Journal of Multimedia Information System VOL. 10, NO. 2, June 2023 (pp. 101-108): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2023.10.2.101

Temporal Facial Alignment with Triple Discriminators

Seongmin Lee[1], Hyunse Yoon[1], Sanghoon Lee[1], Jiwoo Kang[2*]

Abstract

Facial alignment is one of the most important features of 3D facial models. However, when fitting the 3D Morphable Model (3DMM) into a large expression or pose facial image, there is ambiguity as to whether facial shape deformation is caused by identity or expression. To address this, in this paper, we propose a stable and accurate facial alignment framework by introducing multiple stability discriminators. The proposed discriminators determine the regressed camera, face identity, and expression parameters simultaneously from an image. The proposed framework for facial alignment consists of a facial alignment network and stability discriminators: identity, expression, and temporal discriminators. The facial alignment network is trained to predict camera, face identity, and expression parameters based on an image. The stability discriminator is trained to distinguish whether the facial deformation generated by the estimated facial identity and expression parameters is stable. Meanwhile, the discriminator distinguishes whether the deformation between adjacent frames is consistent. By utilizing these stability discriminators, the proposed facial alignment network demonstrates precise and consistent performance in aligning faces in both static and dynamic domains. To verify the performance of the proposed discriminators, the large-scale facial tracking dataset, 300 Videos in the Wild (300VW) dataset, is used for qualitative and quantitative evaluations. The experimental results show that the proposed method outperforms state-of-the-art methods, demonstrating the strong benefits of our method in accurate facial alignment over time.

Key Words: Facial Alignment, Facial Tracking, Temporal Stability.

I. INTRODUCTION

3D facial models have been widely used in various facial applications, such as facial animation, facial synthesis, facial reconstruction, facial recognition, and facial tracking. To use 3D facial models, facial alignment, which is the process of moving and deforming a facial model to an image, is an essential pre-processing step. Since the human face has a regularized structure of facial components such as eyes, lips, and nose, facial alignment is performed efficiently using this as prior. However, in traditional facial alignment methods, there is alignment instability in large pose and expression changes. In the cases of large changes in pose and expression, it is unclear whether the change in facial shape is caused by identity, expression, or pose. If this ambiguity is expanded to the temporal domain, unnatural facial shape changes and jittering artifacts arise, and significant quality degradation occurs visually. To address this, in this paper, we propose a stabilized facial alignment framework in identity, expression, and temporal changes.

The 3D Morphable Model (3DMM), a statistical model of 3D faces, is the most widely used representative model to obtain the 3D face from a facial image in various face-related applications. Since the first 3DMM was introduced [1], variants of 3DMM have been built by decomposing facial scans of various identities and expressions using Principal Component Analysis (PCA) to represent an arbitrary human face. It can efficiently represent a 3D face from a target facial image. However, when fitting the 3DMM into a large expression or pose facial image, there is ambiguity in the facial shape whether facial shape deformation is caused by identity or expression. It does not cause large visual degradation in a static domain; however, it occurs in large visual degradation, such as unnatural facial shape changes and jittering artifacts in a temporal domain.

Recently, with the expansion of the Generative Adversarial Network (GAN) in deep learning, it has been found that using discriminators leads to a network with higher performance [2]. GAN is composed of two networks: a generator and a discriminator. The discriminator is trained to determine whether the input data distribution is close to the ground-truth data distribution or the generated data distribution. At the same time, the generator is trained to fool the discriminator, by generating more accurate data. Motivated

Manuscript received May 24, 2023; Revised June 01, 2023; Accepted June 03, 2023. (ID No. JMIS-23M-05-019)
Corresponding Author (*): Jiwoo Kang, +82-2-2077-7445, jwkang@sookmyung.ac.kr
[1]Department of Electronical and Electronic Engineering, Yonsei University, Seoul, Korea, lseong721@yonsei.ac.kr, hsyoon97@yon sei.ac.kr, slee@yonsei.ac.kr
[2]Divion of Artificial Intelligence Engineering, Sookmyung Women's University, Seoul, Korea, jwkang@sookmyung.ac.kr

by this, we propose a stable and accurate facial alignment framework by introducing stability discriminators that determine that the regressed camera and facial shape parameters are stable. The proposed facial alignment framework consists of a facial alignment network and stability discriminators. The facial alignment network is trained to regress camera, face identity, and expression parameters from an image. The stability discriminator is trained to discriminate whether the facial deformation generated from the estimated facial identity and expression parameters is stable and whether the deformation between neighbor frames is stable. Using these stability discriminators, the proposed facial alignment network shows accurate and stable facial alignment performance in both the static and temporal domains. The 300 Videos in the Wild (300VW) dataset [3], which provides large-scale facial tracking data, is used for qualitative and quantitative evaluations. In the experimental results, the proposed method shows significant improvements over state-of-the-art methods for temporal facial alignment. The results demonstrate that the proposed method enables accurate facial tracking with multiple discriminators by stabilizing facial locations and shapes over time.

II. RELATED WORKS

2.1. 3D Morphable Model

From the first 3DMM introduced by Blanz [1], various 3DMMs have been proposed [4-6]. From facial scans collected from multiple subjects, the features of 3D facial scans for identity, expression, and texture have been encoded using PCA decomposition. Since each facial scan has a different topology, mesh registration is required to find vertex correspondences. In [1], optical was used to find the correspondence between the vertex between facial scans. In [4], for an accurate alignment, non-rigid registration method, warping based on Thin-Plate Splines (TPS) [7] and non-rigid Iterative Closest Point (ICP) [8] was used. In [5], a multilinear facial model was proposed to represent facial identity and expression using the Singular Value Decomposition (SVD). In [6], based on the multilinear model, a bilinear facial model of identity and expression was constructed by deforming the facial scan into the template model with expression. Due to the many efforts to build an accurate 3DMM, an arbitrary 3D face can be accurately and effectively represented using 3DMM.

2.2. 3D Facial Alignment

3D facial alignment aims to fit a 3DMM to a facial image. The first 3D facial alignment method [9] is performed by fitting the 3DMM, minimizing the pixel-wise difference between the target facial image and a rendered image of the 3DMM. In recent years, regression-based 3D facial alignment has been introduced that minimizes the difference between the target 2D landmark and the projected 2D landmark of 3DMM [10-13]. These methods have shown performance improvement; however, there remain two major challenges. First, self-occlusion arises due to a large pose or expression. Due to self-occlusion, facial semantic information is lost, and unreliable facial alignment may occur. Second, in the temporal sequences, temporal instability arises due to a large and fast motion. The results of facial alignment may look reliable in the static shot, but in the temporal sequence, jittering artifacts on facial alignment usually occur. To address these problems, in this paper, we propose novel stabilization discriminators that guide changes in the stabilized facial shape in large poses, expressions, and motion.

III. METHOD

The proposed method utilizes a facial alignment network with 3DMM to produce accurate 3D facial alignment. Multiple discriminators are employed to ensure consistent facial alignments with an individual's identity and expression over time. Fig. 1 illustrates the overall framework of the proposed method for facial alignment.

3.1. 3D Morphable Model

A 3DMM represents an arbitrary 3D face using bases decomposed through PCA. Using the 3DMM, the 3D face (**S**) can be represented by parameters for both identity and expression, as follows:

$$\boldsymbol{S} = \overline{\boldsymbol{S}} + \boldsymbol{A}\ \alpha, \tag{1}$$

where $\boldsymbol{F} \in R^{3\times N}$ is the 3D face with N vertices, $\mathbf{F}$ is the mean face, $\mathbf{A}$ represents the 3D shape bases, and α is the shape parameter corresponding to $\mathbf{A}$. The facial shape bases, denoted by $\boldsymbol{A}$, are decomposed into two parts: $\boldsymbol{A} = [\boldsymbol{A_{id}}, \boldsymbol{A_{exp}}]$. Here, $\boldsymbol{A_{id}}$ is trained using 3D facial meshes with a neutral expression, while $\boldsymbol{A_{exp}}$ is computed as the difference between the facial mesh with expression and the neutral facial mesh. The facial shape parameter $\alpha = [\alpha_{id}, \alpha_{exp}]$ is divided into two components: α_{id} and α_{exp}, which correspond to the facial identity and expression bases, respectively. In this paper, $\boldsymbol{A_{id}}$ and $\boldsymbol{A_{exp}}$ are from the Basel Face Model [4] ($\alpha_{\mathrm{id}} \in \mathbb{R}^{199}$) and FaceWareHouse [6] ($\alpha_{\mathrm{exp}} \in \mathbb{R}^{300}$), respectively. By applying a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$, a 2D translation vector $\mathbf{t} \in \mathbb{R}^{3}$, a focal length scalar f, and a projection matrix $\mathbf{P}$, the 3D face is projected onto image coordinates $\mathbf{v}$ as follows:

Journal of Multimedia Information System VOL. 10, NO. 2, June 2023 (pp. 101-108): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2023.10.2.101

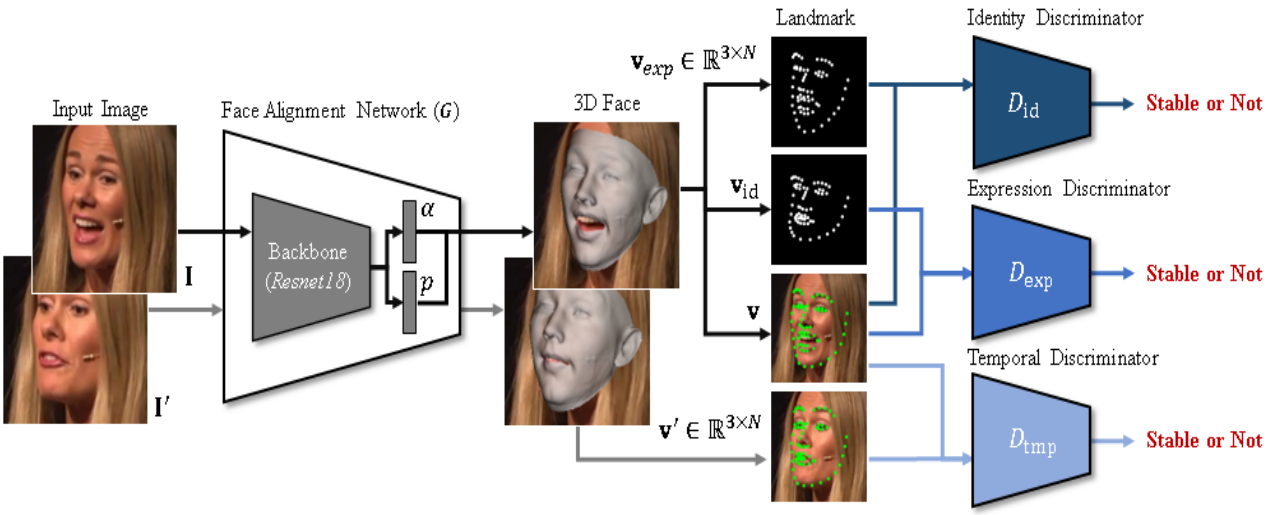


Fig. 1. Overview of the proposed facial alignment framework, which comprises the facial alignment network, denoted as $\boldsymbol{G}$, and discriminators for identity ($\boldsymbol{D_{\mathrm{id}}}$), expression ($\boldsymbol{D_{\mathrm{exp}}}$), and temporal ($\boldsymbol{D_{tmp}}$) factors. The Facial Alignment Network is used to align a 3D Morphable Model (3DMM) with facial landmarks and to estimate the shape parameters shape ($\boldsymbol{\alpha}$) and location parameters ($\boldsymbol{p}$) by aligning the 3DMM with the facial landmarks. Discriminators are trained to determine whether changes in facial shape are consistent with the individual's identity and expression over time. For temporal discrimination, the facial alignment network is trained using images of a current frame ($\mathbf{I}$) and the previous frame ($\mathbf{I}'$).

$$\boldsymbol{v} = f\boldsymbol{P} \ * \ \boldsymbol{R} \ * \ \boldsymbol{S} + \boldsymbol{t}, \tag{2}$$

where $*$ is the matrix multiplication operator.

We adopt *ResNet-50* as the facial alignment network G to estimate the camera and shape parameters. We divided the 2048 output features from the final pooling layer into two sections: 128 features for camera parameters and 1920 features for facial shape parameters. Two fully connected layers, each with 512 hidden nodes, are added to each component to estimate the camera parameters, $[\boldsymbol{R},\ \boldsymbol{t},\ f]$, and the shape parameters, $[\alpha_{\mathrm{id}},\ \alpha_{\mathrm{exp}}]$.

3.2. Facial Alignment Network

Given a 2D image $\mathbf{I}$, the facial alignment network G encodes the camera parameters $p = [f, \boldsymbol{R}, \boldsymbol{t}]$ and shape parameter α. Then, the projected landmark of the 3D face is estimated using a landmark index vector $\boldsymbol{l} \in \mathbb{R}^{68}$. The proposed facial alignment network is trained using the L_{land} loss for landmark s and regularization loss for shape and expression parameters. The landmark loss, denoted as L_{land}, is defined as follows:

$$L_{\mathrm{land}} = |\boldsymbol{v}(:, \boldsymbol{l}) \ - \ \boldsymbol{U}|_2, \tag{3}$$

where $\boldsymbol{U}$ is the labeled ground-truth 2D landmark locations of the input image. To prevent the generation of undesired facial shapes and expressions, L_2 regularization terms are applied to the shape and expression parameters. Specifically, the shape regularization term is denoted $L_{\mathrm{id}} = |\alpha_{\mathrm{id}}|_2$, while the expression regularization term is denoted as $L_{\mathrm{exp}} = |\alpha_{\mathrm{exp}}|_2$. The loss of facial alignment is defined by combining these losses in the following manner:

$$L_{G_{\mathrm{align}}} = L_{\mathrm{land}} + \lambda_{\mathrm{id}} L_{\mathrm{id}} + \lambda_{\mathrm{exp}} L_{\mathrm{exp}}, \tag{4}$$

where λ_{id} and λ_{exp} are the balancing factors between losses and we set to $\lambda_{\mathrm{id}} = \lambda_{\mathrm{exp}} = 0.0001$ in our experiments. Sufficiently small values of the balancing factors help avoid parameter overfitting without sacrificing alignment accuracy.

3.3. Identity, Expression, and Temporal Discriminators

To train the facial alignment network to stabilize in both the temporal and static domains, we propose three different discriminators: identity, expression, and temporal cues. The identity and expression discriminator stabilizes the facial alignment network in a static domain by distinguishing between the estimated changes in facial shape based on the identity and expression parameters. The temporal discriminator stabilizes facial alignment over time by distinguishing changes in facial shape over time.

The identity discriminator determines whether the estimated changes in facial shape correspond to the desired facial shape based on the regressed facial identity parameter. The identity discriminator is trained by calculating the difference between facial landmarks and estimated landmarks without considering identity. To calculate the difference in facial landmarks, the facial shape without identity needs to be estimated as follows:

$$\boldsymbol{S}_{\exp} = \overline{\boldsymbol{S}} + \boldsymbol{A}_{\exp}\alpha_{\exp}. \quad (5)$$

$$\boldsymbol{v}_{\exp} = f\boldsymbol{P} * \boldsymbol{R} * \boldsymbol{S}_{\exp} + \boldsymbol{t}. \quad (6)$$

The facial landmarks can be detected from the projected facial vertices. However, since the landmark is located in the image coordinates, we normalize both the ground truth and the estimated landmark to the range of [0, 1] before calculating the difference. The difference to be used as input for the identity discriminator is computed by using the normalized landmarks as follows:

$$\boldsymbol{x}_{\mathbf{id}} = \boldsymbol{U} - \boldsymbol{v}_{\exp}(:, \boldsymbol{l}). \quad (7)$$

$$\boldsymbol{x}_{\boldsymbol{z},\mathrm{id}} = \boldsymbol{v}(:, \boldsymbol{l}) - \boldsymbol{v}_{\exp}(:, \boldsymbol{l}), \quad (8)$$

where $\boldsymbol{x_{id}}$ is the difference calculated from the ground-truth landmark and $\boldsymbol{x_{z,id}}$ is the difference calculated from the estimated landmark. To train the identity discriminator, $\boldsymbol{x_{id}}$ is used as the real distribution and $\boldsymbol{x_{z,id}}$ is used as fake distribution. The loss of identity discrimination is defined as follows:

$$\begin{aligned} L_{D_{\mathrm{id}}} = {} & E_{\boldsymbol{x}_{\mathrm{id}}}\left[\log\left(D_{\mathrm{id}}(\boldsymbol{x}_{\mathrm{id}})\right)\right] \\ & + E_{\boldsymbol{x}_{\boldsymbol{z},\mathrm{id}}}\left[\log\left(1 - D_{\mathrm{id}}(\boldsymbol{x}_{\boldsymbol{z},\mathrm{id}})\right)\right]. \end{aligned} \quad (9)$$

Similar to the identity discriminator, the expression discriminator is trained to determine the change in facial shape based on the validity of the expression parameter. The facial shape without expression, denoted as $\boldsymbol{S_{id}}$, is calculated by replacing $\boldsymbol{A}_{\exp}$ and $\alpha_{\exp}$ with $\boldsymbol{A}_{\mathrm{id}}$ and α_{id} in (5), and the projected facial shape, denoted as $\boldsymbol{v_{id}}$, is then calculated using equation (2). Then, the difference between the facial landmarks with and without expression is computed by replacing $\boldsymbol{v}_{exp}$ with $\boldsymbol{v}_{id}$ in equation (7) and (8). To train the expression discriminator, we use the differences between the calculated landmarks without expression and the ground truth landmarks ($\boldsymbol{x_{exp}}$) as real distribution. Likewise, the differences between the estimated landmarks with expression and the ground truth landmarks ($\boldsymbol{x_{z,exp}}$) is used as fake distributions. Therefore, the expression discriminator loss is defined as follows:

$$\begin{aligned} L_{D_{\exp}} = {} & E_{\boldsymbol{x}_{\exp}}\left[\log\left(D_{\exp}(\boldsymbol{x}_{\exp})\right)\right] \\ & + E_{\boldsymbol{x}_{\boldsymbol{z},\exp}}\left[\log\left(1 - D_{\exp}(\boldsymbol{x}_{\boldsymbol{z},\exp})\right)\right]. \end{aligned} \quad (10)$$

The identity and expression discriminators stabilize the facial alignment network in a static domain. For improved temporal stabilization performance, we propose a temporal discriminator that can accurately judge the validity of any temporal changes in facial shape. The variation in facial landmarks between the current and previous frames is utilized as input for the temporal discriminator. Facial temporal changes are determined by calculating the difference between the current and previous frames as follows:

$$\boldsymbol{x}_{\mathrm{tmp}} = \boldsymbol{U} - \boldsymbol{U}'. \quad (11)$$

$$\boldsymbol{x}_{\boldsymbol{z},\mathrm{tmp}} = \boldsymbol{v}(:, \boldsymbol{l}) - \boldsymbol{v}'(:, \boldsymbol{l}), \quad (12)$$

where $\boldsymbol{v}'$ and $\boldsymbol{U}'$ are the projected vertices and the ground-truth landmark of the previous frame, respectively. The temporal discriminator loss is defined as follows:

$$\begin{aligned} L_{D_{\mathrm{tmp}}} = {} & E_{\boldsymbol{x}_{\mathrm{tmp}}}\left[\log\left(D_{\mathrm{tmp}}(\boldsymbol{x}_{\mathrm{tmp}})\right)\right] \\ & + E_{\boldsymbol{x}_{\boldsymbol{z},\mathrm{tmp}}}\left[\log\left(1 - D_{\mathrm{tmp}}(\boldsymbol{x}_{\boldsymbol{z},\mathrm{tmp}})\right)\right]. \end{aligned} \quad (13)$$

These multiple discriminators are trained to distinguish whether identity, expression, and temporal changes are valid. The facial alignment network, on the other hand, is trained to fool these discriminators. The total adversarial loss for these discriminators, namely $D_{\mathrm{id}}, D_{\exp}$, and D_{tmp}, is defined as follows:

$$L_D = \lambda_{\mathrm{id}} L_{D_{\mathrm{id}}} + \lambda_{\exp} L_{D_{\exp}} + \lambda_{\mathrm{tmp}} L_{D_{tmp}}, \quad (14)$$

where λ_{id}, $\lambda_{\exp}$, and λ_{tmp} are the balancing factors. The total loss for the facial alignment network (**G**) is defined by combining the alignment and adversarial losses as follows:

$$L_{G_{\mathrm{id}}} = E_{\boldsymbol{x}_{\boldsymbol{z},\mathrm{id}}}\left[\log\left(D_{\mathrm{id}}(\boldsymbol{x}_{\boldsymbol{z},\mathrm{id}})\right)\right]. \quad (15)$$

$$L_{G_{\exp}} = E_{\boldsymbol{x}_{\boldsymbol{z},\exp}}\left[\log\left(D_{\exp}(\boldsymbol{x}_{\boldsymbol{z},\exp})\right)\right]. \quad (16)$$

$$L_{G_{\mathrm{tmp}}} = E_{\boldsymbol{x}_{\boldsymbol{z},\mathrm{tmp}}}\left[\log\left(D_{\mathrm{tmp}}(\boldsymbol{x}_{\boldsymbol{z},\mathrm{tmp}})\right)\right]. \quad (17)$$

$$L_G = L_{G_{\mathrm{align}}} + \lambda_{\mathrm{id}} L_{G_{\mathrm{id}}} + \lambda_{\exp} L_{G_{\exp}} + \lambda_{\mathrm{tmp}} L_{G_{\mathrm{tmp}}}. \quad (18)$$

In our experiments, we set balancing factors to $\lambda_{\mathrm{id}} = \lambda_{\exp} = \lambda_{\mathrm{tmp}} = 0.1$ for discriminators and facial alignment network. The same network structure is used for all discriminators. From the landmark difference, the two fully connected layers with 256 hidden nodes are used to deter mine the stability, a single scalar value ranging from 0 to 1.

IV. EXPERIMENTS

4.1. Implementation Details

The 300VW dataset [3], which provides large-scale facial tracking data, is used for both qualitative and quantitative evaluations. The 300VW dataset comprises 114 videos,

totaling 218,595 frames, each with 68-point landmark labels. Out of 114 videos, 50 are allocated for training purposes, while the remaining 64 are designated for testing. The test videos are divided into three categories (A, B, and C), with C being the most challenging test set. For training purposes, each frame is cropped using a ground-truth landmark and resized to 256×256 *pixels* to be used as input for the facial alignment network. To enhance the network's resilience to temporal changes, the frame interval between the current and previous frames is randomly increased within the range of 1 to 6. After a roughly aligned network is formed, each frame is cropped using a landmark estimated from the previous frame. In the testing phase, the first frame is cropped based on the landmarks detected using a conventional landmark detection algorithm called MTCNN [14]. From the second frame onward, each subsequent frame is cropped using a landmark estimated from the previous frame. The proposed method used in all experiments was trained for 500 epochs using Tensorflow (version 2.10.0), (CUDNN version 8.1), and CUDA (version 11.2). We used the Adam optimizer for optimization and trained the model on a single NVIDIA 2080Ti (11GB) with a batch size of 20.

We used a learning rate of 0.001 during the initial training phase, which gradually decreased to 0.00001.

4.2. Performance Evaluation

For evaluation, we compare our method with other state-of-the-art facial alignment methods: 3DDFA [15] and RingNet [16], DSFNet [17]. For quantitative comparison, we measure the Normalized Mean Error (NME) of the 2D facial landmarks. NME is calculated as the average normalized landmark error divided by the facial bounding size based on previous facial alignment methods [18-19]. The size of the facial bounding box is defined as the square root of the product of the width and height of the rectangular hull calculated from all the landmarks. The quantitative measurements are summarized in Table 1. Also, Fig. 2 visualizes some examples of the 3D facial alignment results. In the experimental results, the proposed method outperforms other methods in every case. Especially, our method has a distinct advantage in the tracking challenging case (300VW-C) compared to other methods.

Table 1. 2D Facial alignment accuracy (%) on 300VW dataset.

Type	300VW-A	300VW-B	300VW-C
3DDFA	2.913	3.035	3.387
RingNet	2.845	2.983	3.343
DSFNet	2.799	2.878	3.214
Ours	**2.495**	**2.549**	**2.734**

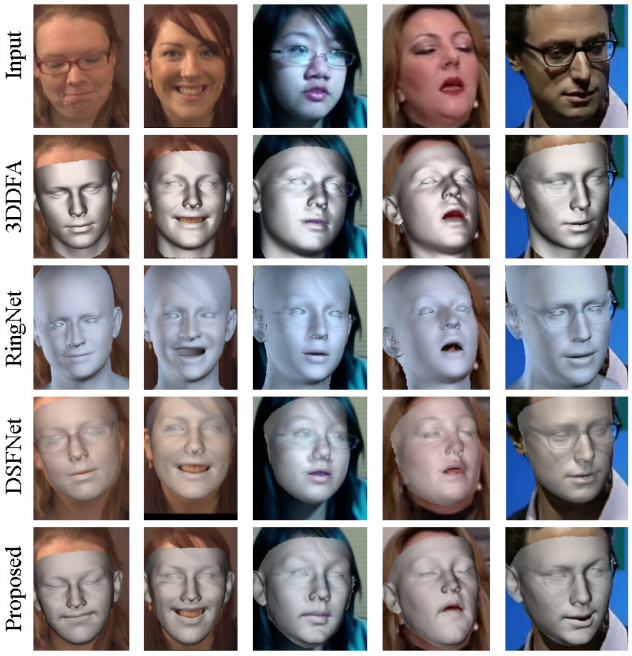


Fig. 2. Facial alignment results on 300 Videos in the Wild (300VW) dataset. The proposed method shows outperformed results over other methods. In particular, our method can capture expression details accurately.

We performed four ablation tests on the discriminator to verify the primary contributions of the proposed multiple discriminators. For the baseline, we trained the facial alignment network without using any discriminator. On the baseline, each discriminator is used to evaluate its own performance. The ablation tests are conducted by measuring the accuracy of facial alignment, which is represented by the Normalized Mean Error (NME). The results of these tests are presented in Table 2.

The results show that discriminations on an individual's identity, expression, and temporal changes give distinct performance gains. In particular, temporal discrimination is

Table 2. Ablation tests according to the discriminator type.

Type	300VW-A	300VW-B	300VW-C
Without all	3.731	4.060	4.677
With $\boldsymbol{D_{id}}$	3.315	3.575	3.912
With $\boldsymbol{D_{exp}}$	3.133	3.284	3.665
With $\boldsymbol{D_{tmp}}$	3.078	3.192	3.504
With $\boldsymbol{D_{id}}$, $\boldsymbol{D_{exp}}$	3.099	3.227	3.601
With $\boldsymbol{D_{id}}$, $\boldsymbol{D_{tmp}}$	2.912	3.085	3.311
With $\boldsymbol{D_{exp}}$, $\boldsymbol{D_{tmp}}$	2.721	2.801	3.057
With all (Ours)	**2.495**	**2.549**	**2.734**

shown to play the most important role in accomplishing stable alignments in time, while identity discrimination plays the least. By comparing the result in Table 2, it is demonstrated that using multiple discriminations on temporal identity and expression simultaneously gives strong benefits to obtaining stable 3D faces.

V. CONCLUSION

In this paper, we propose a stable and accurate facial alignment framework by introducing multiple stability discriminators. The proposed discriminators determine the regressed camera, face identity, and expression parameters simultaneously from an image. The proposed framework consists of a facial alignment network and multiple discriminators: identity, expression, and temporal discriminators. To verify the performance of the proposed discriminators, the large-scale facial tracking dataset, 300VW dataset, is used for qualitative and quantitative evaluations. The experimental results show significant performance improvements over state-of-the-art methods, demonstrating the strong benefits of our method in accurate facial alignment over time. We believe that our work would be helpful in various facial applications, such as facial recognition [20-21].

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00229451, Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms).

REFERENCES

[1] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques,* 1999, pp. 187-194.

[2] Y. J. Heo, B. G. Kim, and P. P. Roy"Frontal face generation algorithm from multi-view images based on generative adversarial network," *Journal of Multimedia Information System*, vol. 8, no. 2, pp. 85-92, 2021.

[3] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 50-58.

[4] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance,* 2009, pp. 296-301.

[5] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, "Face transfer with multilinear models," *ACM SIGGRAPH 2006 Courses*, 2006, p. 24-es.

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413-425, 2013.

[7] F. Bookstein and W. Green, "A thin-plate spline and the decomposition of deformations," *Mathematical Methods in Medical Imaging,* vol. 2, pp. 14-28, 1993.

[8] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[9] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063-1074, 2003.

[10] J. Kang and S. Lee, "A greedy pursuit approach for fitting 3D facial expression models," *IEEE Access*, vol. 8, pp. 192 682-192 692, 2020.

[11] A. Jourabloo and X. Liu, "Pose-invariant face alignment via CNN-based dense 3D model fitting," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 187-203, 2017.

[12] J. Kang, S. Lee, and S. Lee, "Competitive learning of facial fitting and synthesis using uv energy," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 5, pp. 2858-2873, 2021.

[13] J. Kang, H. Song, K. Lee, and S. Lee, "A selective expression manipulation with parametric 3D facial model," *IEEE Access*, vol. 11, pp. 17066-17084, 2023.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

[15] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 152-168.

[16] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7763-7772.

[17] H. Li, B. Wang, Y. Cheng, M. Z. Kankanhalli, and R. T. Tan, " DSFNet: Dual Space Fusion Network for Oc-

clusion-Robust 3D dense face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4531-4540.

[18] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4188-4196.

[19] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78-92, 2017.

[20] A. K. Y. Yoon, K. C. Park, S. M. Park, D. K. Oh, H. Y. Cho, and J. H. Jang, et al., "Three-dimensional face recognition based on feature points compression and expansion," *Journal of Multimedia Information System*, vol. 6, no. 2, pp. 91-98, 2019.

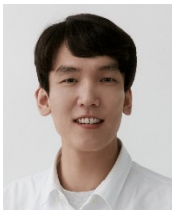
[21] A. K. Y. Yoon, K. C. Park, B. C. Lee, J. H. Jang, "A study on overcoming disturbance light using polarization filter and performance improvement of face recognition system," *Journal of Multimedia Information System*, vol. 7, no. 4, pp. 239-248, 2020.

AUTHORS

Seongmin Lee was born in South Korea in 1992. He received the B.S. degree in Electronic and Electrical Engineering from Hongik University, Seoul, South Korea, in 2018. He is currently pursuing the M.S. and Ph.D. degrees in Electrical and Electronic Engineering with the Multidimensional Insight Laboratory, Yonsei University, Seoul. His research interests are in the area of computer vision, computer graphics, and deep learning.

Hyunse Yoon was born in South Korea in 1997. He received the B.S. degree in Computer Science from the University of Hong Kong, Hong Kong, in 2020. He is currently pursuing the M.S. and Ph.D. degrees in Electrical and Electronic Engineering with the Multidimensional Insight Laboratory, Yonsei University, Seoul. His research interests are in the area of sensors, computer vision, computer graphics, and deep learning.

Sanghoon Lee received the B.S. in Electronic Engineering from Yonsei University in 1989 and the M.S. in Electronic Engineering from KAIST in 1991. From 1991 to 1996, he worked for Korea Telecom. He received his Ph.D. in Electronic Engineering from the University of Texas at Austin in 2000. From 1999 to 2002, he worked for Lucent Technologies. In March 2003, he joined the faculty of the Department of Electrical and Electronics Engineering, Yonsei University, Seoul, Korea, where he is a Full Professor. He was an Associate Editor of the IEEE Trans. Image Processing (2010−2014), an Editor of the Journal of Communications and Networks (JCN) (2009−2015), an Associate Editor of IEEE Signal Processing Letters (2014−2018), a Guest Editor for IEEE Trans. Image Processing (2013) and Journal of Electronic Imaging (2015), and a Senior Area Editor of IEEE Signal Processing Letters (2014−2022). He has been a Chair of the IEEE P3333.1 Quality Assessment Working Group (2011−), an Associate Editor of the IEEE Trans. Multimedia (2022−) and a Member of the Senior Editorial Board of the IEEE Signal Processing Magazine (2022−). He was a Chair and Member of the APSIPA IVM Technical Committee (2018−2019) and (2014–2017), respectively, and a Member in the Technical Committees of the IEEE IVMSP (2014−2020). He currently serves as a Member in the Technical Committees of the IEEE MMSP (2016−). He has participated in the international activities as a General Chair of the 2013 IEEE IVMSP Workshop, a Technical Program Co-chair of the IEEE International Conference on Multimedia and Expo (ICME) 2018, the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2018, the International Conference on Information Networking (ICOIN) 2014, the Global 3D Forum 2012 and 2013, and the Exhibition Chair of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018, the Local Arrangements Chair of IEEE International Symposium On Broadband Multimedia Systems and Broadcasting (BMSB), 2012. He received a 2012 Special Service Award from the IEEE Broadcast Technology Society, a 2013 Special Service Award from the IEEE Signal Processing Society, a Humantech Thesis Award of Samsung Electronics, 2013, an IEEE Seoul Section Student Paper Contest Award, 2012, a Qualcomm Innovation Award, Qualcomm, 2012, and a Best Student Paper Award of International Conference on Quality of Multimedia Experience 2018. His research interests include deep learning, image&video quality of experience, computer vision and computer graphics.



Jiwoo Kang received the B.S. in 2011 from Yonsei University, Seoul, South Korea, in electrical and electronic engineering. He received the M.S. and the Ph.D. in 2019, both at once, through the integrated Ph.D. program, in electrical and electronic engineering, Yonsei University. He worked as a Researcher in Yonsei University from September 2019 to November 2020, and as a Research Professor at Y-BASE R\&E Institute of Yonsei University from December 2020 to February 2022. He has been an assistant professor in the Division of Artificial Intelligence Engineering at Sookmyung Women's University, Seoul, since March 2022. His research interests include computer graphics, computer vision, and image processing.