2017 IEEE International Conference on Computer Vision

# Visual Transformation Aided Contrastive Learning for Video-based Kinship Verification

Hamdi Dibeklioğlu[1,2]
[1] Pattern Recognition & Bioinformatics Group, Delft University of Technology, Delft, The Netherlands
[2] Department of Computer Engineering, Bilkent University, Ankara, Turkey
h.dibeklioglu@tudelft.nl

## Abstract

*Automatic kinship verification from facial information is a relatively new and open research problem in computer vision. This paper explores the possibility of learning an efficient facial representation for video-based kinship verification by exploiting the visual transformation between facial appearance of kin pairs. To this end, a Siamese-like coupled convolutional encoder-decoder network is proposed. To reveal resemblance patterns of kinship while discarding the similarity patterns that can also be observed between people who do not have a kin relationship, a novel contrastive loss function is defined in the visual appearance space. For further optimization, the learned representation is fine-tuned using a feature-based contrastive loss. An expression matching procedure is employed in the model to minimize the negative influence of expression differences between kin pairs. Each kin video is analyzed by a sliding temporal window to leverage short-term facial dynamics. The effectiveness of the proposed method is assessed on seven different kin relationships using smile videos of kin pairs. On the average,* $93.65\%$ *verification accuracy is achieved, improving the state of the art.*

## 1. Introduction

Identifying relatives of a person from visual attributes is a common practice for humans. For instance, when we see a woman and a young kid standing next to each other in a social gathering, we can decide whether the woman is the mother of that kid based on their visual similarity. To this end, humans use an aggregate of different features such as facial appearance and color of hair and eyes [6]. In contrast to strong capability of humans for kinship recognition, automatic identification of kin relationships from facial images is a very challenging task. This is a relatively new research problem, and it has gained a large interest in recent years due to the fact that unobtrusive and rapid computer vision solutions for detecting kinship would have several potential applications such as searching missing child/parent, social media analysis, and family album organization [20].

Most studies in the literature approach kinship analysis as a verification problem [27, 34] while a few works define it as a recognition task [33, 13]. In kinship verification, given a pair of face images, the aim is to identify whether the subjects shown in these images have a kin relationship or not. Kinship recognition, on the other hand, aims at classifying the type of kin relationship such as Mother-Daughter, Mother-Son, Father-Daughter, Father-Son, etc. [13]. This paper focuses on kinship verification.

Recent findings of [24, 8] indicate that temporal appearance and dynamics of facial expressions can also be hereditary. For instance, it has been shown that blind-born people display similar facial expressions with those of their sighted relatives [24]. Yet, all previous studies except two works [8, 3], solely focus on static face images for kinship verification instead of exploring temporal patterns of kin relationships. This paper aims to reveal kinship patterns hidden in facial appearance and short-term expression dynamics in a combined manner. To this end, smile videos of kin pairs are employed, rather than static face images.

In contrast to earlier studies, this paper proposes to learn an efficient representation by modeling a visual mapping that can transform facial appearance of a subject to a very similar form of his/her kin's face while reducing the similarity patterns that can also be observed between non-kins. The model is then fine-tuned in a supervised way to verify kinship between a given pair of subjects. Optimizing facial representation in the visual space can be thought as an extension of unsupervised pre-training (e.g. deep autoencoders). Earlier studies have explored the action of unsupervised pre-training as a regularizer. Such studies state that direct supervised optimization in the feature space disproportionately affects the models due to their extreme flexibility. According to [10], this is the reason of effectiveness of initializing representation by pre-training since supervised training cannot escape from the basin of attraction

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.269

defined by the initialization. As a result, compared to traditional regularizers, effectiveness of pre-training does not fade away. In other words, unsupervised pre-training effectively restricts/simplifies the form of the prediction function by learning a sparse representation. Regularization also achieves the same objective, expressed as a function of the weights. While traditional pre-training initializes the representation so as to achieve the reconstruction of corrupted data samples, the proposed approach *further* restricts the latent representation in a way that it can capture visual resemblance of kin pairs while discarding the similarity patterns between non-kins. Thus, it can theoretically be claimed that the proposed approach would provide a stronger regularization by simplifying the kinship verification model (during pre-training) in a different but related space.

This study is the first exploration of visual transformation aided deep representation learning for kinship verification. The proposed method is evaluated in a detailed manner, compared with several baselines including the recent image- and video-based kinship verification approaches from the literature, and state-of-the-art results are reported.

## 2. Related Work

The journey of automatic kinship recognition has started with the work of Fang *et al.* [11], where facial features such as skin color, position and shape of face parts, and histogram of gradients are employed for verification. Then, several studies have aimed to engineer powerful facial appearance representations such as Spatial Pyramid LEarning-based descriptors [38], DAISY descriptors [12], Gabor-based Gradient Orientation Pyramid [39], Self Similarity Representation [16], semantic-related attributes [32], SIFT flow based genetic Fisher vector feature [26], etc.

With the increase in number and content size of kinship image databases [31, 20, 13, 28], the research focus has shifted to representation and similarity learning [36, 40]. Following the dramatic improvements in deep learning, recent studies have started to utilize deep architectures such as gated autoencoders [7], stacked autoencoders [30], convolutional neural networks (CNN) [37, 21], and convolutional Siamese networks [17] for modeling kinship patterns. Metric learning has also been adopted for both engineered [20, 15, 35, 34] and deep learned features [30, 17].

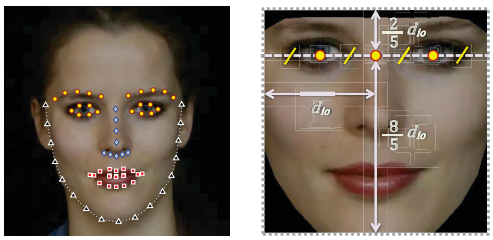
Differently from the aforementioned methods, [8] has explored the temporal facial expression patterns of kinship using spatio-temporal features and landmark displacement dynamics extracted from smile videos, showing that the use of expression dynamics beside the appearance information improves the verification accuracy. A more recent study [3] has complemented the spatio-temporal features with deep learned appearance features, providing a further accuracy improvement. Yet, no other study in the area has focused on temporal analysis for kinship verification.

In contrast to prior studies, the current study proposes to optimize facial feature representation by learning the visual transformation between kin pairs. By defining a contrastive loss in the visual appearance space, the influence of facial similarity patterns that can be observed between subjects who do not share a kin relationship, are reduced. To minimize the effect of expression differences between kin pairs, an expression matching procedure is employed. The representation is then fine-tuned, and the verification model is learned in a supervised manner by using a contrastive loss [17] defined in the feature space. To leverage short-term dynamics of facial expressions, each kin video is analyzed by a sliding temporal window and compared with the matching expression sub-sequences in its pair video.

## 3. Method

The aim of the proposed method is to learn an efficient facial representation to enhance kinship verification by revealing the facial resemblance patterns between kin pairs which are not observed between non-kins. In contrast to most approaches in the literature, the proposed method analyzes facial expression videos, rather than images, in order to exploit the resemblance between facial expressions of kin pairs. To this end, videos of enjoyment smiles are used since it is one of the most frequently shown facial expression.

The method assumes that the given pair of input videos show the entire duration of a smile expression. The flow of the system is summarized as follows. Initially, 68 facial landmarks are tracked in the videos (see Fig. 1). Once face images and the corresponding landmarks are normalized in terms of pose and scale, shape-based features are extracted using the tracked points to represent the surface deformations on eyes & eyebrows and mouth regions. Using the extracted features, each "$2m+1$"-frame sub-sequence (obtained by a sliding window) of the input videos is matched to a "$2m+1$"-frame sub-sequence of its pair-video so as to have a very similar facial expression. Matched sub-sequences are then fed to a coupled convolutional encoder-decoder network that learns the transformation between facial appearance of the given pair of subjects that have a kin relationship (see Fig. 2). Designed architecture has two parallel identical encoder-decoder networks with weight sharing. To reveal the facial resemblance patterns between kin pairs, these encoder-decoder networks are trained in a way that each network outputs a face image that is similar to its input's kin-pair while minimizing its resemblance to non-kins. To this end, an appearance-based contrastive loss is defined. After training the network based on visual similarity, decoding blocks are removed and a classification layer with contrastive loss is connected to the full connection blocks. Modified network is then fine-tuned in a supervised manner by using a feature-based contrastive loss [17]. In the test phase, "$2m+1$"-frame sub-sequences of a given smile




(a) (b)

Figure 1. (a) Landmarks on facial boundary, eyes & eyebrows, nose, and mouth regions, and (b) a normalized/cropped face.

pair are matched based on expression similarity. Each of the matched sub-sequence pairs are processed by the fine-tuned network. The verification result is obtained by fusing the posterior probabilities for all matched sub-sequence pairs.

### 3.1. Facial Landmark Tracking and Alignment

To normalize face images in terms of rotation and scale as well as measuring regional deformations for expression matching, 68 landmarks on facial boundary (17 points), eyes & eyebrows (22 points), nose (9 points), and mouth (20 points) regions are tracked as shown in Fig. 1(a). To this end, a state-of-the-art tracker [2] is used. The tracker uses an extended version of Conditional Local Neural Fields (CLNF) [1], where individual point distribution and patch expert models are learned for eyes, lips and eyebrows. Detected points by individual models are then fit to a joint point distribution model. To handle pose variations, CLNF employs a 3D representation of facial landmarks.

The tracked 3D coordinates of the landmarks $l' = \{l'^X, l'^Y, l'^Z\}$ are normalized by removing the global rigid transformations such as translation, rotation and scale. The movement of the normalized landmarks are smoothed by the 4253H-twice method [29] to reduce the tracking noise. Since the normalized face is frontal with respect to the camera, the depth dimension (Z) is ignored, and each landmark is represented as $l = \{l^X, l^Y\}$. To shape-normalize facial texture, each face image is warped using piecewise linear warping so as to transform the X and Y coordinates of the detected landmarks $l'$ onto those of normalized landmarks $l$. Obtained images are then scaled by setting the inter-ocular distance ($d_{io}$) to 48 pixels, and cropped around the facial boundary and eyebrows as shown in Fig. 1(b). As a result, each normalized face image has a resolution of $96 \times 96$ pixels. Images are then converted to gray scale.

### 3.2. Expression Matching

This paper presents a method that learns an efficient facial representation for kinship verification by exploiting the transformation between facial appearances of kin pairs. To minimize the influence of expression differences in such a transformation, the most similar facial expressions of kin pairs (and those of pairs with no kin relation) are proposed to be matched. To this end, the change in facial surface deformations should be described effectively first. Since previous research [5, 9] shows that facial landmark displacements can successfully describe expression dynamics, a shape-based representation is used in this study.

To leverage regional properties, a separate descriptor is computed for each of eyes & eyebrows, and mouth regions using the corresponding landmarks (Fig. 1(a)). Let $l_{f,i,t}$ denotes the $i^{th}$ landmark (of $N_f$ landmarks) in facial region $f = \{\text{ey}, \text{mt}\}$ at frame $t$ of a given smile video, where "ey" and "mt" indicate regions of eyes & eyebrows and mouth, respectively. Then, a regional shape descriptor $\mathcal{S}_{f,t}$ for frame $t$ can be computed as a set of Euclidean distances between all possible landmark pairs in region $f$:

$$\mathcal{S}_{f,t} = \Big\{ s \in \mathbb{R} \mid s = \|l_{f,j,t} - l_{f,k,t}\|_2,\ j > k, \\ j,k \in \{1,2,3,\dots N_f\} \Big\}, \tag{1}$$

where the length of $\mathcal{S}_{\text{f},t}$ is equal to $\binom{N_f}{2} = \frac{N_f!}{N_f!\,(N_f-1)!}$.

Regional representations of eyes & eyebrows ($\mathcal{S}_{\text{ey},t}$) and mouth ($\mathcal{S}_{\text{mt},t}$) are concatenated to describe the combined expression $\mathcal{S}_{\text{comb},t}$. Dimensionality of the obtained vectors is reduced to 15-d using the Principal Component Analysis (PCA) so as to retain $99.5\%$ of the variance. Resulting frame-based expression descriptor with reduced dimensionality is hereafter denoted as $\bar{\mathcal{S}}_{\text{comb},t}$. To include short-term temporal information in the analysis, expression displayed from frame $t-m$ to $t+m$ can be defined as:

$$\mathcal{D}_t = \left[\ \bar{\mathcal{S}}_{\text{comb},t-m} \quad \bar{\mathcal{S}}_{\text{comb},t-m+1} \quad \dots \quad \bar{\mathcal{S}}_{\text{comb},t+m}\ \right] \tag{2}$$

Note that, a "$2m+1$"-frame sub-sequence of the expression from $t-m$ to $t+m$ is indicated by frame $t$ for simplicity.

Using the extracted descriptors $\mathcal{D}_t$, each "$2m+1$"-frame sub-sequence of the input videos can be matched to a sub-sequence of its pair-video so as to have a very similar facial expression. Let $\mathcal{D}_t^V$ denotes the expression descriptor for the sub-sequence from frame $t-m$ to $t+m$ of a video $V$. If $V_1$ and $V_2$ show the pair of input videos, and $T_1$ and $T_2$ denote the length (number of frames) of $V_1$ and $V_2$, respectively, a matching sub-sequence at time step $t_2^*$ in $V_2$ for $t_1$ in $V_1$ can be determined as follows:

$$t_2^* = \underset{t \in \{m+1,\dots,T_2-m\}}{\arg\min} \|\mathcal{D}_{t_1}^{V_1} - \mathcal{D}_t^{V_2}\|_2 \tag{3}$$

Notice that this is not a one-to-one mapping. So, a single time step in $V_1$ can be matched with several time steps in $V_2$. Therefore, the procedure is repeated by changing the direction of comparison to find a matching time step $t_1^*$ in $V_1$ for each time step $t_2$ of $V_2$. Two sets of matching sub-sequences are obtained as $\{t_1, t_2^*\}$ and $\{t_1^*, t_2\}$, where $t_1, t_1^* \in \{1+m, \dots, T_1-m\}$ and $t_2, t_2^* \in \{1+m, \dots, T_2-$

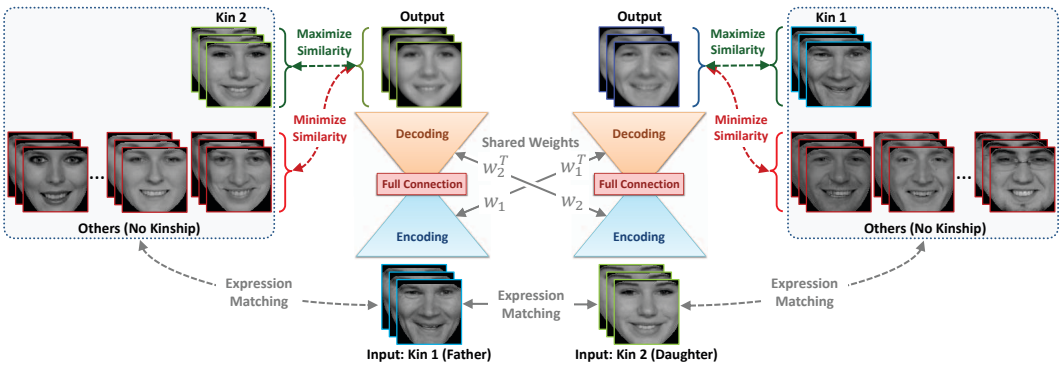



Figure 2. Overview of the proposed method for learning contrastive mapping of facial appearance between kin pairs.

$m\}$. Resulting pairs of "$2m+1$"-frame image sequences are used as inputs in the proposed method.

### 3.3. Appearance-based Pre-training

For a more reliable kinship verification, the proposed method aims to learn a facial appearance representation that can capture the patterns of visual kinship resemblance while discarding the similarity patterns that can be observed between people who do not have a kin relationship, by employing a contrastive loss function. To this end, a Siamese-like architecture is designed by combining two identical convolutional encoder-decoder networks in parallel as shown in Fig. 2. In contrast to the conventional Siamese architectures, the proposed pre-training network optimizes the parameters by employing a loss function defined in the image space, rather than the feature space. Instead of solely relying on label information (kin or non-kin), the proposed approach exploits the visual similarity patterns between kins and between non-kins. Such an approach can provide a strong regularization in the model, yielding a more accurate kinship verification.

Inputs of the designed network are a pair of matching facial image sequences of two subjects with a kin relationship. While the first encoder-decoder learns the transformation from kin 1 to kin 2, other one learns the transformation from kin 2 to kin 1. These sub-networks cross-share the weights such that the decoding weights of the second sub-network are defined as the transposed encoding weights of the first sub-network. As shown in Table 1, the encoder blocks contain four convolutional layers followed by a fully connected layer. A set of $5 \times 5$-pixel filters are used in all convolutional layers. To add non-linearity to the model, rectified linear unit (ReLU) is applied to the output of each convolutional layers. The encoder blocks contains three max-pooling layers which are applied after the second, third, and fourth convolutional layers. Max-pooling is applied with a $2 \times 2$ window such that the output of max-pooling layer is downsampled with a factor of 2. The final layer of the encoding blocks is the fully connected layer that aims to aggregate information obtained from all neurons from the third max-pooling layer. The decoder blocks are the symmetric of encoder blocks such that max-pooling layers are replaced with max-unpooling and convolutions are replaced with transposed convolutions. Note that, similar to the encoding, the transposed convolutional layers are followed by ReLU in the decoder blocks except the last layer.

Table 1. Configuration of the proposed network. "conv", "trconv", and "fc" denote convolution, transposed convolution, and full connection layers in the network, respectively. Numbers next to the name of each layer indicate the order of the corresponding layer. For brevity, ReLU layers are discarded in the table.

| Layer | Kernel Size | Stride | Output Size |
|---|---|---|---|
| conv1-1 | $5 \times 5$ | 1 | $92 \times 92 \times 16$ |
| conv1-2 | $5 \times 5$ | 1 | $88 \times 88 \times 32$ |
| pool1 | $2 \times 2$ | 2 | $44 \times 44 \times 32$ |
| conv2 | $5 \times 5$ | 1 | $40 \times 40 \times 64$ |
| pool2 | $2 \times 2$ | 2 | $20 \times 20 \times 64$ |
| conv3 | $5 \times 5$ | 1 | $16 \times 16 \times 128$ |
| pool3 | $2 \times 2$ | 2 | $8 \times 8 \times 128$ |
| fc4 | $8 \times 8$ | 1 | $1 \times 1 \times 512$ |
| trconv-fc4 | $8 \times 8$ | 1 | $8 \times 8 \times 128$ |
| unpool3 | $2 \times 2$ | 2 | $16 \times 16 \times 128$ |
| trconv3 | $5 \times 5$ | 1 | $20 \times 20 \times 64$ |
| unpool2 | $2 \times 2$ | 2 | $40 \times 40 \times 64$ |
| trconv2 | $5 \times 5$ | 1 | $44 \times 44 \times 32$ |
| unpool1 | $2 \times 2$ | 2 | $88 \times 88 \times 32$ |
| trconv1-2 | $5 \times 5$ | 1 | $92 \times 92 \times 16$ |
| trconv1-1 | $5 \times 5$ | 1 | $96 \times 96 \times (2m+1)$ |

To optimize the parameters of the proposed network, an appearance-based contrastive loss function is proposed. Let $\mathcal{K}_1 = \{I_{1,1}, \ldots, I_{1,2m+1}\}$ and $\mathcal{K}_2 = \{I_{2,1}, \ldots, I_{2,2m+1}\}$ denote the given input pair of "$2m+1$"-frame image sequences, and $\hat{\mathcal{K}}_1 = \{\hat{I}_{1,1}, \ldots, \hat{I}_{1,2m+1}\}$ and $\hat{\mathcal{K}}_2 = \{\hat{I}_{2,1}, \ldots, \hat{I}_{2,2m+1}\}$ denote the corresponding output pair of image sequences, respectively. If $C$ is the cosine distance between two images, then, a kinship loss can be defined to learn the transformation of facial appearance between kin pairs by maximizing the similarity between $\mathcal{K}_1$ and $\hat{\mathcal{K}}_2$, and

between $\mathcal{K}_2$ and $\hat{\mathcal{K}}_1$ as follows:

$$\ell_{\text{kin}} = \Theta(\hat{\mathcal{K}}_1, \mathcal{K}_2) + \Theta(\hat{\mathcal{K}}_2, \mathcal{K}_1) \text{ , where}$$
$$\Theta(\hat{\mathcal{K}}_i, \mathcal{K}_j) = \frac{1}{2m+1} \sum_{n=1}^{2m+1} C(\hat{I}_{i,n}, I_{j,n}) \,. \tag{4}$$

In a similar manner, a non-kinship loss is defined to minimize the resemblance of the output image sequences to the matching "$2m+1$"-frame sequences of other subjects (with no kinship). To this end, one matching sequence with $\mathcal{K}_1$ and one with $\mathcal{K}_2$ are obtained for each of the subjects (with the same gender) in the database who do not have a kin relation with the input subjects as described in Section 3.2. A fraction of the most similar sequences are determined as $\{\mathcal{P}_{1,1}, \mathcal{P}_{1,2}, \ldots, \mathcal{P}_{1,N_{s_1}}\}$ for $\mathcal{K}_1$ and as $\{\mathcal{P}_{2,1}, \mathcal{P}_{2,2}, \ldots, \mathcal{P}_{2,N_{s_2}}\}$ for $\mathcal{K}_2$, where $N_{s_1}$ and $N_{s_2}$ denote the number of obtained most similar subjects to the input subject 1 and 2, respectively. Then, the non-kinship loss is computed as:

$$\ell_{\text{non-kin}} = -\frac{1}{N_{s_2}} \sum_{n=1}^{N_{s_2}} \Theta(\hat{\mathcal{K}}_1, \mathcal{P}_{2,n}) - \frac{1}{N_{s_1}} \sum_{n=1}^{N_{s_1}} \Theta(\hat{\mathcal{K}}_2, \mathcal{P}_{1,n}). \tag{5}$$

Thus, the model can be trained by minimizing the contrastive loss that is defined as:

$$\ell_{\text{contrastive}} = \lambda\, \ell_{\text{kin}} + (1-\lambda)\, \ell_{\text{non-kin}} \,, \tag{6}$$

where $\lambda$ is the tradeoff parameter between the kinship loss and the non-kinship loss.

### 3.4. Fine-tuning and Classification

Once the model is pre-trained, decoding blocks of the network are removed. Instead, a binary classification layer is connected to the full connection blocks to fine-tune the network parameters and distinguish between kin and non-kin pairs. Note that the obtained architecture is a Siamese network with two encoding blocks without weight sharing.

For the classification layer, the contrastive learning method that is proposed by Li *et al.* [17], is employed. Let $x_1$ and $x_2$ be the "$2m+1$"-frame input sequences, and $G_{w_1}(x_1)$ and $G_{w_2}(x_2)$ be the representation obtained from the corresponding encoding block (responses of the fully connected layer) for subject 1 and subject 2, respectively, where $w_1$ and $w_2$ denote the weights of the encoding blocks. Then, the distance between $G_{w_1}(x_1)$ and $G_{w_2}(x_2)$ is:

$$\gamma(x_1, x_2) = \|G_{w_1}(x_1) - G_{w_2}(x_2)\|_1. \tag{7}$$

To classify kin and non-kin pairs, a distance threshold $\tau$ is defined such that $\tau > 1$ and $y\,\big(\tau - \gamma(x_1, x_2)\big) > 1$ , where $y \in \{-1 \text{ (non-kin)}, 1 \text{ (kin)}\}$ is the kinship label for the given subject pair. Thus, a contrastive loss that enforces the model to pull kin pairs to each other while pushing apart non-kin pairs in the feature space, can be defined as follows:

$$\hat{\ell}_{\text{contrastive}} = \delta\Big(1 - y\,\big(\tau - \gamma(x_1, x_2)\big)\Big) \,, \tag{8}$$

where $\delta(\Phi) = \frac{1}{\beta}\log(1 + e^{\beta\Phi})$ is the generalized logistic loss function, and $\beta$ denotes the sharpness parameter. By minimizing $\hat{\ell}_{\text{contrastive}}$, the weights of the modified network are fine-tuned for the kinship verification task. The posterior probability of being kin $p(y = 1 \,|\, x_1, x_2)$ is estimated using the sigmoid of $\gamma(x_1, x_2)$ distance.

In the test phase, when videos $\{V_1, V_2\}$ of a subject pair are given, every "$2m+1$"-frame sequence of $V_1$ is matched with the most similar sequence in $V_2$ using a temporal sliding window (see Section 3.2). For each of the matched subsequence pairs $\{x_1(t), x_2(t^*)\}$, the posterior probability of being kin $p(y = 1 \,|\, x_1(t), x_2(t^*))$ is estimated. Then, a kinship probability for $V_1$ is computed as:

$$\mathcal{Q}(V_1 \to V_2) = \frac{1}{T_1 - 2m} \sum_{n=m+1}^{T_1-m} p(y = 1 \,|\, x_1(t_n), x_2(t_n^*)) \,, \tag{9}$$

where $T_1$ denotes the number of frames in $V_1$. Next, the process described above, from sequence matching to computation of the kinship probability, is repeated for $V_2$ by changing the direction of comparison. Once $\mathcal{Q}(V_2 \to V_1)$ is computed, the kin relation between subject 1 and subject 2 is verified if $\big(\mathcal{Q}(V_1 \to V_2) + \mathcal{Q}(V_2 \to V_1)\big) > 1$.

## 4. Database

To train and evaluate the proposed architecture for video-based kinship verification, the kinship partition [8] of the UvA-NEMO Smile Database [9] is used. It has spontaneous/posed enjoyment smiles of 95 subject pairs who have a kin relationship. Ages of subjects vary from 8 to 74 years. Videos have a resolution of $1920\times1080$ pixels at a rate of 50 frames per second. There are 152 subjects in the database. There is no spontaneous smile videos for 15 of the subjects while six subjects do not have a posed one. Each of the remaining subjects has one or two posed/spontaneous enjoyment smiles. 1031 pairs of smile videos are obtained by using different video combinations of each kin relation. These pairs consist of Sister-Sister (S-S), Brother-Brother (B-B), Sister-Brother (S-B), Mother-Daughter (M-D), Mother-Son (M-S), Father-Daughter (F-D), and Father-Son (F-S) relationships. Table 2 shows the number of pairs of subjects, videos, and matched sequences for each kin relationship.

To assess the proposed model for image-based kinship verification, the Kinship Face in the Wild datasets (KFW-I and KFW-II) [19, 20] are employed. In KFW-I, there are 156 F-S, 134 F-D, 116 M-S, and 127 M-D image pairs. KFW-II has 250 image pairs for each of these relations. To

Table 2. Distribution of subject, video, and matched sub-sequence (for $m = 2$) pairs in the UvA-NEMO Smile Database.

| Relation | Number of Pairs | | |
|---|---|---|---|
| | Subject | Video | Matched Sequence |
| Mother-Daughter | 22 | 262 | 86,390 |
| Mother-Son | 14 | 162 | 55,585 |
| Father-Daughter | 10 | 120 | 44,175 |
| Father-Son | 20 | 188 | 66,652 |
| Sister-Sister | 9 | 109 | 43,579 |
| Brother-Brother | 8 | 54 | 22,361 |
| Sister-Brother | 12 | 136 | 46,480 |
| All | 95 | 1031 | 365,222 |

increase the number of training pairs for the experiments, combinations of original and mirror images are matched.

# 5. Experiments

To evaluate the proposed method, and to assess the effectiveness of using appearance-based pre-training, contrastive loss, expression matching, and short-term temporal information in video-based kinship verification, the UvA-NEMO Smile Database is employed. Furthermore, in Section 5.5, to evaluate the reliability of the method for image-based kinship verification, KFW-I and KFW-II datasets are used (see Section 4). While kinship pairs are used as positive samples, random pairs that do not have a kin relation are used as negative samples. These random pairs are specifically constructed for each subset. For instance, a negative F-S pair is prepared by replacing the son with another male child while the father is retained. Numbers of positive and negative (random) pairs are kept same to have a balanced dataset. A separate verification model is trained for each of the M-D, M-S, F-D, F-S, S-S, B-B, and S-B relationships. S-S and B-B relations are not used for image-based assessment since they are not available in KFW-I/II. Each experiment is repeated three times using a different random set of negative samples at each time. Average (over repeated experiments) of the obtained mean (over different relations) correct verification rates are reported. In a similar manner, performance of the system is also measured using the area under ROC curve (AUC). The same set of random negative samples are used in each experiment for a fair comparison.

All experiments are conducted using a two level leave-one-subject-pair-out cross-validation scheme. Each time videos of a test pair are separated, the system is trained and parameters are optimized using leave-one-subject-pair-out cross-validation on the remaining pairs. The proposed networks are implemented using Lua/Torch for GPU, and trained using the standard stochastic gradient descent (SGD). Fixed learning rate is optimized during cross-validation (considered values: $\{0.01, 0.005, 0.001\}$).

Table 3. Effect of using different levels of frame-neighborhood ($m$) in kinship verification, and the correlation of smile amplitude between matched pairs.

| Temporal Width | Matching Correlation | AUC | Accuracy (%) |
|---|---|---|---|
| 1 ($m = 0$) | 0.87 | 0.90 | 89.51 |
| 3 ($m = 1$) | 0.85 | 0.95 | 92.93 |
| 5 ($m = 2$) | 0.83 | 0.96 | 93.65 |
| 7 ($m = 3$) | 0.81 | 0.92 | 91.21 |

## 5.1. Temporal Dynamics & Expression Matching

The proposed method exploits short-term temporal dynamics of expressions for a better kinship verification. To this end, "2m+1"-frame sequences (neighborhood of m frames) of given videos are matched based on expression similarity. Obtained sequence matches are used for verification. To assess the effectiveness of temporal information as well as the influence of temporal width of matched segments, different frame-neighborhood levels ($m$) are evaluated. As shown in Table 3, the matched sequences ($m > 0$) perform consistently better than the matched frames. While the AUC of using sequence-pairs reaches to 0.96, that of using frame-pairs is only 0.90. This shows the effectiveness of employing short-term expression dynamics in the analysis, and confirms the findings in the literature that suggest the expression dynamics display hereditary patterns [24, 8].

The optimal level of frame-neighborhood is found as $m = 2$ (five-frame sequences), achieving an accuracy of 93.65%. Since the minimum validation error is also achieved by five-frame sequence pairs, $m$ is set to 2 in the remainder of the experiments. Increasing the sequence length up to five frames improves the accuracy, showing the benefit of temporal information. Yet, further length increase causes loss in accuracy. The reasons may be that: (1) Complexity of the model is not increased for longer sequences; (2) Matching is getting much more challenging with longer sequences, yielding inaccurate matching, thus expression effects cannot be well minimized; (3) Expression similarity of face pairs may appear only in instant/short responses. The use of random (not matched) five-frame sequence pairs is also evaluated, and it can only reach 82.74% accuracy where that of using matched pairs is 93.65%.

Next, the reliability of expression matching is evaluated. Since the UvA-NEMO Smile Database does not include per-frame annotations of expression intensity, following [9], the smile amplitude of each frame is estimated as the average distance of the right and left lip corners to the lip center, normalized by the length of the lip. The average correlation of smile amplitude between matched segments of each video pair is computed. Table 3 reports the obtained correlation coefficients. Although increasing the frame-neighborhood decreases the matching quality, the matched

Table 4. Effect of the non-kinship loss using different percentages of the most similar non-kins, in kinship verification.

| Non-kinship Loss | % of Non-kins | AUC | Accuracy (%) |
|---|---|---|---|
| ✗ | 0 | 0.91 | 90.28 |
| ✓ | 10 | 0.96 | 93.65 |
| ✓ | 50 | 0.94 | 92.15 |
| ✓ | 100 | 0.92 | 91.31 |

pairs are highly correlated in terms of smile amplitude. Also notice that the decrease in matching correlation does not directly affect the verification performance.

## 5.2. Assessment of Non-kinship Loss

During pre-training, a non-kinship loss is employed for contrastive learning so as to optimize the latent representation by minimizing the importance of the similarity patterns that can also be observed between non-kins. To this end, a fraction of the most similar non-kin samples to the input samples are used. To evaluate the effectiveness of this approach and observe the influence of the non-kin fraction, the method is compared with modified models that use different amount (%) of the most similar non-kin samples and with a model that does not use a non-kinship loss.

As shown in Table 4, the use of non-kinship loss employing 10% of the most similar non-kins provides the highest accuracy that is, in terms of AUC, 4% and 2% (absolute) better than using all and 50% of non-kins, respectively. These findings may suggest that using a large amount of non-kins samples in image-based loss, would enforce the model to focus on general/prominent differences since the loss is averaged over all non-kins. Thus, the learned latent representation may not manage to capture subtle but important characteristics. Results also show that not using non-kinship loss performs worst, confirming the effectiveness of the proposed non-kinship loss. 10% of the most similar kin samples are used in the remainder of the experiments since this setting achieves the minimum test/validation error.

## 5.3. Assessment of Contrastive Learning

One of the main contributions of this study is to learn an efficient representation for kinship verification using a limited number of subject pairs. To this end, an appearance-based contrastive loss is proposed so as to provide a stronger regularization in the model based on the fact that the supervised verification task is defined in a different space. Furthermore, the proposed model is fine-tuned using a contrastive loss defined in the feature space to enhance the reliability of kinship verification. To assess the influence of contrastive loss on representation learning, and to compare the effectiveness of contrastive losses defined in the visual appearance and in the feature spaces, four different models are trained and compared. Table 5 reports the kinship verification performance of these models, where "standard" and "contrastive" pre-training indicate the sole use of $\ell_{\text{kinship}}$ and $\ell_{\text{contrastive}}$ in the proposed visual-similarity based pre-training architecture, respectively (see Section 3.3). Similarly, "contrastive" fine-tuning indicates the use of $\hat{\ell}_{\text{contrastive}}$ in the proposed feature-similarity based fine-tuning architecture (see Section 3.4). "Standard" fine-tuning represents a modified fine-tuning architecture that replaces the presented classification layer with the logistic regression.

Table 5. Effect of using contrastive loss in kinship verification.

| Pre-training (appearance-based) | Fine-tuning (feature-based) | AUC | Accuracy (%) |
|---|---|---|---|
| Contrastive | Contrastive | 0.96 | 93.65 |
| Contrastive | Standard | 0.94 | 92.72 |
| Standard | Contrastive | 0.91 | 90.28 |
| Standard | Standard | 0.88 | 86.30 |

As the first and the last rows of Table 5 show, using contrastive loss in both of the visual-similarity based pre-training and feature-similarity based fine-tuning provides $8.3\%$ AUC improvement compared to the use of standard loss. This indicates the success of contrastive learning for kinship verification. When the contrastive loss functions that are defined in the appearance space and in the feature space are compared, appearance-based (visual-similarity based) contrastive loss is found to be more effective than the feature-based (see the second and third rows in Table 5). This finding may indicate the additional regularization provided by the optimization in appearance space, while the supervised verification task is defined in the feature space.

## 5.4. Comparison to Video-based Methods

The proposed method is compared with the state-of-the-art video-based kinship verification approaches from the literature [8, 3]. The method proposed by Dibeklioğlu *et al.* [8] extracts spatio-temporal features (CLBP-TOP) [25], from different facial regions. To represent temporal dynamics, a set of statistical descriptors are extracted from regional facial movements. Boutellaa *et al.* present two approaches in [3]. First one extracts deep features for each frame of smile videos using the pre-trained VGG-face CNN model [22]. Then, each video is represented by averaging the features of all frames. The second approach represents smiles by combining the deep representation with several spatio-temporal descriptors.

An additional baseline has been implemented, where matched sequences of five frames are used, the proposed transformation network is replaced with the stacked denoising autoencoders (SDAE; 4 hidden layers), and the fine-tuning is kept same. Furthermore, two state-of-the-art image-based methods, namely, Neighborhood Repulsed Metric Learning (NRML) [20] and Similarity Metric based CNN (SMCNN) [17] have been modified, using the pro-

Table 6. Accuracy (%) of different video-based methods. * shows the methods that have been modified for video-based verification.

| Method | Representation | M-D | M-S | F-D | F-S | S-S | B-B | S-B | Mean |
|---|---|---|---|---|---|---|---|---|---|
| *Proposed Method* | Short-term Temporal Appearance (Deep) | **93.64** | **92.24** | **93.83** | **93.35** | **94.18** | **95.71** | **92.58** | **93.65** |
| SDAE | Short-term Temporal Appearance (Deep) | 85.24 | 85.44 | 87.48 | 87.89 | 86.33 | 87.88 | 84.54 | 86.40 |
| SMCNN: Li *et al.* [17]* | Static Appearance (Deep) | 83.58 | 81.46 | 85.15 | 84.81 | 84.64 | 86.43 | 85.84 | 84.56 |
| NRML: Lu *et al.* [20]* | Static Appearance | 76.36 | 77.94 | 79.48 | 80.99 | 78.24 | 78.43 | 76.78 | 78.32 |
| Boutellaa *et al.* [3] | Static Appearance (Deep) + Temporal Appearance | 91.23 | 90.49 | 93.10 | 88.30 | 88.93 | 94.74 | 90.07 | 90.98 |
| Boutellaa *et al.* [3] | Static Appearance (Deep) | 90.24 | 85.69 | 89.70 | 92.69 | 88.92 | 92.82 | 88.47 | 89.79 |
| Dibeklioğlu *et al.* [8] | Displacement Dynamics + Temporal Appearance | 67.54 | 75.00 | 75.00 | 78.95 | 75.00 | 70.00 | 68.75 | 72.89 |

posed expression matching (single-frame) and frame-score fusion, for video-based kinship verification, and compared with the proposed model. NRML learns a metric that minimizes the distance between kin pairs while maximizing that of non-kins. For NRML, Learning-Based features [4] are employed. SMCNN is a convolutional Siamese network with metric constraints and shared weights.

As shown in Table 6 the proposed method, outperforms all the competitor methods. The use of handcrafted features that represent temporal appearance and facial dynamics [8] can only reach an accuracy of $72.89\%$, where that of the proposed method is $93.65\%$. Accuracy of employing deep face features [3] obtained from pre-trained VGG [22] model is $89.79\%$. When temporal appearance representation is combined with these deep-learned features, the verification accuracy of [3] reaches to $90.98\%$. Notice that the deep representation used in [3] is learned using 13 linear convolutional layers and trained on 2.6 million images of 2622 people. Still, the accuracy of the proposed method in this study is $2.67\%$ (absolute) higher than that of [3]. Accuracy of the modified single-image based methods SMCNN and NRML are 84.56% and 78.32%, respectively. Lower accuracy of these methods can be explained by the fact that they do not exploit temporal information and solely rely on static images. Compared to the SDAE-based approach, the proposed method provides an additional accuracy of $7.25\%$ (absolute). Notice that the SDAE-based approach is a modified version of our method which replaces the proposed encoder/decoder blocks with SDAE. All these findings indicate the effectiveness of the proposed method even it is trained with a limited number of kin pairs (subjects). Additionally, the accuracy of the method for each kin relationship is analyzed. It is shown that the best results are obtained for B-B and S-S pairs, respectively. Such a finding can be explained by the age and gender resemblance.

### 5.5. Comparison to Image-based Methods

To assess the effectiveness of the proposed model for image-based kinship verification, a modified version of the method has been prepared by removing the expression matching step, and setting the level of frame-neighborhood ($m$) to 0. The modified method is evaluated on KFW-I and KFW-II datasets, and compared with the state-of-the-art image-based methods in terms mean verification accuracy. As shown in Table 7, the proposed model outperforms all the competitors, and reaches an accuracy of 80.5% and 82.3% on KFW-I and KFW-II, respectively.

Table 7. Accuracy (%) of different image-based methods.

| Method | KFW-I | KFW-II |
|---|---|---|
| *Proposed Method* | **80.5** | **82.3** |
| Block-based Neighbor. Repulsed Metric [23] | 78.7 | 80.6 |
| Local Large-Margin Multi-Metric [14] | - | 80.0 |
| Similarity Metric Based CNN [17] | 72.7 | 79.3 |
| Neighborhood Repulsed Correlation Metric [34] | 66.3 | 78.7 |
| Ensemble Similarity [40] | 78.6 | 75.7 |
| Scalable Similarity [41] | 77.6 | 74.8 |
| Asymmetric Metric [18] | 78.4 | 80.9 |
| Neighborhood Repulsed Metric [20] | 64.3 | 75.7 |

## 6. Conclusion

In this paper, a deep contrastive learning architecture for video-based kinship verification has been proposed. The proposed architecture employs an appearance-based pre-training step followed by a feature-based supervised fine-tuning. For pre-training, a novel contrastive loss that is defined in the visual-appearance space, has been introduced. To minimize the influence of facial expression differences between given subject pairs, an expression matching procedure has been proposed. Furthermore, to capture temporal similarity patterns of kin expressions, facial image sequences of kin pairs have been used in the analysis instead of solely focusing on static appearance.

Each of the proposed components employed in the presented kinship verification framework such as representation optimization in the visual-appearance space, contrastive learning, frame matching, and the use of short-term facial dynamics, has been evaluated on the UvA-NEMO Smile Database in a detailed manner. Furthermore, the proposed method has been contrasted with the state-of-the-art methods for video- and image-based kinship verification. The experimental results have confirmed the effectiveness of each of the proposed components as well as showing the reliability of the method for kinship verification even with employing a limited number of subject pairs for training.

## References


[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[3] E. Boutellaa, M. B. López, S. Ait-Aoudia, X. Feng, and A. Hadid. Kinship verification from videos using spatio-temporal texture features and deep learning. In *International Conference on Biometrics*, pages 1–7, 2016.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2707–2714, 2010.

[5] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(2):121–132, 2004.

[6] M. F. Dal Martello and L. T. Maloney. Where are kin recognition signals in the human face? *Journal of Vision*, 6(12), 2006.

[7] A. Dehghan, E. G. Ortiz, R. Villegas, and M. Shah. Who do I look like? Determining parent-offspring resemblance via gated autoencoders. In *International Conference on Computer Vision*, pages 1757–1764, 2014.

[8] H. Dibeklioğlu, A. Ali Salah, and T. Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *International Conference on Computer Vision*, pages 1497–1504, 2013.

[9] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. 2012.

[10] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[11] R. Fang, K. D. Tang, N. Snavely, and T. Chen. Towards computational models of kinship verification. In *IEEE International Conference on Image Processing*, pages 1577–1580, 2010.

[12] G. Guo and X. Wang. Kinship measurement on salient facial features. *IEEE Trans. on Instrumentation and Measurement*, 61(8):2322–2325, 2012.

[13] Y. Guo, H. Dibeklioğlu, and L. van der Maaten. Graph-based kinship recognition. In *International Conference on Pattern Recognition*, pages 4287–4292, 2014.

[14] J. Hu, J. Lu, Y.-P. Tan, J. Yuan, and J. Zhou. Local large-margin multi-metric learning for face and kinship verification. *IEEE Trans. on Circuits and Systems for Video Technology*, 2017.

[15] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Asian Conference on Computer Vision*, pages 252–267, 2014.

[16] N. Kohli, R. Singh, and M. Vatsa. Self-similarity representation of weber faces for kinship classification. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 245–250, 2012.

[17] L. Li, X. Feng, X. Wu, Z. Xia, and A. Hadid. Kinship verification from faces via similarity metric based convolutional neural network. In *International Conference Image Analysis and Recognition*, pages 539–548, 2016.

[18] J. Lu, J. Hu, V. E. Liong, X. Zhou, A. Bottino, I. U. Islam, T. F. Vieira, X. Qin, X. Tan, S. Chen, et al. The FG 2015 kinship verification in the wild evaluation. In *IEEE International Conference on Automatic Face and Gesture Recognition Workshops*, 2015.

[19] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, and G. Wang. Neighborhood repulsed metric learning for kinship verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[20] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2014.

[21] I. Önal-Ertuğrul and H. Dibeklioğlu. What will your future child look like? Modeling and synthesis of hereditary patterns of facial dynamics. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 33–40, 2017.

[22] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[23] B. Patel, R. Maheshwari, and B. Raman. Evaluation of periocular features for kinship verification in the wild. *Computer Vision and Image Understanding*, 160:24–35, 2017.

[24] G. Peleg, G. Katzir, O. Peleg, M. Kamara, L. Brodsky, H. Hel-Or, D. Keren, and E. Nevo. Hereditary family signature of facial expression. *Proceedings of the National Academy of Sciences*, 103(43):15921–15926, 2006.

[25] T. Pfister, X. Li, G. Zhao, and M. Pietikainen. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *IEEE International Conference on Computer Vision Workshops*, pages 868–875, 2011.

[26] A. Puthenputhussery, Q. Liu, and C. Liu. SIFT flow based genetic fisher vector feature for kinship verification. In *IEEE International Conference on Image Processing*, pages 2921–2925, 2016.

[27] X. Qin, X. Tan, and S. Chen. Tri-subject kinship verification: Understanding the core of a family. *IEEE Trans. on Multimedia*, 17(10):1855–1867, 2015.

[28] J. P. Robinson, M. Shao, Y. Wu, and Y. Fu. Families in the wild (fiw): Large-scale kinship image database and benchmarks. In *ACM Internatiol Conference on Multimedia*, pages 242–246, 2016.

[29] P. F. Velleman. Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association*, 75(371):609–615, 1980.

[30] M. Wang, Z. Li, X. Shu, and J. Tang. Deep kinship verification. In *International Workshop on Multimedia Signal Processing*, 2015.

[31] S. Xia, M. Shao, and Y. Fu. Kinship verification through transfer learning. In *International Joint Conference on Artificial Intelligence*, 2011.

[32] S. Xia, M. Shao, and Y. Fu. Toward kinship verification using visual attributes. In *International Conference on Pattern Recognition*, pages 549–552, 2012.

[33] S. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo. *IEEE Trans. on Multimedia*, 14(4):1046–1056, 2012.

[34] H. Yan. Kinship verification using neighborhood repulsed correlation metric learning. *Image and Vision Computing*, 60:91–97, 2017.

[35] H. Yan, J. Lu, W. Deng, and X. Zhou. Discriminative multimetric learning for kinship verification. *IEEE Trans. on Information Forensics and Security*, 9(7):1169–1178, 2014.

[36] H. Yan, J. Lu, and X. Zhou. Prototype-based discriminative feature learning for kinship verification. *IEEE Trans. on Cybernetics*, 45(11):2535–2545, 2015.

[37] K. Zhang, Y. Huang, C. Song, H. Wu, and L. Wang. Kinship verification with deep convolutional neural networks. In *British Machine Vision Conference*, 2015.

[38] X. Zhou, J. Hu, J. Lu, Y. Shang, and Y. Guan. Kinship verification from facial images under uncontrolled conditions. In *ACM International Conference on Multimedia*, pages 953–956, 2011.

[39] X. Zhou, J. Lu, J. Hu, and Y. Shang. Gabor-based gradient orientation pyramid for kinship verification under uncontrolled environments. In *ACM International Conference on Multimedia*, pages 725–728, 2012.

[40] X. Zhou, Y. Shang, H. Yan, and G. Guo. Ensemble similarity learning for kinship verification from facial images in the wild. *Information Fusion*, 32:40–48, 2016.

[41] X. Zhou, H. Yan, and Y. Shang. Kinship verification from facial images by scalable similarity fusion. *Neurocomputing*, 197:136–142, 2016.