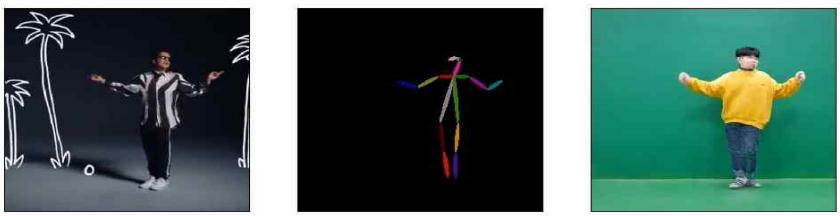
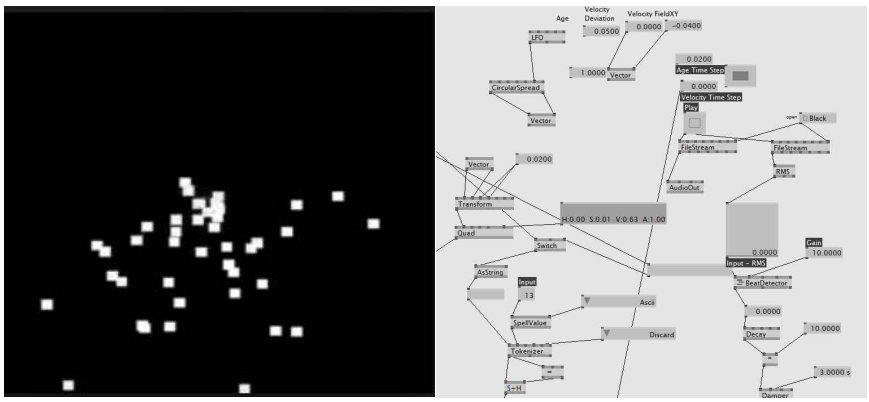


**Figure 5. Synthesised Process sample**

3.3. Special Effect Granting

The proposed system has special effects in the VVVV environment for images that have been synthesized. Before the special effects were given, special effects were given after the background was removed using an image editing tool, such as a documentary, to improve the quality. The special effects of this system are partitioned images using the VVVV embedded node, Particles, and RMS is used to track the amplitudes of music in images that the user is trying to track and automatically adjust the size and scatter of the particles according to the amplitude. In addition, the color information of the particle was randomly changed to super units to avoid monotony.




**(a) Sample of Particle in VVVV** **(b) nodes for insert particle on video**

**Figure 6. Bit Tracking to Particle Nodes**

The final result of the proposed system is shown in [Figure 7].

**Figure 7. Result of Proposed System**

## 4. Conclusion

The purpose of this study is to create a video that simulates the dance moves of the object that users admire by using deep learning synthesis technology, giving the user the feeling of dancing that he or she admires. The experiment resulted in the creation of a composite image that gives the user the same feeling that he or she actually did without actually taking the desired dance action, which was particularly effective in conveying the same emotion as the user dancing in a different space.
However, the synthesis technology of the system proposed in this paper has some limitations. First, if none of the user-entered motion in the image was similar to the motion in the image being tracked, serious noise was generated. Second, because quality of resolution of composite images has decreased, there is a problem of increasing sense of heterogeneity when viewed from large screen. Finally, the imaging synthesis technology of the proposed system required a lot of computations, which took a lot of time to compose the images for one minute and a half. Accordingly, this paper foresees the development of neural network structures that enhance the system of deducing the motion of original images and the behavior of trace images, algorithms that produce less resolution loss during image synthesis, and the synthesis time reduction algorithm through optimization of code used.

## References

[1] Cho Sung-hee and Kim Eun-jung. The Fourth Industrial Revolution : A Classification of Reality-Virtual Media Connection System and Case Studies on Dance Performing Arts (2018). the Fourth industrial revolution Written by the Korea Content Association, 18(9), 544-554.
[2] Lim Yi Jun, Lim Chan, Interactive advertising with VVVV and Kinect, Art and Humanities Convergence Multi-media paper, Vol.8, No10, p.481~489
[3] Young Eun Kim, Mi Gyung Lee, Sang Hun Nam, Jin Wan Park, User Interface of Interactive Media Art in a Stereoscopic Environment, Lecture Notes in Computer Science, vol. 8018, pp. 219-227, (2013).
[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative Adversarial Networks, Department of Informatics and Operational Research University of Toronto, (2014)
[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro. ,Video-to-Video Synthesis, Computer Vision and Pattern Recognition (2018).
[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros, Everybody Dance Now. Computer Vision and Pattern Recognition (2014).
[7] Qifei Wang, Gregorij Kurillo, Ferda Ofli, Ruzena Bajcsy ,Evaluation of Pose Tracking Accuracy in the First 8and Second Generations of Microsoft Kinect, IEEE International Conference on Healthcare Informatics (2015).
[8] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition(2014).
[9] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Computer Vision and Pattern Recognition (2017).
[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro, High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, Computer Vision and Pattern Recognition (2018).