Facial Dynamics Interpreter Network: What Are the Important Relations Between Local Dynamics for Facial Trait Estimation?

Seong Tae Kim and Yong Man Ro[(✉)]

School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
{stkim4978,ymro}@kaist.ac.kr

Abstract. Human face analysis is an important task in computer vision. According to cognitive-psychological studies, facial dynamics could provide crucial cues for face analysis. The motion of a facial local region in facial expression is related to the motion of other facial local regions. In this paper, a novel deep learning approach, named facial dynamics interpreter network, has been proposed to interpret the important relations between local dynamics for estimating facial traits from expression sequence. The facial dynamics interpreter network is designed to be able to encode a relational importance, which is used for interpreting the relation between facial local dynamics and estimating facial traits. By comparative experiments, the effectiveness of the proposed method has been verified. The important relations between facial local dynamics are investigated by the proposed facial dynamics interpreter network in gender classification and age estimation. Moreover, experimental results show that the proposed method outperforms the state-of-the-art methods in gender classification and age estimation.

Keywords: Facial dynamics · Interpretable deep learning
Relation between local dynamics · Facial trait estimation

1 Introduction

Analysis of human face has been an important task in computer vision because it plays a major role in soft biometrics, and human-computer interaction [7,33]. Facial behavior is known to benefit perception of the identity [32,34]. In particular, facial dynamics play crucial roles for improving the accuracy of facial trait estimation such as age estimation or gender classification [6,9].

In recent progress of deep learning, convolutional neural networks (CNN) have shown outstanding performance on many fields of computer vision. Several research efforts have been devoted to developing spatio-temporal feature representation in various applications such as action recognition [13,21,23,40] and activity parsing [26,42]. In [23], a long short-term memory (LSTM) network has

been designed on top of CNN features to encode dynamics in video. The LSTM network is a variant of recurrent neural network (RNN), which is designed to capture long-term temporal information in sequential data [19]. By using the LSTM, the temporal correlation of CNN features was effectively encoded.

Recently, a few research efforts have been made regarding facial dynamic feature encoding for a facial analysis [6,9,24,25].It is generally known that the dynamic features of local regions are valuable for facial trait estimation [6,9]. Usually, the motion of facial local region in facial expression is related to the motion of other facial regions [39,43]. However, to the best of our knowledge, there are no studies that utilize relations between facial motions and interpret the important relations between local dynamics for facial trait estimation.

In this paper, a novel deep network has been proposed for interpreting relations between local dynamics in facial trait estimation. To interpret the relations between facial local dynamics, the proposed deep network consists of a facial local dynamic feature encoding network and a facial dynamics interpreter network. By the facial dynamics interpreter network, the importance of relations for estimating facial traits is encoded. The main contributions of this study are summarized in following three aspects:

1. We propose a novel deep network which estimates facial traits by using relations between facial local dynamics of smile expression.
2. The proposed deep network has been designed to be able to interpret the relations between local dynamics in facial trait estimation. For that purpose, the relational importance is devised. The relational importance is encoded from the relational features of facial local dynamics. The relational importance is used for interpretation of important relations in facial trait estimation.
3. To validate the effectiveness of the proposed method, comparative experiments have been conducted on two facial trait estimation problems (*i.e.* age estimation and gender classification). In the proposed method, the facial trait estimation is conducted by combining the relational features based on the relational importance. By exploiting the relational features and considering the importance of relations, the proposed method could more accurately estimate facial traits compared with the state-of-the-art methods.

2   Related Work

Age Estimation and Gender Classification. A lot of research efforts have been devoted to development of automatic age estimation and gender classification techniques from face image [2,4,4,16,22,28,29,29,38,38,41]. Recently, deep learning methods show notable potential in various face analysis tasks. One of the main focus of these methods is to design suitable deep network structure for some specific tasks. Parkhi et al. [31] reported VGG-style CNN learned from large-scale static face images. Deep learning based age estimation method and gender classification method have been reported but they were mostly designed on static face image [22,27,28,41].

Facial Dynamic Analysis. The temporal dynamics of face have been ignored in both age estimation and gender classification. Recent studies have reported that facial dynamics could be an important cue for facial trait estimation [6,8–10,17]. With aging, the face loses muscle tone and underlying fat tissue, which creates wrinkles, sunken eyes and increases crow's feet around the eyes [9]. Aging also affects facial dynamics along with appearance. As a human being gets older, the elastic fibers of the face show fraying. Therefore facial dynamic features of local facial regions are important cues for age estimation. In cognitive-psychological studies [1,5,18,36], evidence for gender-dimorphism in the human expression has been reported. Females express emotions more frequently compared with males. Males have a tendency to show restricted emotions and to be unwilling to self-disclose intimate feelings [6]. In [6], Dantcheva et al. used dynamic descriptors extracted from facial landmarks for gender classification. However, there are no studies for learning relations of dynamic features in facial trait estimation.

Relational Network. In this paper, we propose a novel deep learning architecture for analyzing relations of facial dynamic features in facial trait estimation. A relational network has been reported in visual question and answering (VQA) [35]. In [35], the authors defined an object as a neuron on feature map obtained from CNN and designed a neural network for relational reasoning. However, it was designed for image-based VQA. In this paper, the proposed method automatically encodes the importance of relations by considering the locational information on face. By utilizing the importance of relations, the proposed method could interpret the relations between facial dynamics in facial trait estimation.

3   Proposed Facial Dynamics Interpreter Network

Overall structure of the proposed facial dynamics interpreter network is shown in Fig. 1. The aim of the proposed method is to interpret the important relations between local dynamics in facial trait estimation from expression sequence. The proposed method largely consists of the facial local dynamic feature encoding network, the facial dynamics interpreter network, and interpretation on important relation between facial local dynamics. The details are described in the following subsections.

3.1   Facial Local Dynamic Feature Encoding Network

Given a face sequence, appearance features are computed by CNN on each frame. For the purpose of appearance feature extraction, we employ the VGG-face network [31] which is trained with large-scale face images. The pre-trained VGG face model is used to get off-the-shelf CNN features in this study. With given CNN features, the proposed facial dynamics interpreter network has been investigated. The output of convolutional layer in the VGG-face network is used as feature map of facial appearance representation.

Based on the feature map, the face is divided into N_0 local regions. The location of local regions was determined to interpret the relation of local dynamics
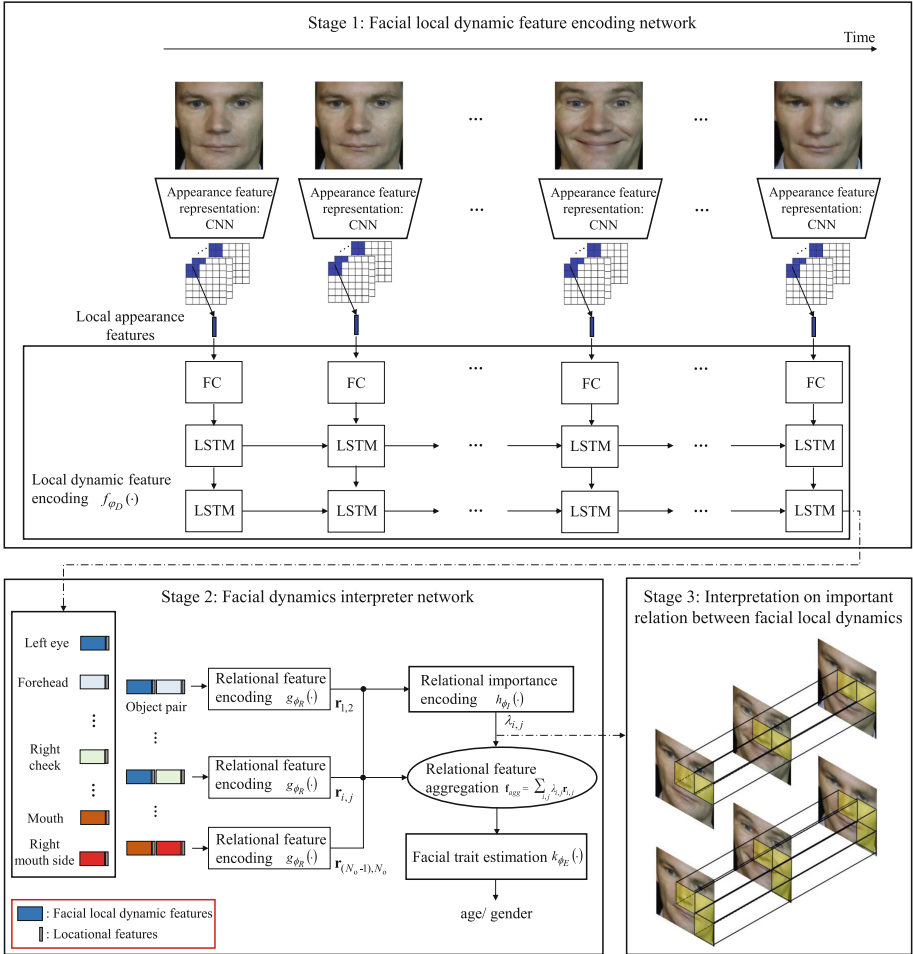
Fig. 1. Overall structure of the proposed facial dynamics interpreter network.

base on semantically meaningful facial local region (*i.e.* left eye, forehead, right eye, left cheek, nose, right cheek, left mouth side, mouth, and right mouth side in this study). Note that each face sequence is automatically aligned based on the landmark detection [3]. Let $\mathbf{x}_i^t$ denote the local appearance features of i-th facial local part at t-th time step. To encode local dynamic features, an LSTM network has been devised with fully-connected layer on top of the local appearance features $\mathbf{X}_i = \left\{ \mathbf{x}_i^1, \ldots, \mathbf{x}_i^t, \ldots, \mathbf{x}_i^T \right\}$ as followings:

$$\mathbf{d}_i = f_{\phi_D}(\mathbf{X}_i), \tag{1}$$

where $\mathbf{d}_i$ denotes the facial local dynamic feature of i-th local part and f_{ϕ_D} is a function with learnable parameters ϕ_D. f_{ϕ_D} consists of the fully-connected layer and the LSTM layers as shown in Fig. 1. T denotes the length of face sequence.

The LSTM network could deal with the different length of sequences. The various dynamic related features including variation of appearance, amplitude, speed, and acceleration could be encoded from the sequence of local appearance features. The detailed configuration of the network used in the experiments will be presented in Sect. 4.1.

3.2   Facial Dynamics Interpreter Network

We extract object features (*i.e.* facial local dynamic features and locational features) for pairs of objects. The locational features are defined as the central position of the object (*i.e.* facial local region). For the purpose of telling the location information of objects to the facial dynamics interpreter network, the local dynamic features and the locational features are embedded and defined as object features $\mathbf{o}_i$. The object feature can be written as

$$\mathbf{o}_i = [\mathbf{d}_i, p_i, q_i], \tag{2}$$

where $[p_i, q_i]$ denotes the normalized central position of i-th object.

The design philosophy of the proposed facial dynamics interpreter network is to make the functional form of a neural network which captures the core relations for facial trait estimation. The importance of the relation could be different for each pair of object features. The proposed facial dynamics interpreter network is designed to encode relational importance in facial trait estimation. The relational importance could be used for interpreting the relation between local dynamics in facial trait estimation.

Let $\lambda_{i,j}$ denote a relational importance between i-th and j-th object feature. The relational feature, which represents latent relation between two objects for facial trait estimation, can be written as

$$\mathbf{r}_{i,j} = g_{\phi_R}(\mathbf{s}_{i,j}), \tag{3}$$

where g_{ϕ_R} is a function with learnable parameters ϕ_R. $\mathbf{s}_{i,j} = (\mathbf{o}_i, \mathbf{o}_j)$ is relation pair from i-th and j-th facial local parts. $\mathbf{S} = \{\mathbf{s}_{1,2}, \cdots, \mathbf{s}_{i,j}, \cdots, \mathbf{s}_{(N_0-1),N_0}\}$ is a set of relation pairs where N_0 denotes the number of objects in face. $\mathbf{o}_i$ and $\mathbf{o}_j$ denote the i-th and j-th object features, respectively. The relational importance $\lambda_{i,j}$ for relation between two object features $(\mathbf{o}_i, \mathbf{o}_j)$ is encoded as:

$$\lambda_{i,j} = h_{\phi_I}(\mathbf{r}_{i,j}), \tag{4}$$

where h_{ϕ_I} is a function with learnable parameters ϕ_I. In this paper, h_{ϕ_I} is defined with $\phi_I = \{(\mathbf{W}_{1,2}, \mathbf{b}_{1,2}), \cdots, (\mathbf{W}_{(N_0-1),N_0}, \mathbf{b}_{(N_0-1),N_0})\}$ as followings:

$$h_{\phi_I}(\mathbf{r}_{i,j}) = \frac{\exp(\mathbf{W}_{i,j}\mathbf{r}_{i,j} + \mathbf{b}_{i,j})}{\sum_{i,j} \exp(\mathbf{W}_{i,j}\mathbf{r}_{i,j} + \mathbf{b}_{i,j})}. \tag{5}$$

The aggregated relational features $\mathbf{f}_{agg}$ are represented by

$$\mathbf{f}_{agg} = \sum_{i,j} \lambda_{i,j}\mathbf{r}_{i,j}. \tag{6}$$

Algorithm 1. Calculating relational importance of N_I objects

---

Input

$\boldsymbol{\lambda}$: the set of relational importance of two facial local parts
$\left\{ \lambda_{1,2}, \cdots, \lambda_{i,j}, \cdots, \lambda_{(N_0-1),N_0} \right\}$

N_I: the number of objects for interpretation

Output: $\boldsymbol{\chi}$: Relational importance of N_I objects

Let N_R denote total number of relations with N_I objects $N_R =_{N_o}\mathrm{C}_{N_I}$

for *P from 1 to N_R* **do**

    Let $\chi_P^{N_I}$ denote P-th relational importance of N_I object features (

    $\mathbf{o}_{p_1}, \mathbf{o}_{p_2}, \cdots, \mathbf{o}_{p_{(N_I-1)}}, \mathbf{o}_{p_{N_I}}$)

    Compute $\chi_P^{N_I}$

    $\chi_P^{N_I} = \lambda_{p_1,p_2} + \cdots + \lambda_{p_1,p_{N_I}} + \cdots + \lambda_{p_{(N_I-1)},p_{N_I}}$

end

$\boldsymbol{\chi} = \left\{ \chi_1^{N_I}, \cdots, \chi_P^{N_I}, \cdots, \chi_{N_R}^{N_I} \right\}$

Finally, the facial trait estimation can be performed with

$$\mathbf{y} = k_{\phi_E}(\mathbf{f}_{agg}), \tag{7}$$

where $\mathbf{y}$ denotes estimated result and k_{ϕ_E} is a function with parameters ϕ_E. k_{ϕ_E} and g_{ϕ_R} are implemented by fully-connected layers.

3.3   Interpretation on Important Relations Between Facial Local Dynamics

The proposed method is useful for interpreting the relations in facial trait estimation. The relational importance calculated in Eq. (4) is utilized to interpret the relations of facial local dynamics. Note that the high relational importance values mean that the relational features of corresponding facial local parts are important for estimating facial traits. The pseudocodes for calculating relational importance of N_I objects are given in Algorithm 1. By analyzing the relational importance, important relations for estimating facial traits could be explained. In Sects. 4.2 and 4.3, we discuss the important relations for age estimation and gender classification, respectively.

4   Experiments

4.1   Experimental Settings

Database. To evaluate the effectiveness of the proposed facial dynamics interpreter network, comparative experiments were conducted. For generalization purpose, we verified the proposed method on both age estimation and gender classification tasks. Age and gender were known as representative facial traits [28]. The public UvA-NEMO Smile database was used for both tasks [10,11].

The UvA-NEMO smile database has been known as the largest smile database [12]. The database consists of 1,240 smile videos collected from 400 subjects. Among 400 subjects, 185 subjects are female and remaining 215 subjects are male. The ages of subjects range from 8 to 76 years. For evaluating the performance of age estimation, we used the experimental protocol defined in [9–11]. The 10-fold cross-validation scheme was used to calculate the performance of the proposed method. Each fold was divided in a way where there was no subject overlap [9–11]. Each time an independent test fold was separated and it was only used for calculating the performance. The remaining 9-folds were used to train the deep network and optimize hyper-parameters. To evaluate the performance of gender classification, we followed the experimental protocol used in [6].

Evaluation Metric. For age estimation, the mean absolute error (MAE) [41] was utilized for evaluation. The MAE could measure the error between the predicted age and the ground-truth. The MAE was computed as follows:

$$\epsilon = \frac{\sum_{n=1}^{N_{test}} \|\hat{\mathbf{y}}_n - \mathbf{y}_n^*\|_1}{N_{test}}, \tag{8}$$

where $\hat{\mathbf{y}}_n$ and $\mathbf{y}_n^*$ denote predicted age and ground-truth age of n-th test sample, respectively. N_{test} denotes the number of the test samples. For the case of gender classification, classification accuracy was used for evaluation. We reported the MAE and classification accuracy averaged over all test folds.

Implementation Details. The face images used in the experiments were automatically aligned based on the two eye locations detected by the facial landmark detection [3]. The face images were cropped and resized to 96×96 pixels. For the appearance representation, the frontal 10 convolutional layers and 4 max-pooling layers of VGG-face network was used. As a result, $6 \times 6 \times 512$ size of feature map was obtained from each face image. Each facial local region was defined on the feature map with size of $2 \times 2 \times 512$. In other words, there were 9 objects in face sequence ($N_0 = 9$). The fully-connected layer with 1024 units and the stacked LSTM layers were used for f_{ϕ_D}. We stacked two LSTMs and each LSTM had 1024 memory cells. Two-layer full-connected layers consisting of 4096 units (with dropout [37]) per layer was used for g_{ϕ_R} with RELU [30]. h_{ϕ_I} was implemented by a fully-connected layer and softmax function. Two-layer fully-connected layers consisting of 2048, 1024 units (with dropout, RELU, and batch normalization [20]) and one fully-connected layer (1 neuron for age estimation and 2 neurons for gender classification) were used for k_{ϕ_E}. The mean squared error was used for training the deep network in age estimation. The cross-entropy loss was used for training the deep network in gender classification.

4.2 Age Estimation

Interpreting Relations Between Facial Local Dynamics in Age Estimation. To understand the mechanism of the proposed facial dynamics interpreter network in age estimation, the relational importance calculated from
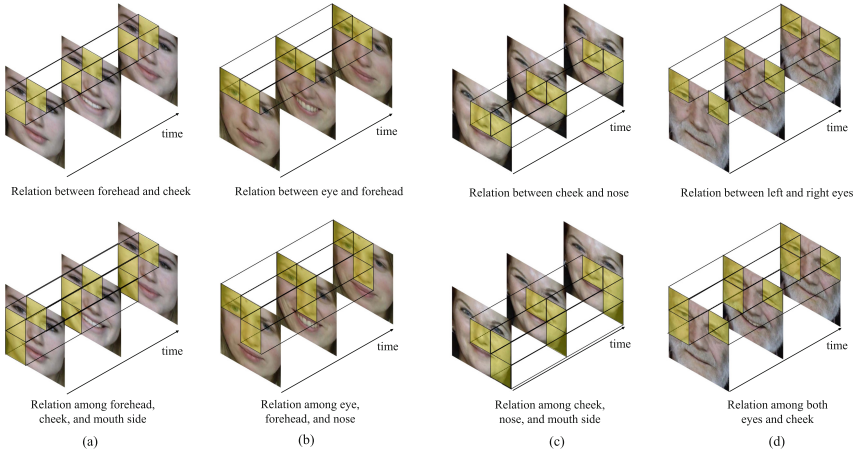
Fig. 2. Example of facial dynamic interpretation in age estimation. Most important relations are visualized with yellow box for the relation between 2 objects in upper side and the relation among 3 objects in bottom side. (a) age group of [13–19], (b) age group of [20–36], (c) age group of [37–66], (d) age group of 66+.

each sequence was analyzed. Figure 2 shows the important relations where the corresponding pair has high relational importance values. We showed the difference of important regions over different ages by presenting the important relations over age groups. Ages were divided into five age groups (8–12, 13–19, 20–36, 37–65, and 66+) according to [15]. To interpret the important relations between each age group, the relational importance values encoded from test set were averaged in each age group, respectively. Four groups were visualized with example face images (there was no subject to be permitted for reporting in age group of [8–12]). As shown in the figure, when estimating age group of [66+], the relation between two eye regions was important. The relation between two eye regions could represent discriminative dynamic features according to crow's feet and sunken eyes, which could be important factors for estimating ages of the older people. In addition, when considering three objects, the relation among left eye, right eye, and left cheek had highest relational importance in age group of [66+]. There was a tendency to symmetry about the relational importance. For example, the relation among left eye, right eye, and right cheek was included in top-5 high relational importance among 84 relations in age group of [66+]. Although the relation of action unit (AU) for determining specific facial expressions has been reported [14], the relation of the motions for estimating age or classifying gender was not investigated. In this study, the facial dynamics interpreter network was designed to interpret the relation of motions in facial trait estimation. It was found that the relation of dynamic features related with AU 2 and AU 6 was highly used by the deep network for estimating ages in range [66+].

In addition, to verify the effect of important relations, we made perturbation on the dynamic features as shown in Fig. 3. For the sequence of 17 years
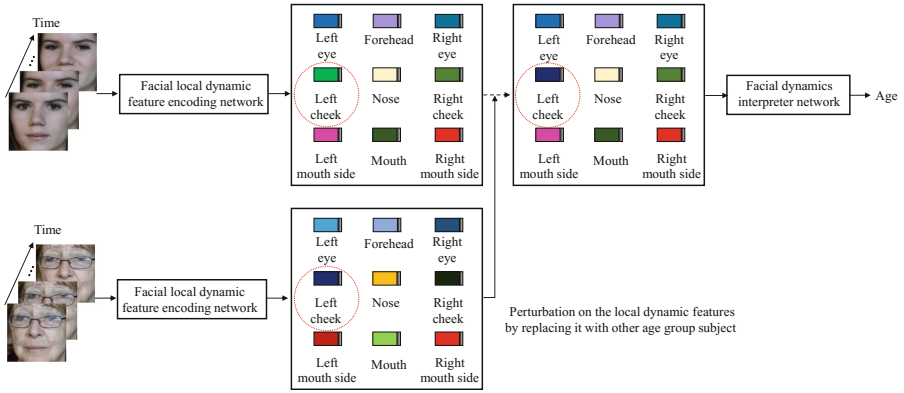
Fig. 3. Perturbing the local dynamic features by replacing it with the local dynamic features of another age group subject (*e.g.*, older subject).
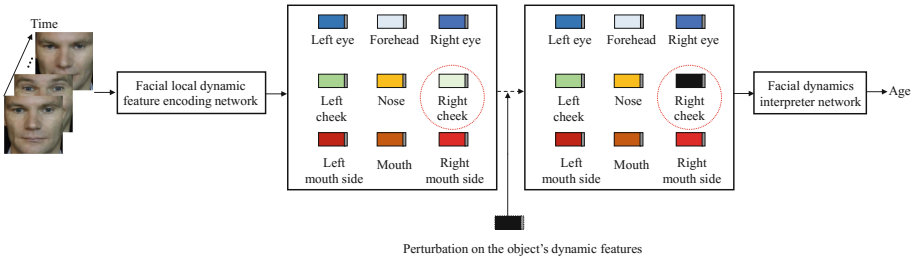

Fig. 4. Perturbing local dynamic features by replacing it with zero vector.

old subject, we changed the local dynamic features of left cheek region with that of 73 years old subject in the experiment. Note that the cheek constructed important pairs for estimating age group of [13–19] as shown in Fig. 2(a). By the perturbation, the absolute error was changed from 0.41 to 2.38. In the same way, we changed the dynamic features of other two regions (left eye and right eye) one by one. The other two regions constructed relatively less important relations and achieved the absolute error of 1.40 and 1.81 (left eye and right eye, respectively). The increase of absolute errors was less than the case which made perturbation on the left cheek. It showed that the relations with the left cheek were important for estimating age compared with the relations with eye in age group of [13–19].

For the same sequence, the facial dynamics interpreter network without the use of relational importance was also analyzed. For the facial dynamics interpreter network without the use of relational importance, the absolute error of the estimated age was increased by perturbation on the local dynamic feature of the left cheek from 1.20 to 7.45. When conducting perturbation on the left eye and the right eye, the absolute errors were 1.87 and 4.21, respectively. The increase of absolute error became much larger when conducting perturbation on

Table 1. Mean absolute errors (MAE) measured after perturbing local dynamic features of different location for subjects with age groups of [37–66].

Perturbation location	MAE (standard error)
Not contaminated	5.05 (±0.28)
Less important parts (left eye, forehead, right eye)	7.56 (±0.35)
Important part (right cheek)	9.00 (±0.41)

Table 2. Mean absolute error (MAE) of age estimation on UvA-NEMO smile database for analyzing the effectiveness of locational features and relational importance. L.F. and R.I. denote locational features and relational importance, respectively.

Method			MAE (years)		
			Posed	Spontaneous	All
Aggregation of local dynamic features using regional importance			4.27	4.25	4.26
	L.F.	R.I.			
Facial dynamics interpreter network			4.05	4.07	4.06
	✔		3.95	4.05	4.00
	✔	✔	**3.83**	**3.90**	**3.87**

the left cheek. Moreover, the increase of error was larger when the facial dynamics interpreter network did not use relational importance. In other words, the facial dynamics interpreter network with the relational importance was more robust to feature contamination because it adaptively encoded the relational importance from the relational features as in Eq. (4).

In order to statistically analyze the effect of contaminated features in the proposed facial dynamics interpreter network, we also evaluated the MAE when conducting perturbation on each dynamic features of facial local parts with zero vector as shown in Fig. 4. For 402 videos which were collected from the subjects in age group of [37–66] in the UvA-NEMO database, the MAE was calculated as shown in Table 1. As shown in the table, the perturbation on most important facial region (*i.e.* right cheek in age group of [37–66]) had more influenced the accuracy of age estimation compared with the case which made perturbation on less important parts (*i.e.* left eye, forehead, and right eye in age group of [37–66]). The difference of MAE between the cases which made perturbation on important part and less important parts was statically significant ($p < 0.05$).

Assessment of Facial Dynamics Interpreter Network for Age Estimation. We evaluated the effectiveness of the facial dynamics interpreter network. First, the effects of relational importance and locational features were validated for age estimation. Table 2 shows the MAE of the facial dynamics interpreter network with locational feature and relational importance. To verify the

Table 3. Mean absolute error (MAE) on the UvA-NEMO smile database compared with other methods.

Method	MAE (years)
VLBP [17]	15.70
Displacement [10]	11.54
BIF [16]	5.78
BIF + Dynamics [9]	5.03
IEF [2]	4.86
IEF + Dynamics [9]	4.33
Holistic dynamic approach	4.02
Proposed method	**3.87**
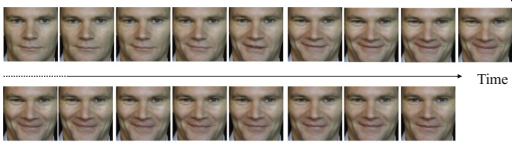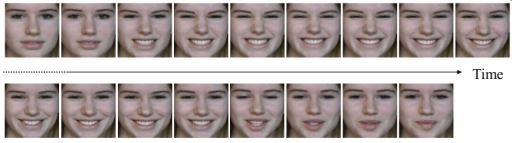
Face sequence	Ground truth	Proposed method	Holistic dynamic approach
Time	47	43.91	40.79
Time	15	15.85	13.43

Fig. 5. Examples of the proposed method on age estimation. For visualization purpose, face sequences are displayed in 5 frames per sec.

effectiveness of the relational features, the aggregation of local dynamic features using regional importance were compared. In the aggregation of local dynamic features using regional importance approach, facial local dynamic features were aggregated with regional importance in unsupervised way. As shown in the table, using the relational features improved the accuracy of age estimation. Moreover, the locational features could improve the performance of the age estimation by making the network know the location information of the object pairs. The locational features of the objects were meaningful as the objects of the face sequence were automatically aligned by the facial landmark detection. By utilizing both the relational importance and the locational features, the proposed facial dynamics interpreter network achieved the lowest MAE of 3.87 over all test set. It was mainly due to the reason that the importance of relations for age estimation was different. By considering the importance of the relational features, the accuracy of age estimation was improved. Moreover, we further analyzed the MAE of the age estimation according to the spontaneity of the smile expression.
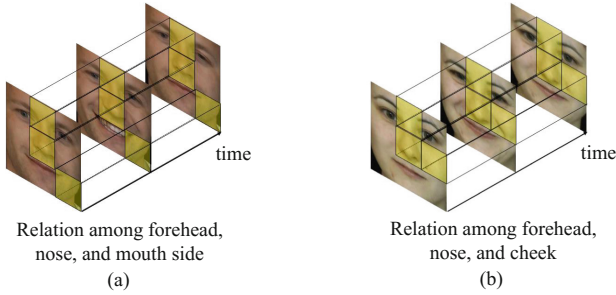
Relation among forehead,
nose, and mouth side
(a)

Relation among forehead,
nose, and cheek
(b)

Fig. 6. Example of interpreting important relations between facial dynamics in gender classification. Most important relations are visualized with yellow box for the relation between 3 objects for recognizing male (a) and for recognizing female (b).

The MAE of the facial dynamic interpreter network was slightly lower in posed smile ($p > 0.05$).

To assess the effectiveness of the proposed dynamics interpreter network (with locational features and relational importance), the MAE of the proposed method was compared with the state-of-the-art methods (please see Table 3). The VLBP [17], displacement [10], BIF [16], BIF with dynamics [9], IEF [2], IEF with dynamics [9], and holistic dynamic approach were compared. In the holistic dynamic approach, appearance features were extracted by the same VGG-face network used in the proposed method and the dynamic features were encoded with the LSTM network on the holistic appearance feature without dividing the face into local parts. It was compared because it has been widely used architecture for a spatio-temporal encoding [13,24,25] As shown in the table, the proposed method achieved lowest MAE. The MAE of the proposed facial dynamics interpreter network was lower than the MAE of the IEF + Dynamics and the difference was statistically significant ($p < 0.05$). It was mainly attributed to the fact that the proposed method encoded the latent relational features from object features (facial local dynamic features and locational features) and effectively combined the relational features based on the relational importance. Examples of age estimation from the proposed method and the holistic dynamic approach are shown in Fig. 5.

4.3   Gender Classification

Interpreting Relations Between Facial Local Dynamics in Gender Classification. In order to interpret important relations in gender classification, the relational importance values encoded from each sequence were analyzed. Figure 6 shows the important relations where the relational importance had high values at classifying gender from face sequence. As shown in the figure, the relation among forehead, nose, and mouth side was important in making decisions on males. Note that there was a tendency to symmetry about the relational importance. For determining male, the relation among forehead, nose, and right

Table 4. Accuracy of gender classification on the UvA-NEMO smile database to analyze the effectiveness of the locational feature and relational importance. L.F. and R.I. denote locational features and relational importance, respectively.

Method			Classification accuracy (%)
Aggregation of local dynamic features using regional importance			88.87
	L.F.	R.I.	
			88.79
Facial dynamics interpreter network	✔		89.35
	✔	✔	**90.08**

Table 5. Accuracy of gender classification on the UvA-NEMO smile database compared with other methods.

Method	Classification accuracy				(%)
	Spontaneous		Posed		All
	Age ≤ 19	Age > 19	Age ≤ 19	Age > 19	
how-old.net + dynamics (tree) [6]	60.80	93.46	N/A	N/A	N/A
how-old.net + dynamics (SVM) [6]	N/A	N/A	60.80	92.89	N/A
COTS + dynamics (tree) [6]	76.92	93.00	N/A	N/A	N/A
COTS + dynamics (bagged trees) [6]	N/A	N/A	76.92	92.89	N/A
Image-based CNN [27]	80.72	89.85	80.58	92.36	86.94
Holistic dynamic approach	74.38	93.52	77.51	93.91	87.10
Proposed method	80.17	94.65	85.14	95.18	**90.08**

mouth side and the relation among forehead, nose, and left mouth side were top-2 important relations among 84 relations of three objects. For the case of female, the relation among forehead, nose, and cheek was important. It could be related to the observation that the females express emotions more frequently compared with males and the males have a tendency to show restricted emotions compared with the females. In other words, the females have a tendency to make smiles bigger than males by using muscles of cheek regions. Therefore, the relations among cheek and other face parts were important for recognizing females.

Assessment of Facial Dynamics Interpreter Network for Gender Classification. We also evaluated the effectiveness of the proposed facial dynamics interpreter network for gender classification. First, the classification accuracy of the facial dynamics interpreter network with relational importance and locational features are summarized in Table 4. For comparison, the aggregation of local dynamic features using regional importance was compared. The proposed facial dynamics interpreter network achieved the highest accuracy by using both locational features and relational importance. The locational features

and the relational importance in the facial dynamics interpreter network were also important for gender classification.

Table 5 shows the classification accuracy of the proposed facial dynamics interpreter network compared with other methods on UvA-NEMO database. Two types of appearance based approach named "how-old.net" and "commercial off-the-shelf (COTS)" were combined with a hand-crafted dynamic approach for gender classification [6]. How-old.net was a website (http://how-old. net/) launched by Microsoft for online age and gender recognition. COTS was a commercial face detection and recognition software, which included a gender classification. The dynamic approach calculated the facial local region's dynamic descriptors such as amplitude, speed, and acceleration as described in [6]. In holistic dynamic approach, appearance features were extracted by the same VGG-face network used in the proposed method and the dynamic features were encoded on the holistic appearance features. An image based method [27] was also compared to validate the effectiveness of utilizing facial dynamics in gender classification. The accuracy of how-old.net + dynamics and COTS + dynamics were directly from [6] and the accuracy of the image-based CNN and the holistic dynamic approach were calculated in this study. By exploiting the relations between local dynamic features, the proposed method achieved the highest accuracy compared with other methods. The performance difference between the holistic approach and the proposed method was statistically significant ($p < 0.05$).

5   Conclusions

According to cognitive-psychological studies, facial dynamics could provide crucial cues for face analysis. The motion of facial local regions from facial expression is known that it is related to the motion of other facial regions. In this paper, the novel deep learning approach which could interpret the relations between facial local dynamics was proposed to interpret relations of local dynamics in facial trait estimation from the smile expression. Facial traits were estimated by combining relational features of facial local dynamics based on the relational importance. By comparative experiments, the effectiveness of the proposed method was verified for facial trait estimation. The important relations between facial dynamics were interpreted by the proposed method in gender classification and age estimation. The proposed method could accurately estimate facial traits (age and gender) compared with the state-of-the-art methods. We will attempt to extend the proposed method to other facial dynamic analysis such as spontaneity analysis [11] and video facial expression recognition [24].

References

1. Adams Jr., R.B., Hess, U., Kleck, R.E.: The intersection of gender-related facial appearance and facial displays of emotion. Emot. Rev. **7**(1), 5–13 (2015)
2. Alnajar, F., Shan, C., Gevers, T., Geusebroek, J.M.: Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions. Image Vis. Comput. **30**(12), 946–953 (2012)
3. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1859–1866 (2014)
4. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recogn. Lett. **36**, 228–234 (2014)
5. Cashdan, E.: Smiles, speech, and body posture: how women and men display sociometric status and power. J. Nonverbal Behav. **22**(4), 209–228 (1998)
6. Dantcheva, A., Brémond, F.: Gender estimation based on smile-dynamics. IEEE Trans. Inf. Forensics Secur. **12**(3), 719–729 (2017)
7. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? a survey on soft biometrics. IEEE Trans. Inf. Forensics Secur. **11**(3), 441–467 (2016)
8. Demirkus, M., Toews, M., Clark, J.J., Arbel, T.: Gender classification from unconstrained video sequences. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 55–62. IEEE (2010)
9. Dibeklioğlu, H., Alnajar, F., Salah, A.A., Gevers, T.: Combining facial dynamics with appearance for age estimation. IEEE Trans. Image Process. **24**(6), 1928–1943 (2015)
10. Dibeklioğlu, H., Gevers, T., Salah, A.A., Valenti, R.: A smile can reveal your age: enabling facial dynamics in age estimation. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 209–218. ACM (2012)
11. Dibeklioğlu, H., Salah, A.A., Gevers, T.: Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 525–538. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_38
12. Dibeklioğlu, H., Salah, A.A., Gevers, T.: Recognition of genuine smiles. IEEE Trans. Multimed. **17**(3), 279–294 (2015)
13. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
14. Ekman, P.: Facial action coding system (FACS). In: A Human Face (2002)
15. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 256–263. IEEE (2009)
16. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 112–119. IEEE (2009)
17. Hadid, A.: Analyzing facial behavioral features from videos. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 52–61. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25446-8_6
18. Hess, U., Adams Jr., R.B., Kleck, R.E.: Facial appearance, gender, and emotion expression. Emotion **4**(4), 378 (2004)

19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
20. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
21. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
22. Juefei-Xu, F., Verma, E., Goel, P., Cherodian, A., Savvides, M.: DeepGender: occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 68–77 (2016)
23. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
24. Kim, D.H., Baddar, W., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect. Comput. (2017). https://doi.org/10.1109/TAFFC.2017.2695999
25. Kim, S.T., Kim, D.H., Ro, Y.M.: Facial dynamic modelling using long short-term memory network: analysis and application to face authentication. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–6. IEEE (2016)
26. Lea, C., Reiter, A., Vidal, R., Hager, G.D.: Segmental spatiotemporal CNNs for fine-grained action segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 36–52. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_3
27. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–42 (2015)
28. Li, S., Xing, J., Niu, Z., Shan, S., Yan, S.: Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 222–230 (2015)
29. Makinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 541–547 (2008)
30. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
31. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC, 1, p. 6 (2015)
32. Pilz, K.S., Thornton, I.M., Bülthoff, H.H.: A search advantage for faces learned in motion. Exp. Brain Res. **171**(4), 436–447 (2006)
33. Reid, D., Samangooei, S., Chen, C., Nixon, M., Ross, A.: Soft biometrics for surveillance: an overview. In: Machine Learning: Theory and Applications, pp. 327–352. Elsevier (2013)
34. Roark, D.A., Barrett, S.E., Spence, M.J., Abdi, H., O'Toole, A.J.: Psychological and neural perspectives on the role of motion in face recognition. Behav. Cogn. Neurosci. Rev. **2**(1), 15–46 (2003)

35. Santoro, A., et al.: A simple neural network module for relational reasoning. In: Advances in Neural Information Processing Systems, pp. 4974–4983 (2017)
36. Simon, R.W., Nath, L.E.: Gender and emotion in the united states: do men and women differ in self-reports of feelings and expressive behavior? Am. J. Sociol. **109**(5), 1137–1176 (2004)
37. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
38. Toews, M., Arbel, T.: Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. IEEE Trans. Pattern Anal. Mach. Intell. **31**(9), 1567–1581 (2009)
39. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1683–1699 (2007)
40. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
41. Uřičař, M., Timofte, R., Rothe, R., Matas, J., et al.: Structured output SVM prediction of apparent age, gender and smile from deep features. In: Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognision Workshop (CVPRW 2016), pp. 730–738. IEEE (2016)
42. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702 (2015)
43. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit and holistic expression recognition. IEEE Trans. Image Process. **25**(8), 3931–3946 (2016)