A Greedy Pursuit Approach for Fitting 3D Facial Expression Models

JIWOO KANG AND SANGHOON LEE, (Senior Member, IEEE)
Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
Corresponding author: Sanghoon Lee (slee@yonsei.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2020R1A2C3011697.

ABSTRACT We propose a novel fitting strategy for the expression of blendshapes. Rather than employing all of the expression blendshapes to approximate the target points, only a subset of blendshapes selected to represent an expression on the target face is utilized, which efficiently reduces redundancy among the expression models. An expression correlation map is proposed to measure the redundancies between the blendshapes under the assumption that each expression changes the facial shape regionally, which enables a few less-correlated expressions to be obtained using a greedy pursuit approach. It is demonstrated that a subset of blendshapes that represents the target more expressively and semantically can be obtained non-parametrically using the proposed selection method, which enables natural facial shapes to be reliably generated without regularization, while also coping well with target-specific or unusual expressions. The experimental results from public datasets exhibit an increase in the quality of the facial shapes and expressions over baseline methods and state-of-the-art facial fitting approaches.

INDEX TERMS 3D model, expression fitting, blendshapes, greedy pursuit, expression similarity.

I. INTRODUCTION

Blendshapes, a linear model for facial expressions, has become a standard approach for generating 3D facial models [1], with the blendshape models used in recent works [2]–[6] typically utilizing the facial action coding system (FACS) [7]. Because of its simplicity and expressiveness, the blendshape model has been widely used in various applications [2], [4], [8]–[11] for face editing [12], facial pose estimation [13], facial feature extraction [14], and face classification [15].

There are two common types of representation used to linearly combine expression models to produce a target facial shape: global and delta blendshapes [1], [16]. In the global approach, the target shape is represented by positively weighting each facial blendshape and then linearly combining their full shapes. Although the global blendshape approach has the advantage of producing stable shapes while preventing the generation of unexpected expressions, it cannot correctly generate a face that has more than two expressions concurrently if the corresponding model is not included in the expression models.

The delta or local blendshape approach adds flexibility by using differences in the shape when compared to a neutral expression model rather than using whole facial shapes. Thus, a face with both eyes closed can be represented by adding two delta faces with their left and right eyes closed to the neutral expression face. While this flexibility makes it possible to approximate expressions much more precisely, stable representation is no longer guaranteed. This drawback occurs because delta expression models are not orthogonal to (or independent of) each other. In other words, the activation of one expression model may lead to a change in the facial shape generated by activating another expression model.

This leads to two main problems with the delta expression approach. First, too many personal factors can be reflected in the facial shape generated when fitting the model. When fitting expression models to a target face, it is assumed that these models have the same identity as the target and that shape deformation stems merely from the target expression. In practice, this assumption is not always correct. Therefore, the result is biased toward personal shapes rather than expressions, and its application can be somewhat limited. In addition, personal bias reduces the interpretability of a blendshape model. Because the contributions of each expression model can be obtained while being normalized during blendshape
fitting, the expression of the target facial shape can be interpretable and controllable to some extent, which is the primary advantage of blendshape representation. However, if personal factors are involved in the fitting, they act like noise, thus it is difficult to interpret the blendshape expressions numerically. Second, artifacts can make the facial shape generated by fitting look different from the target face. Unlike global blendshapes, because the delta blendshape approach changes the facial shape locally, the validity of the resulting facial shape cannot always be ensured. In other words, it is not guaranteed that the facial shape lies within the range of the expression blendshapes, which often causes over-fitting and unstable shapes during the facial-fitting process.

Efforts have been made to overcome the issues associated with the delta blendshape approach. \( L_1 \) and \( L_2 \) regularization have been used in several studies [8], [17]–[19] to prevent the overfitting of the expression models to the target face. Although norm-based regularization is a useful technique for preventing unexpected facial shapes and to make the weights of the blendshapes rather sparse, they can also lead to the facial shape being less expressive than the target face. Other methods have employed prior constraints for blendshape manipulation [20]–[24]. While these approaches have performed well, their quality strongly depends on the constraints, which are selected either manually or by using sufficient training examples.

In this paper, a novel expression fitting strategy for blendshape models is proposed. This method finds a set of expressions that are less correlated with each other rather than merely minimizing the fitting error. To identify the target expressions, we propose an expression correlation map that measures redundancy among the expression models. Using the correlation map, the degree to which the activation of one expression model influences the activation of other expressions is calculated, and a few less-correlated expressions that adequately represent the target face from among the candidates are chosen (Fig. 1).

By fitting the selected expressions to the target face, the accuracy of facial fitting is considerably higher than the baseline methods that optimize the facial coefficients of all expressions and state-of-the-art methods that use complete RGB information to reconstruct the facial shape of the target. In addition, by utilizing a sparse set of delta blendshapes that are less redundant in relation to each other for fitting, expressive and reliable facial shapes can be generated without regularization.

The main contributions of this paper can be summarized as follows:

- The fitting method utilizes selective expressions to semantically generate a facial shape while avoiding expression blendshapes that will be canceled out by others. We validate the performance of this strategy by comparing it with the baseline and state-of-the-art methods using the public datasets;
- The expression selection method sequentially selects expressions that are similar to the target while considering the redundancy of the chosen expressions, which ensures more robust and stable fitting optimization;
- The expression correlation map measures the relative redundancy among the expression models under the assumption that each expression changes the facial shape regionally.

**FIGURE 1.** An example of the proposed expression selection method. Once the magnitude of each expression is measured, the proposed method selects an expression most similar to the target and deducts its similarities by measuring the positive (colored in red) and negative (blue) redundancies from the remaining candidates. Using the selection method, a subset of blendshape expressions and the approximate values for the corresponding coefficients can be obtained. A fitting optimization for these selected expressions instead of using an entire set of the blendshapes can lead to a more accurate and stable representation of the target facial shape.
II. FITTING EXPRESSION BLENDSHAPES TO AN IMAGE

Expression blendshape weights are approximated by minimizing the distance between the facial landmark points of an image and the projected points of the corresponding landmarks on the 3D facial blendshape [2], [21]. Thus, before fitting the blendshape to the face on an image, the facial landmark points need to be labeled in advance. Seventy-four landmark positions on the face, consisting of 59 internal and 15 contour landmarks based on Labeled Faces in the Wild (LFW) [25], are automatically located by employing the supervised descent method (SDM) [26] in our implementation.

Under the assumption of an ideal pinhole camera, a perspective projection operator is defined for convenience that transforms 3D position \( p_m \) in model coordinates to 2D position \( p_l \) in image coordinates using \( p_l = \Pi_Q(p_m) \). Facial blendshape mesh \( B \) can be represented as a linear combination of \( n_b \) blendshape expression models \( B = [B_1, \ldots, B_{n_b}] \) using \( B = B_e^T = \sum_{i=1}^{n_b} e_i B_i \) where \( e = [e_1, \ldots, e_{n_b}] \) is a vector of expression coefficients [1], [24]. The landmarks corresponding to vertices are labeled manually in a neutral face model. Because the vertex order of every expression model is aligned with each other, the corresponding points in the other models can be determined immediately. To employ rotation \( R \) and translation \( t \) on the facial model \( B \), facial mesh \( F \) can be represented as \( F(e; R, t) = R(B_e^T) + t \). Let \( \{k_e\} (1 \leq k \leq n_l) \) denote the 2D facial landmarks on a given image, where \( n_l \) is the number of landmarks. Fitting the expression blendshapes to the image can be achieved by minimizing:

\[
E_{\text{fit}}(e; R, t) = \sum_{k=1}^{n_l} \left\| \Pi_Q \left(F(e; R, t)^{\{k_e\}}\right) - k_e \right\|^2
\]

where \( \{k_e\} \) is the corresponding vertex index for the \( k^{th} \) facial landmark on the mesh. Similar to [27], the energy is minimized using the coordinate-descent method by alternately finding \( R, t, \) and \( e \). When determining the optimum \( e \), the proposed method employs particular expression models rather than using all expression blendshapes, which distinguishes the proposed method from other blendshape fitting methods. Let \( s = \{s_1, s_2, \ldots, s_{n_s}\} \) denote an index set of particular expression models composed of \( n_s \) expression indices that will be used to fit \( (n_s \ll n_e) \). Thus, the expression fitting is formulated as

\[
e^* = \arg \min_{e_s} E_{\text{fit}}(e_{s1}, \ldots, e_{s_{n_b}})
\]

where \( e_s = \{e_{s1}, \ldots, e_{s_{n_b}}\} \) is a set of the expression coefficients composed of the selected expressions.

III. GREEDY PURSUIT SELECTIVE FITTING

A. GREEDY PURSUIT SELECTION OF EXPRESSIONS

We make an effort to selectively identify a subset of unit expressions that compose a facial shape of the target. This is achieved by initially activating each model alone and subsequently choosing the expressions in an order that minimizes the fitting error in (1). In a sense, this is no different to previous approaches that utilize delta blendshapes because the expression models are correlated with each other. Thus, ensuring that the selected expression models are less correlated with each other is an essential step toward obtaining a natural shape.

A correlation map is introduced to measure the relative redundancy between the expression blendshapes. Using this map, it is possible to avoid choosing expressions that are highly correlated with previously selected expressions. As a result, the selected number of facial expressions \( n_e \) is much lower than the total number of expressions \( n_e \). Once the subset of expressions is selected, the proposed method fits the expression model to the target in the same way with the previous blendshape fitting methods except that the proposed approach uses the subset of expression blendshapes during the optimization in (2) rather than the entire set of blendshapes. By utilizing the selected expressions, which are less correlated with each other, the blendshape fitting process is more stable and interpretable. Our method can also be easily applied to existing fitting methods by selecting a subset of expressions for the fitting process. In the following subsections, we introduce the expression correlation map, and then, the algorithm used to obtain the selective expression index \( s_e \) for the fitting process is explained in detail.

B. EXPRESSION CORRELATION MAP

Expression correlation is exploited to measure the relative redundancy of an expression model in comparison to others. In this process, both positive and negative redundancies are considered. In the delta blendshape approach, each expression model differs from the neutral shape, which can be thought of as a motion (or a change in the facial shape). Positive redundancy means that the motions of two expressions have a positive correlation, while the opposite is the case for negative redundancy. Thus, once an expression is chosen, the other expressions that are much positive-redundant to the chosen expression are not supposed to be used to represent the target expression. Selecting an expression with strongly positive redundancy can exaggerate the facial shape when compared with the original expression. Conversely, negative redundancy needs to be decreased to generate a stable facial shape with sparse coefficients.

From a redundancy perspective, when fitting whole blendshapes, the activation of an expression locally cancels out the facial shape that is generated by expressions that have already been activated during optimization. The canceled shape is recovered by the activation of another expression. Finally, the facial shape is obtained when the fitting energy is lowest. Although cancelation reduces the fitting energy,
it dramatically increases the negative redundancy between expressions. This back-and-forth process is a significant reason for the lower sparsity of the expression coefficients; i.e., the number of expressions required to represent the target is higher.

Because the fitting energy is calculated based on sparse facial landmarks in the image, the number of points is insufficient to cover the entire facial region in detail. Minimizing the fitting energy does not guarantee that the generated facial shape accurately reflects the target expression. Rather, it causes the facial shape to be overfitted and unstable. Therefore, ensuring that the expressions are sparse and less correlated by minimizing redundancy can help to obtain a stable facial shape by preserving the semantic meaning of the expressions. Examples of positive and negative redundancies are presented in Fig. 2. The two expressions $s_1$ and $t_1$ in Figs. 2a and 2b are different, but they both have motions for the opening of the mouth. Because $t_1$ has highly positive redundancy with $s_1$ around the mouth, the activation of the positively redundant expression $s_1$ exaggerates the original expression $t_1$, as shown in Fig. 2c. Thus, positively redundant expressions are treated as mutually exclusive in the selection process as much as possible to increase the stability and interpretability of the facial shape. On the other hand, expression $s_2$ in Fig. 2d has wide-open eyes, whereas expression $t_2$ in Fig. 2e has half-closed eyes. Because these represent counter-wise motions, $t_2$ is negatively redundant in relation to $s_2$, as indicated by the blue vertices in Fig. 2e. It is difficult to recognize the expressions in $s_1$ and $t_2$ in Fig. 2f, although both of the expressions are activated due to the cancelation caused by the negative redundancy. This cancelation generates a ‘fake’ expression, which increases the number of fitted expressions and leads to the overfitting of the blendshapes to the target.

Let $B_s$ and $B_t$ be any two expression models for delta blendshapes. To measure the redundancy of $B_s$ and $B_t$, the correlation map is defined as

$$C(B_s, B_t) = \frac{\sum_{i} |\Delta V_i^s \cdot \Delta V_i^t|}{\sqrt{\sum_{i} |\Delta V_i^s|^2} \sqrt{\sum_{i} |\Delta V_i^t|^2}}$$

where $n_v$ is the number of vertices, and $\Delta V_i^s$ and $\Delta V_i^t$ are the differences in the $i^{th}$ vertex of expression models $B_s$ and $B_t$ from the neutral facial shape, respectively. The formulation $(\Delta V_i^s \cdot \Delta V_i^t)$ in (3) calculates the redundancy between the vertex motion pair $\Delta V_i^s$ and $\Delta V_i^t$. If this is positive, it represents positive redundancy, and if it is negative, it represents negative redundancy. The expression correlation is calculated by accumulating the magnitude of both types of redundancy. The correlation is then normalized by the source model vertices to obtain the relative correlation value. The correlation map is not symmetrical, i.e., $C(B_i, B_j) \neq C(B_j, B_i)$ for $i \neq j$. In our implementation, the correlation of each delta blendshape pair is calculated in advance, i.e., $C(B_i, B_j)$ for $i, j \in (2, \ldots, n_v)$. An example of the expression correlation map we use is shown in Fig. 4. Figure 3 presents the positive and negative redundancies and shows the expression correlation according to the change in expression. $C(B_s, B_t,i)$ quantifies the relative expressional redundancy of $B_t$ in relation to the expression of $B_s$.

Let $B_s$ and $B_t$ be any two expression models for delta blendshapes. To measure the redundancy of $B_s$ and $B_t$, the correlation map is defined as

$$C(B_s, B_t) = \frac{\sum_{i} |\Delta V_i^s \cdot \Delta V_i^t|}{\sqrt{\sum_{i} |\Delta V_i^s|^2} \sqrt{\sum_{i} |\Delta V_i^t|^2}}$$

where $n_v$ is the number of vertices, and $\Delta V_i^s$ and $\Delta V_i^t$ are the differences in the $i^{th}$ vertex of expression models $B_s$ and $B_t$ from the neutral facial shape, respectively. The formulation $(\Delta V_i^s \cdot \Delta V_i^t)$ in (3) calculates the redundancy between the vertex motion pair $\Delta V_i^s$ and $\Delta V_i^t$. If this is positive, it represents positive redundancy, and if it is negative, it represents negative redundancy. The expression correlation is calculated by accumulating the magnitude of both types of redundancy. The correlation is then normalized by the source model vertices to obtain the relative correlation value. The correlation map is not symmetrical, i.e., $C(B_i, B_j) \neq C(B_j, B_i)$ for $i \neq j$. In our implementation, the correlation of each delta blendshape pair is calculated in advance, i.e., $C(B_i, B_j)$ for $i, j \in (2, \ldots, n_v)$. An example of the expression correlation map we use is shown in Fig. 4. Figure 3 presents the positive and negative redundancies and shows the expression correlation according to the change in expression. $C(B_s, B_t,i)$ quantifies the relative expressional redundancy of $B_t$ in relation to the expression of $B_s$.
The initial value $t$ after the repeated a maximum of coefficient value is non-zero. Thus, the selection process is executed that all of the expressions are uncorrelated. In the selection step, a few less-correlated expressions are selected by taking the redundancies between the expression blendshapes into account based on correlation map $C$.

Let $e_i$ for $2 \leq i \leq n_i$ be the coefficient of the $i^{th}$ expression after the $t^{th}$ expression selection. In the observation step, the initial value $e_i^0$ is determined by optimizing the fitting error in (1) with all of the other expressions deactivated except the $i^{th}$ expression. $E_i$ is the fitting error corresponding to the expression obtained via optimization. In the selection step, the $i^{th}$ expression that has the minimal $E_i$ and the corresponding coefficient value $e_i^{t-1}$ is selected for the $t^{th}$ iteration ($t \geq 0$) from among the expressions that satisfy $e_i^{t-1} > 0$. Once an expression is chosen, the coefficient is updated to $e_i^t$ based on the correlation map $C$ in (3) for the selection in the following iteration. In detail, in the observation step, the initial coefficient value is obtained for $i \in \{2, \ldots, n_e\}$,

$$e_i^0 = \arg \min_{e_i} E_{\text{fit}}(e_i),$$

$$E_i = E_{\text{fit}}(e_i^0).$$

Note that the optimization in (4) is initialized by deactivating the other expressions except for one for optimization, i.e., $e_j = 0$ for $j \in \{2, \ldots, n_e\}$ and $j \neq i$. In the selection step, selecting an expression and calculating the redundancy of the selected expression in relation to the other expressions are conducted. The expression (index) $s_t \in s$ selected in the $t^{th}$ iteration ($t \geq 0$) and its predicted coefficient $e_{s_t}$ can be determined as follows:

$$s_t = \arg \min_{e_i} E_i,$$

$$e_{s_t} = e_{s_t}^{t-1}.$$  \hspace{1cm} (6)

The minimization in (6) is conducted only for valid expressions, i.e., for $\{i|e_i^{t-1} > 0\}$. The relative redundancy of the remaining blendshapes in relation to the selected expression is measured by scaling the correlation $C$ in (3) by the predicted magnitude. The redundancy is deducted from the magnitude of the expression coefficients by reflecting the correlation of the selected expression with the following selection:

$$e_i^t = \max\left(e_i^{t-1} - (e_{s_t} \cdot C(B_i, B_{s_t})), 0.0\right).$$

The algorithm for the selection method is summarized in Algorithm 1. To help understand our approach, we provide an example of the selection step, in which four blendshape models $B_i$ ($i \in \{2, 3, 4, 5\}$) are fitted to the target. In the observation step, initial coefficient value $e_i^0$ for minimal error $E_i$ for each model is determined using (4) and (5). In the first iteration of the selection step, blendshape $B_3$, which has the lowest $E_i$ among the expressions satisfying $e_i^0 > 0$, is chosen, i.e., $s_1 := 3$ (line 04). The coefficient value of the selected expression is predicted to be the one obtained in the observation step ($e_{s_1} = 0.9$) using (7). Using the correlation map $C(B_3, B_{s_1})$, the relative correlation of the $i^{th}$ expression with the selected expression $s_1$ is calculated. For example, the correlation of $B_2$ with the selected expression blendshape $B_3$ is $C(2, s_1) = 0.5$. The redundancy in relation to the selected expression thus becomes $0.5 \times 0.9 = 0.45$. The calculated redundancy is deducted from the coefficient.
Algorithm 1 Expression Selection Method

**Input:** the fitting energy $E_i$, and expression coefficient value $e_i^0$ for each expression index $i \in (2, \ldots, n_e)$ obtained in the observation step using Eqs. (4) and (5), respectively.

**Output:** the number of selected expressions $n_s$, a sequence of the selected expression indices $s = \{s_1, s_2, \ldots, s_{n_s}\}$, and a set of expression coefficients consisting of the selected expressions $e_s = \{e_{s_1}, \ldots, e_{s_{n_s}}\}$.

**Algorithm 1: Expression Selection Method**

```plaintext
01 $t := 0$
02 while $\{\max_i e_i^t > 0 \mid i \in (2, \ldots, n_e)\}$ do
03 $t := t + 1$
04 $s_t := \arg\min_{i \in (2, \ldots, n_e)} E_i$
05 $e_s := e_s \cup \{e_{s_t}\}$
06 for $i \in (2, \ldots, n_e)$ do
07 $e_i^t := e_i^t - e_{s_t} \cdot C(i, s_t)$ in (3)
08 end do
09 for $i \in (2, \ldots, n_e) \mid e_i^t \leq 0$ do
10 $E_i := \inf$
11 end do
12 $n_s := k$
end do
```

value of $B_2$ (line 08). The coefficient value of a fully correlated expression such as $B_2$ becomes 0. Thus, it is not selected in the following iterations, and $B_5$, which has an expression related to the eyebrows, is selected as the second choice (line 12).

IV. EXPERIMENTAL RESULTS

We validated the performance of the proposed greedy pursuit approach for expression blendshapes by comparing its results with those from state-of-the-art methods that employ parametric facial models for face fitting. We then compared our method to others that have been previously used for delta blendshapes. Following the related works [28], [29], the inter-pupil normalized mean error (NME) was used to measure the distance between the ground-truth facial landmarks for an image and the obtained landmarks.

A. COMPARISONS TO STATE-OF-THE-ART APPROACHES

1) METHOD AND DATASET

We compared our method with the state-of-the-art fitting methods: two convolutional neural network (CNN)-based methods proposed by Zhu et al. [30] and Chang et al. [31] and two optimization-based methods proposed by Cao et al. [4] and Booth et al. [17]. The method presented by Cao et al. [4] used the expression delta blendshape models without decomposition to obtain the coefficients of the expression blendshapes, whereas the other methods utilized a principal components analysis (PCA)-decomposed 3D Morphable Model (3DMM) to reconstruct the facial shapes [32] and expressions [33] of the target in each image. The identity and expression models used for the training variations of the 3DMM used in our experiments were the Basel Face Model [34] and FaceWarehouse [3], respectively. For our method, we used FaceWarehouse to construct a bilinear facial model, which is required to obtain a user-specific expression blendshape set [21]. The proposed approach and Cao’s method only used facial landmarks as an input. On the other hand, other comparison methods used the full information provided by the RGB images. Because the proposed approach and the methods introduced by Booth et al. and Cao et al. need landmarks to obtain the facial shape of the target, they used the landmark points obtained using a state-of-the-art landmark detector [26], [35] instead of the ground-truth landmarks. The methods presented by Zhu et al. and Chang et al. require a bounding box for the face in an image, which was generated using the Viola-Jones face detector [36].

We employed the 300W [37] and Extended Cohn-Kanade (CK+) [38] public datasets for the experiments. Both datasets share the same annotations for 68 landmarks. Of these landmarks, two inner lip corners and 17 facial boundary points can be ambiguous according to the facial pose and expression, so 49 landmarks were used as input for our method and to measure the NME for the performance comparison (Fig. 8). The 300W [37] dataset aligns the multiple facial databases with the same annotations, including AFW [39], LFPW [40], HELEN [41] and IBUG [42]. The methods proposed by Zhu et al. and Chang et al. require a bounding box for the face, thus only images whose facial region had been detected were employed in the experiments. As a result, 270 of 337 images in AFW, 1005 of 1035 images in LFPW, and 40 of 68 images in IBUG were used to train the landmark detector, and 2150 of 2330 images in HELEN and 585 of 593 images in CK+ datasets were used to test the performance of the methods. The HELEN, AFW, LFPW, and IBUG datasets were used to test the Zhu et al. [30] and Booth et al. [17] models. However, using the facial regions and landmarks estimated using detectors instead of using the ground-truths for the evaluation and additional measurements of the CK+ dataset were thought to be sufficient to obtain results for a fair comparison. For Booth et al.’s method, we evaluated performance by employing the 68 landmark sets because performance degradation could be observed when using only 49 inner landmarks.

2) RESULTS AND DISCUSSION

The results obtained from the methods on the HELEN and CK+ datasets are summarized in Table 1. The NME of the SDM used to detect facial landmarks for the inputs for the proposed method, Cao16, and Booth17 were also measured as a baseline. Our method obtained the target expression by fitting the blendshape to the landmarks estimated by the landmark detector and had a lower mean error than did the SDM despite not having any other information outside of the landmarks for the target. It was demonstrated that the
TABLE 1. NME of 49 inner landmarks on the faces in the HELEN and CK+ datasets.

| Dataset | Method | Ours | SDM | Cao16 | Zhu16 | Booth17 | Chang18 |
|---------|--------|------|-----|-------|-------|---------|---------|
| HELEN   | 0.068  | 0.080| 0.096| 0.135 | 0.234 | 0.118   |         |
| CK+     | 0.078  | 0.105| 0.101| 0.126 | 0.197 | 0.104   |         |
| Mean    | 0.070  | 0.085| 0.097| 0.133 | 0.226 | 0.115   |         |
| ± Std.  | 0.021  | 0.036| 0.030| 0.041 | 0.034 | 0.030   |         |

positions of the landmarks were positioned closer to the target by selectively utilizing the blendshape expressions. When compared with the performance of Cao16, where all expressions of the delta blendshapes were used, the result indicates that the increased semantic meaning of the expression blendshapes by the selective utilization helps to better represent the facial shape and expression. On the other hand, fitting the facial model using the points from the landmark detector greatly decreased the performance of Booth17 when compared with facial shapes obtained using the ground-truth landmarks in [17]. The methods proposed by Zhu et al. and Chang et al. successfully identified the target expressions on a coarse scale in most cases. However, they failed to produce detailed expressions for the target. The target expressions produced by the method proposed by Cao et al. from given facial landmarks were insufficient to accurately recover the expression shapes of the target faces. This is clearly illustrated in Figs. 6 and 7, which visualize the facial models fitted using the HELEN and CK+ datasets, respectively.

The results show that detailed expression shapes and poses can be captured for the target using only the inner landmarks with our proposed method. It produces more expressive 3D facial shapes of the target in a non-parametric manner for the given facial points. However, because the proposed method is inherently affected by the landmark detector, facial shapes that have expressions that differ from the target were obtained when completely misaligned landmarks are provided by from the detector, as shown in Fig. 8. Although we focus on the selective fitting process in this work, it could be resolved by extending the fitting procedure to the selection of optimal landmarks for the fitting of an image.

B. FITTING APPROACHES FOR EXPRESSION BLENDSHAPES

1) METHOD AND DATASET

We conducted experiments to verify the performance of the proposed scheme for the selection of expression blendshapes compared to previously used methods. In many previous approaches [2], [18], [21], [43]–[45], the coefficients of the expression blendshapes can be obtained by minimizing the fitting error $E_{\text{fit}}$ in (1) under the assumption that user-specific blendshapes are available. Unlike 3DMM-based methods, these methods do not employ low-dimensional subspace decomposition for the expression blendshapes, such as PCA. Instead, the original shape of each expression blendshape is intended to be preserved as much as possible. Thus, the $L_1$ or $L_2$ regularization term for the expression coefficients was adopted to prevent the facial shape from being overfitted and exaggerated expressions from being generated. In short, the fitting energy of these methods can be formulated as $E = E_{\text{fit}} + \omega_1 \| \epsilon \|_1 + \omega_2 \| \epsilon \|_2$. Note that
\( L_1 \) regularization and \( L_2 \) regularization were employed separately in our experiments for reliable measurement, i.e., either \( \omega_1 \) or \( \omega_2 \) was 0. For simplicity, the results obtained from \( L_1 \) and \( L_2 \) regularization are denoted as \( L_1\text{Reg} \) and \( L_2\text{Reg} \), respectively.

In addition, performance comparisons were conducted with fitting schemes using PCA-decomposed basis vectors, which have been employed in many previous parametric facial models [17], [30], [31], [46]–[48]. In our experiments, two types of PCA-decomposed bases were employed: one obtained by decomposing the whole expression set, including the neutral model, and the second decomposed using the differences in the expression models from the neutral face model [33], which has been widely used in PCA-based facial fitting approaches. For simplicity, the results obtained from these two strategies are denoted as \( \text{PCAExp} \) and \( \text{PCADel} \), respectively.

To reliably assess the expression fitting performance, the expressional shapes were obtained using the same identity shape of the blendshapes for each image. We fitted the model to the neutral face of the target to obtain the identity shape for the target, allowing a user-specific expression blendshape set to be obtained [21]. The proposed method, \( L_1\text{Reg} \) and \( L_2\text{Reg} \) used the user-specific blendshape set directly, whereas \( \text{PCAExp} \) and \( \text{PCADel} \) decomposed the set into PCA space. The regularization term is necessary for PCA-based fitting. Following the standard approach [31], [47], [48], \( L_2 \) regularization to PCA coefficients weighted with PCA eigenvalues was employed. The regularization weights used for \( \text{PCAExp} \) and \( \text{PCADel} \) were denoted as \( \omega_3 \) and \( \omega_4 \), respectively. We employed the FaceWarehouse [3] facial database, consisting of 47 expressions for 150 individuals, as a bilinear facial model for all methods. Because a neutral face for each image was required to generate the user-specific blendshapes, we employed the CK+ dataset [38], in which each sequence begins with a neutral expression and ends with the peak expression. For the 593 sequences in the CK+ dataset, images of the peak expression were used to evaluate model performance. For reliable measurement, we carefully selected 23 points from among the 49 inner landmarks in the CK+ dataset (Fig. 9). All of the methods fitted the facial shapes to the selected points and were evaluated using the NME for all (49pts), unused (26pts), and used (23pts) landmarks.

Unlike the proposed method, which is non-parametric, the other methods need to determine the regularization weights. We used the optimal weight for each method by varying the weights and selecting the one that decreased the error most for all landmarks. The weights used in the experiments were \( \omega_1 = 0.001 \), \( \omega_2 = 0.001 \), \( \omega_3 = 0.005 \), and \( \omega_4 = 0.005 \) for \( L_1\text{Reg} \), \( L_2\text{Reg} \), \( \text{PCAExp} \), and \( \text{PCADel} \), respectively.

2) RESULTS AND DISCUSSION

The results for the expression-fitting methods using the CK+ dataset are summarized in Table 2. The differences in the NME were not significant for the used landmarks. However, the proposed method best predicted the unused landmarks, whereas all the previous approaches failed to accurately predict the unused landmarks, which led to a higher NME for all landmarks. It is clear that a larger regularization weight helps to decrease the error for unused landmarks, but it increases the mean error for used landmarks as a side-effect. Of the previous approaches, \( L_1\text{Reg} \) and \( \text{PCADel} \) exhibited the lowest NME. Figure 9 visualizes the facial shapes obtained from these methods. \( L_1\text{Reg} \) and \( L_2\text{Reg} \) produced similar facial shapes, while \( \text{PCAExp} \) and \( \text{PCADel} \) were also similar to each other. In many cases, the greedy pursuit approach enables our method to produce expressions that are visually unique when compared with the other methods. Indeed, by selecting suitable expressions from the expression blendshape set and by scaling them to fit the target, the proposed method can represent the target shape expressively and semantically. On the other hand, the previous methods combine all of the delta expressions that are similar to the given points while preventing too many expression activations without knowing the proper number of expressions to be matched to the target expression. This leads to a large number of redundant expressions that are dissimilar from the expression of the target.

FIGURE 9. Facial models obtained by fitting 23 of 49 inner landmarks on faces in the CK+ database. The fitted points are marked in red.

| TABLE 2. NME for All (49pts), Unused (26pts) and Used (23pts) landmarks in the CK+ dataset. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
| Ours | \( L_1\text{Reg} \) | \( L_2\text{Reg} \) | \( \text{PCAExp} \) | \( \text{PCADel} \) |
|------|------------------|------------------|-----------------|-----------------|
| All  | 0.048            | 0.056            | 0.058           | 0.056           |
| Unused | 0.049            | 0.060            | 0.062           | 0.058           | 0.061 |
| Used | 0.048            | 0.050            | 0.054           | 0.051           | 0.050 |

Figures 10 and 11 present the results of each method for various regularization weights. In our experiments, employing different regularization weights did not produce a distinct improvement in the expressional shape. A larger weight forces the expressions of the facial shape to decrease overall.
PCA decomposition changes the basis vector globally and helps to represent the expression shape of the course detail. However, the fitting process becomes more unstable when representing the shape at a fine scale. The results demonstrate that overall facial shapes are expressively and semantically obtained by utilizing a suitable number of expression blendshapes in a non-parametric manner for the given points using the proposed method.

V. DISCUSSION AND CONCLUSION
In our method, the subset of expressions is determined using a greedy pursuit approach, which iteratively selects the expression that decreases the fitting error the most from among the remaining expressions that are less redundant than the previously selected expressions. The initial weights of the blendshapes are measured first. For each selection, the weights of the remaining expressions are deducted in proportion to the sum of the positive and the negative redundancies in relation to the selected expression using Eq. (8). Because the weights of the redundant expressions decrease, strongly redundant expressions become non-positive and thus will not be selected in the subsequent cycles following Eqs. (4) and (5).

The positive and negative redundancies in this paper can be understood as the cosine similarity between delta blendshapes. The proposed method attempts to obtain orthogonal delta blendshapes by choosing expressions that are dissimilar to previously selected expressions. A similar approach was employed in EBGAN [49], where cosine similarity was used as a penalty to obtain orthogonal parameters. In most previous fitting methods, the optimization of the transformations $R$ and $t$ and the optimization of expression weight $e$ are conducted alternately (or jointly) until the variables converge. The proposed method follows the same process except that a subset of the expressions is selected before the iterations begin. In previous methods, a non-frontal face tends to become trapped in local minima because $e$ is determined before $R$ and $t$ are sufficiently optimized. However, we observed that the proposed method is robust to the local minimum problem and thus effectively represents the target face in most cases.

There are a number of explanations for these results. First, greedy selection leads to the selection of fewer expressions. Therefore, our method prefers to represent a local expression by combining one or two expression models rather than that of more than three models. It enables our method to robustly estimate the facial pose $R$ and $t$. $L_0$-norm regularization works similarly; however, it does not consider the relevance between the blendshape expressions, resulting in unstable expressions. Second, the proposed method sequentially increases the number of expressions as $R$ and $t$ become more closely fitted to the target. Very few expressions are used in the initial iterations of the fitting optimization. This significantly decreases the possibility of becoming stuck in local minima through hierarchical optimization.

In contrast to many recent studies that use parametric facial models whose expression basis is decomposed into low-dimensional space using PCA, the proposed method directly manipulates 3D facial rigs to represent the target. It enables our method to represent the facial shape semantically while also being easily utilized for the 3D facial rigging of characters in computer games and 3D animations.

The application of the proposed method could be further extended to emotion recognition and action unit detection [50], [51]. In addition, because our method identifies a few less-correlated expressions, it could be utilized in keyframe extraction [52] or in facial video interpolation [53]. It could also be used to capture facial performances [4] by training machine learning models and using the obtained coefficients as the ground-truth values. We are looking to advance our work in these directions in the future.
[46] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, “Learning a model of facial shape and expression from 4D scans,” ACM Trans. Graph., vol. 36, no. 6, pp. 1–17, Nov. 2017.

[47] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, “Learning detailed face reconstruction from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 5553–5562.

[48] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2387–2395.

[49] J. Zhao, M. Mathieu, and Y. LeCun, “Energy-based generative adversarial network,” in Proc. Int. Conf. Learn. Represent. (ICLR), San Juan, Puerto Rico, May 2016, pp. 1–17.

[50] W.-S. Chu, F. De la Torre, and J. F. Cohn, “Selective transfer machine for personalized facial action unit detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2013, pp. 3515–3522.

[51] J. Kang, S. Lee: Greedy Pursuit Approach for Fitting 3D Facial Expression Models

JIWOO KANG was born in South Korea, in 1987. He received the B.S. degree, in 2011, and the M.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2019. He is currently holding a postdoctoral position in electrical and electronic engineering with the Multidimensional Insight Laboratory, Yonsei University. His research interests include computer graphics, computer vision, and image analysis.

SANGHOON LEE (Senior Member, IEEE) received the B.S. degree in E.E. from Yonsei University, Seoul, South Korea, in 1989, the M.S. degree in E.E. from KAIST, in 1991, and the Ph.D. degree in E.E. from the University of Texas at Austin, in 2000. From 1991 to 1996, he worked with Korea Telecom. From 1999 to 2002, he worked with Lucent Technologies. He joined the Faculty of the Department of Electrical and Electronics Engineering, Yonsei University, in March 2003, where he is currently a Full Professor. His research interests include deep learning, image and video quality of experience, computer vision, and computer graphics. From 2014 to 2019, he was a member in the Technical Committees of the IEEE IVMSP. Since 2016, he has been a member in the Technical Committees of the IEEE MMSP. He currently serves as a BOG member of APSIPA. He received the 2015 Yonsei Academic Award, the 2010, 2015, 2016, 2017 Yonsei Outstanding Accomplishment Awards, the 2012 Special Service Award from the IEEE Broadcast Technology Society, the 2013 Special Service Award from the IEEE Signal Processing Society, the Humantech Thesis Award of Samsung Electronics, in 2013, the IEEE Seoul Section Student Paper Contest Award, in 2012, the Qualcomm Innovation Award, Qualcomm, in 2012, and the Best Student Paper Award of QoMEX (International Conference on Quality of Multimedia Experience) 2018. Since 2011, he has been a Chair of the IEEE P3333.1 Quality Assessment Working Group. Since 2018, he has been a Chair of the APSIPA IVM Technical Committee. 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 and the Local Arrangements Chair of IEEE International Symposium On Broadband Multimedia Systems and Broadcasting (BMSB), in 2012. Since 2020, he has been the Vice EiC of the APSIPA newsletters. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, from 2010 to 2014, an Editor of the Journal of Communications and Networks (JCN), from 2009 to 2015, and a Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, in 2013, and the Journal of Electronic Imaging, in 2015. Since 2014, he has been an Associate/Senior Area Editor of the IEEE SIGNAL PROCESSING LETTERS.

** **