졸업논문 소개

Blendshape-based Retargeting for 3D Facial Expression

이솔 / 광운대학교 Intelligent Computing Lab.

Implementing avatar facial movements has been attempted in various ways in the movie, game, and virtual environment industries. In recent film productions, facial movements are primarily captured by marking an actor's face and tracking those markers to move the avatar's face. However, this method comes with the inconvenience of attaching or marking the actor's face with markers. Moreover, there is a growing need for technological research to enhance the realism of digital humans in the virtual environment industry. Consequently, recent research on avatar facial animation is being actively conducted from various perspectives, with a particular focus on studies of facial expression transfer techniques based on deep learning. Learning-based 3D facial expression creation methods have the advantage of reducing time and production costs and have enabled the creation of sophisticated animations thanks to recent advances in deep learning technologies.

Traditional methods for animating faces include the rigging method and the blendshape technique. The rigging method is where control points are inserted into a 3D model, and weights are assigned to each vertex influenced by these control points to manipulate facial expressions. This approach is beneficial for naturally representing complex facial movements. Conversely, the blendshape technique creates new expressions by combining predefined facial expressions. It is effective for simple expression changes and allows for the quick and easy implementation of various expressions. These two methods are often used together to represent the natural expressions of 3D characters. Additionally, animating facial expressions using the deformation transfer method is also possible.

In recent times, alongside traditional methods, several studies have been conducted based on artificial intelligence. For facial feature extraction encoders, autoencoders (AE), variational autoencoders (VAE) combined with the concept of graph convolution networks (GCNs), such as graph convolutional AE

(Graph AE) and graph convolutional VAE (Graph VAE), are primarily used. Moreover, research has been conducted using Diffusionnet for feature extraction. Studies have also been conducted on using 2D AE to generate avatar facial images from human facial footage for 3D retargeting.

However, current 3D facial animation methodologies have three main issues:

1. Not considering the three-dimensional information of the human face can reduce the accuracy and realism of facial expression extraction. Especially when using only 2D information, there can be limitations in accurately representing the depth and volume of specific facial regions depending on the camera angle.
2. Each avatar subject to animation requires independent learning. This issue significantly invests time and effort in manually creating datasets that pair human input expressions with corresponding avatar expressions.
3. If the range of movement between the input and output faces is not considered, the avatar's expression may be too exaggerated or too subtle compared to the input, resulting in low-quality animation.

To address these issues, this paper proposes a method for transferring facial expressions to 3D avatars. This is achieved through estimating blendshape coefficients of facial expression features. The network comprises a facial expression feature extraction encoder and a blendshape coefficient estimation decoder. The facial expression feature extraction encoder extracts features of facial expressions changing from a neutral face in three dimensions. Concurrently, the blendshape coefficient estimation network estimates the blendshape coefficients that make up the expression from the previously extracted facial features. Introducing a mask for each blendshape in the loss function during training significantly enhances the learning effectiveness.

This paper proposes a deep learning network that transfers human facial expressions to targets such as avatars. We utilize a graph convolutional autoencoder (GC AE) to extract features from human facial expressions and then convert these features into blendshape coefficients using a fully-connected (FC) layer. By estimating blendshape coefficients from a person's 3D facial data, we can animate avatars to replicate the same facial expressions as the person. In particular, in this paper, we construct learning data features by combining deformation gradients and displacements (or deltas). We employ graph convolutional networks (GCN) to extract features from a person's 3D facial mesh effectively. We trained the network in a two-stage learning process: firstly, to extract features of facial expressions and secondly, to convert expression features into blendshape coefficients. Introducing masks for each blendshapes into the loss function during the training resulted in significant learning effects. The error between the input and output of the trained AE is less than e-9, demonstrating that the latent space has been sufficiently learned to restore the output data to the same level as the input.

Additionally, the error between the input and the reconstruction using the blend shape model, which

was designed to have the same appearance and topology as the input, and the blendshape coefficients estimated by the proposed network, were less than e-3. Visually, it was confirmed that the same expressions were estimated. Furthermore, we enhanced versatility through additional experiments using the FLAME (Faces Learned with an Articulated Model and Expressions) facial model, which is commonly employed in 2D image-based 3D face generation networks and voice-based 3D facial animation generation networks. This paper demonstrates the feasibility of accurately transferring 3D human facial expressions to other avatars or virtual humans.

The network proposed in this paper was programmed using Python and computations were carried out on a GPU using CUDA in a Linux environment. An NVIDIA RTX 4090 was used as the GPU, and an Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz was used as the CPU. It took 5 hours to train for 100 epochs. The structure of the facial expression retargeting network consists of a facial expression encoder and an expression to blendshape coefficient decoder. The network is composed of graph convolutional layers and fully connected layers. The facial expression encoder is experimentally set with 7 layers of graph convolution layers and 2 layers of fully connected layers. The output dimensions of the graph convolution layers are each the number of vertices multiplied by 256, 128, 128, 64, 32, 16, and 1, respectively. The output of the graph convolution layers thus becomes the number of vertices, which is then flattened and fed into the fully connected layers. The output dimensions of the fully connected layers are respectively twice the dimension of the latent space and the number of dimensions in the latent space. The decoder part of the facial expression encoder has a symmetric structure to the encoder. The expression to blendshape coeff. decoder is experimentally set with 4 layers of fully connected layers. The output dimensions of the fully connected layers are set as 300, 300, 300, and 51, respectively, aligning the output dimensions with the number of blendshapes. For the layers with an output dimension of 300, ReLU is used as the activation function, and for the final layer, tanh is employed.

When constructing the retargeting loss using only blendshape coefficients, an error of approximately 1.18×10^{-2} was observed. Training using only the vertex loss reconstructed from blendshape coefficients showed an error of about 7.13×10^{-5}. In contrast, using the proposed blendshape mask to construct the loss resulted in a significantly lower error of approximately 1.39×10^{-9} and provided visually the most stable results. The advantage of converting extracted expression features into blendshape coefficients is that it eliminates the need for individual training for each avatar. Additionally, when compared to Apple and Google, which estimate blendshape coefficients in a similar manner, the results of our paper are comparable to the superior results of Apple. However, there is a limitation: since avatars' blendshapes can be modeled differently by modelers in terms of the extent of mouth opening or smiling, avatars modeled with excessive or minimal movements might show inappropriate movements. Therefore, our next research goal is retargeting that takes into account the modeling of avatars.

이 솔

- 2022년 2월 : 광운대학교 전자재료공학과 학사
- 2024년 2월 : 광운대학교 전자재료공학과 석사