creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Master's Thesis
석사 학위논문

3D Face Reconstruction for VR HMDs in Eye Region

Seoyoun Ji(지 서 연 池 敍 娟)

Department of
Information and Communication Engineering

DGIST

2020

Master's Thesis
석사 학위논문

3D Face Reconstruction for VR HMDs in Eye Region

Seoyoun Ji(지 서 연 池 敍 娟)

Department of
Information and Communication Engineering

DGIST

2020

3D Face Reconstruction for VR HMDs in Eye Region

Advisor: Professor Hoonsung Chwa
Co-advisor: Professor Sunghyun Cho

by

Seoyoun Ji
Information and Communication Engineering
DGIST

A thesis submitted to the faculty of DGIST in partial fulfillment of the requirements for the degree of Master of Science in the Department of Information and Communication Engineering. The study was conducted in accordance with Code of Research Ethics[1]

12. 22. 2019

Approved by

Professor Hoonsung Chwa (signature)
(Advisor)

Professor Sunghyun Cho (signature)
(Co-Advisor)

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of DGIST, hereby declare that I have not committed any acts that may damage the credibility of my research. These include, but are not limited to: falsification, thesis written by someone else, distortion of research findings or plagiarism. I affirm that my thesis contains honest conclusions based on my own careful research under the guidance of my thesis advisor.

3D Face Reconstruction for VR HMDs in Eye Region

Seoyoun Ji

Accepted in partial fulfillment of the requirements for the degree of Master of Science.

12. 22. 2019

Head of Committee	Prof. Hoonsung chwa (signature)
Committee Member	Prof. Sunghyun Cho (signature)
Committee Member	Prof. Sanghyun park (signature)

MS/IC
201900000

지 서 연. Seoyoun Ji. 3D Face Reconstruction for VR HMDs in Eye Region. Department of Information and Communication Engineering. 2020. 70p. Advisors Prof. Hoonsung Chwa, Co-Advisors Prof. Sunghyun Cho

ABSTRACT

The study of 3D reconstruction using video or photos has been a big issue since the past. As machine learning advances facial wrinkles can be reconstructed in detail and facial expressions can be reconstructed in real-time, but there are no paper about the contents that reconstruct the eyes part intensively. In this paper, I introduce a new method for reconstructing the eye area and a new method for real-time 3d reconstruction using only the head-mounted displays(HMDs) without complicated hardware device environment.

For the reconstruction, neural networks were constructed using machine learning and lightest neural networks were used to operate in real-time. Using HMD, collect dataset for learning and make ground truth reconstruction object manually. Then I make our own dataset for eye region reconstruction. This neural network model enables 3d reconstruction with only one image and allows create expressions such as surprised and frowning to express more diverse expressions while the previous 3d reconstruction models only can realize the expression of closing or opening eyes.

Keywords: 3d reconstruction, real-time eye performance capture, virtual reality

List of Contents

List of Figures

List of Tables

I. INTRODUCTION

For many years, dramatic advancement in computer graphics and display technologies take virtual reality (VR) and augmented reality(AR) to a higher level and allow us to project our own digital avatars into captivating virtual worlds. Also, as people's interest in VR and AR increases there are software on the market using this VR and AR technology, such as Samsung's 'make your own avatar – AR emoji'. However, an immersive and faithful digital presence is unthinkable without the ability to personalized digital avatars that can express precise facial expressions.

In order to create such a realistic and accurate avatar, the latest technologies required enormous hardware display environment. However, this environment setting is difficult to commercialized because of its operation to use for the general public. Therefore, I propose a method to make eye reconstruction more detailed than the conventional method using only head-mounted display(HMD), a simple device. But there are drawbacks to using HMD alone : occlusions. Occlusions introduced by HMD make facial tracking techniques challenges, and even state-of-the-art techniques fail to capture subtle details of the user's facial expressions. So, I introduce a system for HMD users to control a 3d avatar's eye region while producing plausible emotional expressions regardless of occlusions.

State-of-the-art facial reconstruction methods commonly use landmarks detection[1]. However, approaches that directly using tracked landmarks to recover the eye region very vulnerable to occlusions especially while wearing the HMD. In another approach, artists manually draw contours for all frames, and then solve a complex 3D model to fit the data. This is a very computationally intensive process so this method is rarely used in general commercial environment because it is impossible to proceed in real time. Instead of that, it is mainly used when making a movie or animation because of its high quality. To recover detailed 3D eye region facial expressions from video frames or pictures regardless of occlusion, I address this problem by representing a face as a set of facial blendshape meshes instead of facial landmarks. The landmark detection method finds several feature points in the face of an input image or video, displays them in 3d coordinates, and then reconstructs a new 3d mesh with extracted coordinates. However, this method shows a big difference in accuracy depending on the error occurring in converting the feature points to 3d coordinates, and the occlusion can greatly reduce the accuracy. On the other hand, blendshape, also can called by FACS system, is the way that categorize the physical expression of emotions with facial blendshape parameters, so it allows to make an expression through linear combination of expression parameters. With this method, only by finding out the used parameters and corresponding weights, can make high quality of reconstruction model. And it is more free from occlusion than the landmark detection method.

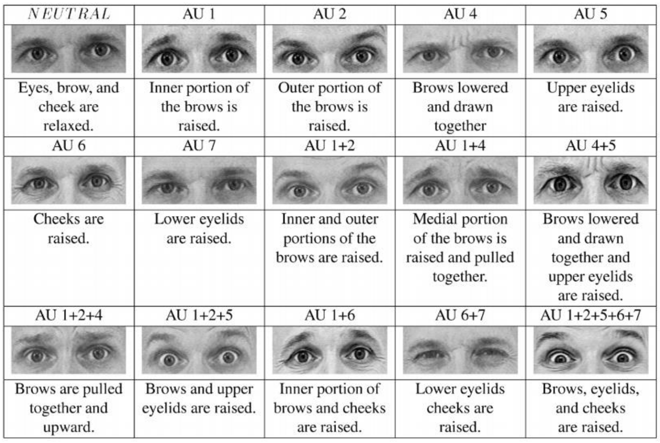
More concretely, let us assume we are given a generic blendshape model as a set of expression parameter meshes b = $\{b_1, \ b_2, \ \dots , \ b_N\}$. My target expression of frame I^t at time t can be formulated as

$$f^t = \sum_i^N (w_i^t * b_i) \tag{1}$$

Where $w^t \in [0,1]^N$ is a corresponding blendshape weight vector at time t. Then, my goal is to determine the value for w that best corresponds to a given image or video frame of a face. I formulate this as learning a mapping function ψ, which predicts a blendshape weight vector of image I^t, that minimizes objective function

$$\mathrm{L}(\psi) = \sum_t \| \psi(I^t) - w^t \|_2^2 \tag{2}$$

$\psi(I^t)$ also can be written as ground truth 3d reconstruction model's parameter weights.



NEUTRAL	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 1. illustration of blendshape. Result is linear combination of face blendshape parameter.

Recently, deep learning algorithms show a great achievement at many classification and regression tasks in computer vision. One key strength of deep learning methods is that they are capable of learning and optimizing high-dimensional functions and are robust to various appearance changes. In this paper, I present a deep learning framework that can extract high fidelity reconstructed eye model from videos and pictures. This neural network model is trained to receive a single HMD image to be reconstructed as an input and to export the FACS parameters and its corresponding weights as a output.

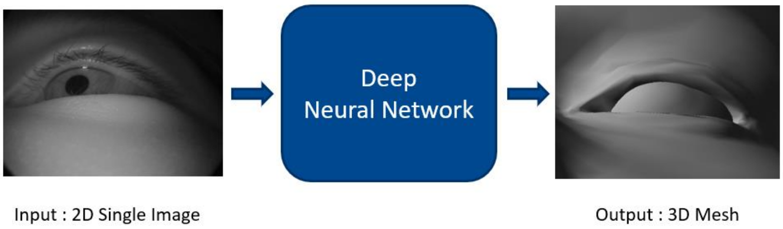


Figure 2. illustration of my approach. The input is a 2d HMD image and the output is a mesh of 3d coordinates.

A crucial aspect of my framework is the variety of expressions that can represent. Compared to the existing 3d reconstruction method that only expressed closed or opened eyes, this framework possible to reconstruction the expression of frowning or half-closed eyes. None of the existing eye dataset is suitable for training this neural network, so the training data to train our network was collected and edited in person.

Our results demonstrate that this framework can efficiently produce more detailed and efficient animations for the eye compared to traditional real-time performance-based facial reconstruction method.

Our contribution is two-fold:

- A regression method using deep convolutional neural networks that produces realistic eye expression. This network can successfully reconstruct half-closed, grimaces, and surprised expressions in addition to closed or open eyes. That means unlike conventional methods, more diverse expressions can be reproduced though this network.
- A set of training dataset that we use for this learning framework. After collecting the real dataset with HMD for eye expression learning, the 3d object mesh that used as ground truth in network training was created by experimenter manually in person.

The rest of this thesis is organized as follows. Section 2 reviews previous approaches closely related to mine, and Section 3 describes each step of my framework in details. Then clarify an experimental environment and result in Section 4, and conclusion in Section 5.

II. RELATED WORK

2.1 Real-time Facial animation

Facial performance capture techniques have been developed in the graphics and vision community to facilitate the production of compelling 3D facial animation [2]. While striving for increased tracking fidelity and realism, production-level methods often rely on complex capture equipment [3] and intensive computations [4]. Fully automatic techniques that do not require any user-specific training such as the regularized landmark meanshift method or the supervised descent algorithm have been recently proposed. While the mapping of sparse 2D facial features to the controls of complex 3D facial models has been explored, but only coarse facial expressions can be recovered. More recently, [5] developed a real-time system that can produce compelling 3D facial animations through a 3D shape regression technique from RGB videos. By directly regressing head motion and facial expressions, rather than regressing 3D facial landmarks and then computing the pose and expression from this data, were able to attain high tracking performance and accuracy, allowing for implementations running in real-time on mobile devices

III. METHOD

The presented approach to 3d reconstruction in eye region is roughly divided into four parts: (1) Using HMD to collect eye region images for the network training, (2) Manually reconstruct the eye region images into 3d object mesh using blendshape method, (3) Pair the parameters and weights used to create the 3d object mesh with each image and use them to train the network, (4) After the training, figure out the parameters and weights for each image that need to be reconstructed using trained network, and reconstruct in real-time using extracted parameters and weights. The remainder of this section describes characteristics of the above in details.

3.1 Collecting Dataset

In research how to reconstruct the eye region of user wearing an HMD, it is very important to collect an appropriate dataset to train the network. However, in recent papers related to face reconstruction, eye reconstruction is usually handled only as part of the task of reconstructing the entire face rather than as an independent task. The task of rebuilding the whole face follows the method of dividing the face into upper and lower faces, reconstructing them separately, and combining them together to finally rebuild the whole face. In most cases, however, the focus is more on reconstructing the lower face. The upper face with eyes improves the quality of reconstruction with texture mapping, but it has the limitation of reconstructing only simple expressions such as closing eyes or opening big eyes except other various expressions. So even if you have an open source dataset for the entire face 3d object mesh, there were no open source datasets specialized for reconstruction of the eyes. And even in the case of a data collection in the state-of-the-art papers, the eye parts lacked a description of how many people were taken, what facial expressions were taken, how many pictures were taken, and whether any of them were wearing glasses. So while I do this research, I built a new dataset specialized for reconstruction of the eyes.

This data must account for variations in the eye expression made by a given user, so I tried to shoot every possible eye expression : neutral, smiling, wide open eyes, half closed eyes, frowning and half closed eyes, closed eyes, process of closing eyes, and process of opening eyes. The expression of closing eyes is not much different for each person with HMD, and the process of closing eyes and the process of opening eyes are similar to that of half-closed eyes, so the closed eyes expression and the process of closing & opening eyes expression were taken only for 2 subjects. The closed eyes expression were photographed 5 times for each of the 2 subjects and the process of closing & opening eyes expression were photographed 10 times for each of the 2 subjects respectively. This process was repeated 4 times. The neutral expression were photographed 10 times for 16 subjects. The rest of expressions, smiling, wide open eyes, half closed eyes, frowning and half closed eyes, were photographed 15 times for 16 subjects. In order to take into account the diversity of users and the environment such as reflection of light, a total 18 subjects including 6 subjects wearing glasses and 12 subjects without glasses. The total collected HMD image data is 1320. The entire process of collecting data using HMD was performed under fluorescent lights. Due to the characteristics of the HMD, built-in cameras that photographed eye region output gray scale image, so no light calibration was performed.

This dataset can be very useful later in other studies because there is no open source dataset that has been reconstructed while intensively capturing eye expressions using HMD.

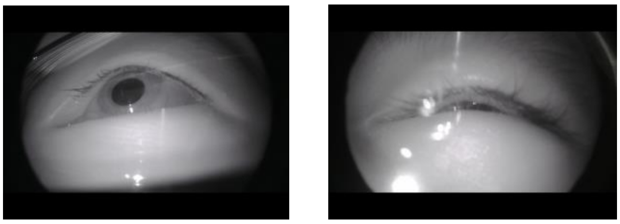


Figure 3. Samples of collected eye images.

3.2 Make Blendshape Parameters and Training dataset for Network

To create a blendshape parameters, I needed a 3d object that was closest to the real person. For this, I used CAESAR body database and D3DFACS dataset, which has 300 identity parameters, 100 expression parameters, and 15 pose parameters for women, men, and general category. With these three types of parameters, anyone’s face can be created. But because of the characteristics of the HMD camera that only shoots the eyes not the whole face, I didn’t use any identity or pose parameters. That means I create a ground truth model for the real HMD image dataset using only expression parameters. After checking each of the 100 expression parameter models, only 14 parameters contained eye expression. However, since these 14 expression parameters can’t represent the closed eyes, I have created and added one expression parameter representing the closed eyes. Using these 15 facial expression parameters, I created a reconstruction model for the real image manually and set the ground truth value to each parameter used and the weights multiplied by that parameters.

After 3d reconstruction of collected real images in this way, I decided to proceed principle component analysis(PCA) to select blendshape parameters specific to the eye reconstruction in more reasonable way. PCA does not analyze the components of each data, but rather analyzes the principal components of the distribution when several data are gathered together to form a distribution. Here, the principal component means a direction vector having the largest variance of data. Therefore, performing PCA on a 2-dimensional data set returns two mutually perpendicular principal component vectors, and performing a PCA on a 3-dimensional data set returns three mutually perpendicular principal component vectors. A representative example of such PCA is face image reconstruction that reconstructs a face by making an eigenfaces. I decided to apply this method to a coordinate system rather than an image to make eye specific blendshape parameters.

To do PCA, I need several datasets to extract the direction vectors. By the way, when proceeding PCA if the lower face part including the mouth is included, it is difficult to extract the eye specific parameters. So I proceed PCA after cropping only eye region of the face. After cropping, eye part was consisted of 1450 vertices in each 3d object mesh file. Since these 1450 vertices have three coordinates of x, y and z, each 3d object mesh could be though of as a 4350 dimensional vector. In other words, each real dataset corresponds to a point in the 4350 dimensional space. Now when PCA is performed on 1320 4350 dimensional point data, the same number of principal component vectors can be obtained as the number of dimensions of the data. The blendshape parameter is the result of reconstructing the principal component vectors into a 3d object mesh.

First, I made synthetic data by randomly weighting the 15 parameters used to make the ground truth mesh. I randomly selected 200,000 15 dimensional weight vectors, and conducted linear combination with 15 expression parameters to create 200,000 synthetic dataset. I extract the PCA parameters from the synthetic dataset and reconstruct the eye expression with the extracted parameters, but an accuracy is highly disappointed. When shooting the facial expressions with HMD, all possible facial expressions were taken, so I proceed PCA with the 1320 3d object mesh created by reconstructing the actual image, not the synthetic dataset. As a result, after reconstructing the eye expression with the extracted parameters, the results showed that the accuracy is highly convincing and there is no difference from the actual ground truth. In the end, I used the real dataset instead of the synthetic dataset to extract the eye specific blendshape parameters through PCA.

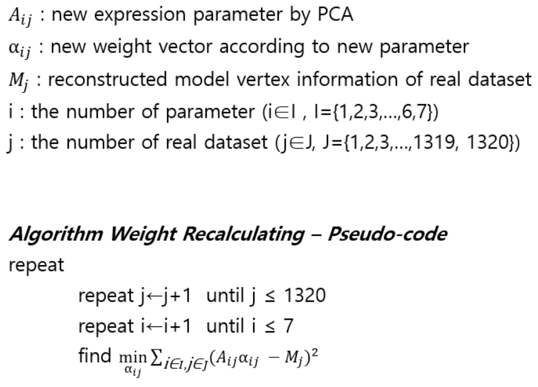
I extract the principal components in the order of greatest variance with PCA and need to pick principal components to use as the blendshape parameters. Since I created a 3d object mesh with 15 parameters, I thought it was meaningless to PCA if the number of new blendshape parameters exceed 15, so I limited the number of new parameters to 15 or less. The reconstruction was carried out by changing the number of principal components used and the loss with ground truth was measured according to the number of principal components used. In this case, loss is specified as the distance between the vertex of ground truth and the reconstructed model in 3 dimensions and the expression to be reconstructed was set to a neutral expression. The loss measurement results are shown in the table below.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Min	9.4	8.4	12.4	0.7	1.3	1.4	0.9	2.4	1.6	2.7	1.1	1.5	0.9	1.5	0
Max	212	182.7	192.1	226.7	61.5	60.3	62.6	62	62.2	61	60.6	60.8	26.7	27.2	1.9
Mean	76.7	75.7	75.3	33	15.5	14.2	13.8	13.2	12.8	12.6	12.6	12.5	7.1	6.9	3
sum	111192	109694.9	109116.2	47892.5	22470.4	20587.7	20022.3	19173.1	18561.3	18333.1	18217.4	18152.4	10327.6	9934.1	439.2

Table 1. loss between ground truth and reconstructed model along with the number of used parameters.

The first row of a table indicates the number of parameters used for reconstruction. The second row represents the smallest value of the loss, the third row represents the largest loss, the fourth row represents the average loss, and the last row represents the total loss. In the case of mean and sum, as the number of parameters used increases, the value decreases gradually. Obviously, the more parameters you use, the smaller the loss from the ground truth. However, in fact, after reconstruction there was no difference with the ground truth when I check it with my eyes. If sort min value in ascending order, 15, 4, 7, 13, 11, 5, 6, 12, 14, 9, 8, 10, 2, 1, 3 and then sort max value in ascending order, 15, 13, 14, 6, 11, 12, 10, 5, 8, 9, 7, 2, 3, 1, 4. Look at min value order, you can see that if you use 4 or 7 parameters, the value of min is lower than the other cases where have more parameters. In the case of using 7 parameters, the max and mean values are lower than those using 4 parameters. As a result, 7 principal components having the largest dispersion are selected as the new blendshape parameters.

Then I computed a new 7 dimensional weight to minimize the least square error compared to the 3 dimensional coordinates of the original ground truth 3d object mesh in order to find the weight for the new parameters.



A_{ij} : new expression parameter by PCA
α_{ij} : new weight vector according to new parameter
M_j : reconstructed model vertex information of real dataset
i : the number of parameter (i∈I , I={1,2,3,...,6,7})
j : the number of real dataset (j∈J, J={1,2,3,...,1319, 1320})

Algorithm Weight Recalculating – Pseudo-code

```
repeat
        repeat j←j+1  until j ≤ 1320
        repeat i←i+1  until i ≤ 7
        find $\min_{\alpha_{ij}} \sum_{i \in I, j \in J} (A_{ij}\alpha_{ij} - M_j)^2$
```

Figure 4. Pseudo code for calculating new weight for new blendshape parameter.

3.3 Learning a Network

For network learning the real image, the parameters that used to reconstruction the image, and the corresponding weights are needed. In this network, the real image enters to input as the target expression, and the parameters and weight values to reconstruct this target expression are out as network output. As the output should be accurate in real time, the network should be light and have good performance, so I modified the input and output parts of the VGG-16 network[4] that show fairly good performance in the image classification field. When training the network, minimize the L2 loss between the ground truth weight value and the predicted weight value from the network output.

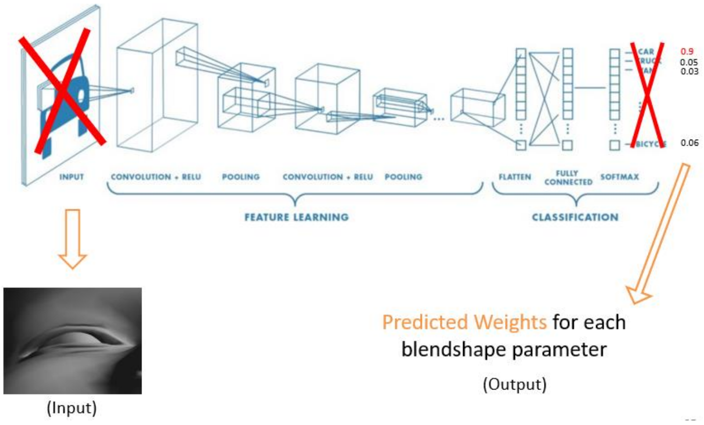


Figure 5. The model trained to regress blendshape weights for the eye region.

3.4 Reconstruction

After learning the network, you can get a 7 dimensional blendshape weight vector if you put the eye image into the network. By linearly combining this weight vector with the parameters, and then adding information about which vertices are connected to which vertices to form a face, I finally get a 3d object mesh that is reconstructed using the network results. Comparisons between the results of the reconstruction using the network output and the ground truth are described in the user study section of the conclusion section.

IV. EXPERIMENTS

I use the data generated in Section 3 to train the network. The network learning performed best with a batch of size 70 and epoch of 300, when 1320 dataset were used without data augmentation. The learning rate starts from 0.01 and it gradually decreases to 0.0001.

4.1 Training result and processing time

Some sample images demonstrating the variety of eye expressions our system is able to animate can be seen in Figure 6. The results demonstrate that this system is able to reconstruct a wide variety of facial expressions related to eye region, including important expressive details such as a user smiling while the process of closing or opening eyes. Other subtle details that are not attainable with existing real-time facial tracking techniques are also seen in these images.

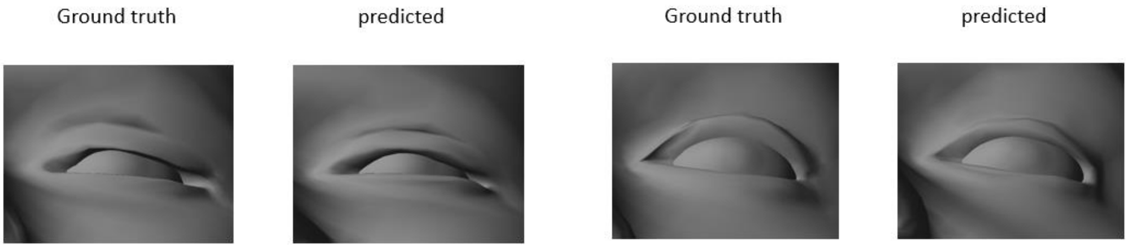


Figure 6. result of network learning. Since it was taken from the bottom of eyes like HMD camera angle (can be seen in Figure 8), it may not look much different from neutral expression, but these are frowned(left) and surprised(right) expression.

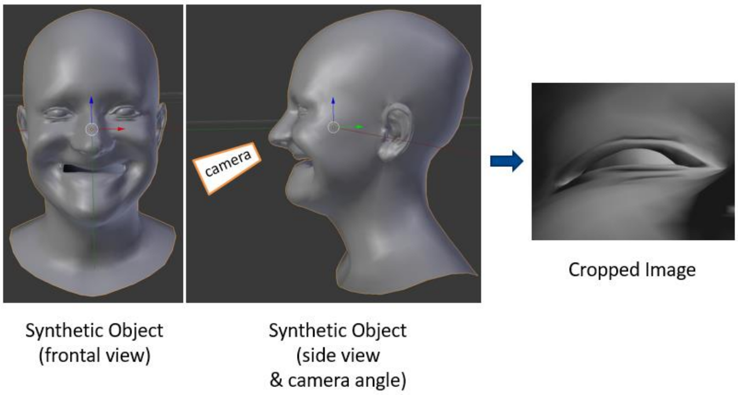


Figure 7. HMD camera angle example.

And also the results of the network learning show that there is no difference between the ground truth and the high performance in terms of accuracy. Now I measure the process time to see if it can be run in real time. To look natural to humans, need at least 30 frames to make a one-second video. Therefore, when measuring the process time, I measured the time for the predicted weight appeared for 30 pictures, and as a result, it took 500ms on average. So it can be said that it runs in real time.

V. CONCLUSION

To evaluate my system as a 3d reconstruction for eye region system, I compare mine with previous face reconstruction methods: [6] and [7].

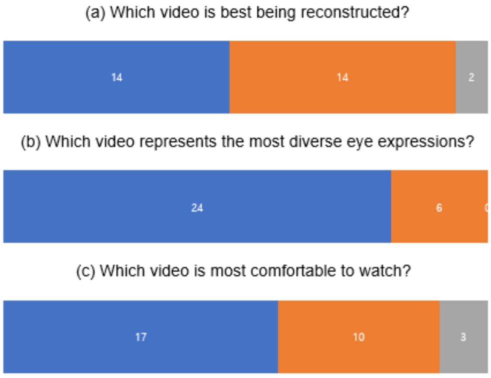


Figure 8. User study result on comparison with other face reconstruction methods.

The current my study didn't proceed until the texture mapping, so even in the user study, only 3d mesh was used except for the texture. However, in the case of [6], because the texture is coated, this part of the experiment was asked for understanding and judged only by the eye expression of the mesh except the texture. Since neither the [7] nor the [6] are open to the public, and there is not enough time to implement both papers, I used the result video of the paper for the user study.

For the user study, I recruited 30 participants. Participants are consist of 10 graduate students majoring in computer science but not related to computer graphics and 20 who do not related to computer science. I showed 3 videos and asked 3 questions, and the result is shown in figure 8. A blue bar is our approach, a orange bar is [7], and a gray bar is [6]. The results of the user study show that our proposed method is more effective to reconstruct the eye expression than the other 2 methods.

I have presented a method for animating a digital avatar in real-time based on the facial expressions of an HMD user. My system is more ergonomic than existing methods such as [1], makes use of more accessible components, and is more straightforward to implement. Furthermore, it achieves higher fidelity animations than can be achieved using existing methods. As such, it makes a significant step towards enabling eye emotional communication in VR, an important step for fully immersive social interaction through digital avatars. This system demonstrates that plausible real-time eye region animation is possible through the use of a deep neural net regressor, trained with animation parameters that not only capture the appropriate expressions of the training subjects, but that also make use of an appropriate psychoacoustic data set.

References

[1] LI, H., TRUTOIU, L., OLSZEWSKI, K., WEI, L., TRUTNA, T., HSIEH, P.-L., NICHOLLS, A., AND MA, C. 2015. Facial performance sensing head-mounted display. ACM Transactions on Graphics (Proceedings SIGGRAPH 2015) 34, 4 (July).

[2] PARKE, F. I., AND WATERS, K. 1996. Computer Facial Animation. A. K. Peters.

[3] LI, H., ADAMS, B., GUIBAS, L. J., AND PAULY, M. 2009. Robust single-view geometry and motion reconstruction. ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2009) 28, 5 (December).

[4] SHI, F., WU, H.-T., TONG, X., AND CHAI, J. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. ACM Trans. Graph. 33, 6, 222:1–222:13.

[5] CAO, C., WENG, Y., LIN, S., AND ZHOU, K. 2013. 3d shape regression for real-time facial animation. ACM Trans. Graph. 32, 4, 41:1–41:10.

[6] Thies, Justus & Zollhöfer, Michael & Stamminger, Marc & Theobalt, Christian & Nießner, Matthias. (2018). FaceVR: Real-Time Gaze-Aware Facial Reenactment in Virtual Reality. ACM Transactions on Graphics. 37. 1-15. 10.1145/3182644.

[7] Olszewski, Kyle & Lim, Joseph & Saito, Shunsuke & Li, Hao. (2016). High-fidelity facial and speech animation for VR HMDs. ACM Transactions on Graphics. 35. 1-14. 10.1145/2980179.2980252.

요 약 문

머리 부분 탑재형 디스플레이를 이용한 실시간 눈 부분 3차원 재건축

카메라로 촬영한 비디오나 사진을 3 차원 재건축하는 연구는 과거부터 큰 이슈였다. 머신러닝이 발전하면서 얼굴 주름도 실제처럼 자세하게 재건축이 가능해지거나 얼굴 표정을 실시간으로 재건축하는 것도 가능해졌지만 눈 부분을 중점적으로 재건축하는 내용의 논문은 발표되지 않았다. 본 논문에서는 눈 부분을 중점적으로 재건축하는 새로운 방법을 소개함과 동시에 복잡한 하드웨어 장치 환경 없이 머리 부분 탑재형 디스플레이만을 이용해 실시간으로 3 차원 재건축하는 새로운 방법을 소개한다. 재건축을 위해 기계학습을 이용해 인공 신경망을 구축하고 실시간으로 동작하기 위해 최대한 가벼운 인공 신경망을 사용했다. 학습을 위한 데이터는 직접 머리 부분 탑재형 디스플레이를 이용해 수집한 뒤 가공하여 사용함으로써 자체적인 데이터셋을 완성하였다. 이 인공 신경망 모델은 오직 한 장의 이미지만으로도 3 차원 재건축이 가능하게 하며 이전 3 차원 재건축 모델들이 눈을 감거나 뜨는 표정만을 구현할 수 있는 반면 찡그리거나 크게 뜨는 등의 표정도 구현하게 함으로써 좀 더 다양한 표정을 나타낼 수 있게 한다.

핵심어: 3차원 재건축, 실시간 눈 표정 포착, 가상 현실