What will Your Future Child Look Like? Modeling and Synthesis of Hereditary Patterns of Facial Dynamics

Itır Önal Ertuğrul[1] and Hamdi Dibeklioğlu[2]
[1] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
[2] Pattern Recognition & Bioinformatics Group, Delft University of Technology, Delft, The Netherlands
itir@ceng.metu.edu.tr, h.dibeklioglu@tudelft.nl

***Abstract*— Analysis of kinship from facial images or videos is an important problem. Prior machine learning and computer vision studies approach kinship analysis as a verification or recognition task. In this paper, first time in the literature, we propose a kinship synthesis framework, which generates smile videos of (probable) children from the smile videos of parents. While the appearance of a child's smile is learned using a convolutional encoder-decoder network, another neural network models the dynamics of the corresponding smile. The smile video of the estimated child is synthesized by the combined use of appearance and dynamics models. In order to validate our results, we perform kinship verification experiments using videos of real parents and estimated children generated by our framework. The results show that generated videos of children achieve higher correct verification rates than those of real children. Our results also indicate that the use of generated videos together with the real ones in the training of kinship verification models, increases the accuracy, suggesting that such videos can be used as a synthetic dataset.**

I. INTRODUCTION

Analysis of kin relations from facial appearance has gained popularity in recent years. This research topic has several potential applications including missing child/parent search, social media analysis, family album organization, and image annotation [1]. Majority of prior studies in kinship analysis focus on *kinship verification* [2], [3], [4]; given a pair of face images, they try to identify whether these two have a kin relationship or not. On the other hand, *kinship recognition* studies aim to classify the type of kin relationship such as Father-Daughter, Mother-Son, etc. [5].

In addition to general appearance of face, style and appearance of expressions can also be inherited. Facial expressions of congenitally blind and deaf phocomelian children, who are incapable of sensing their relatives' face by touching, are shown to be similar to those of their parents [6]. Moreover, [7] reports that a blind-born son, who was abandoned by his mother two days after birth, displays similar facial expressions with the biological mother. Findings of [4] show that the use of expression dynamics extracted from videos together with facial appearance leads to more accurate kinship verification compared to employing only facial appearance. Thus, although facial expressions may comprise learned characteristics, it is clear that they are at least partially inherited.

All of the previous studies approach the kinship analysis as a verification or recognition problem. They model the underlying relationship between a pair of images or videos, yet, what these models learn is not visible to humans. In this study, first time in the literature, we focus on *kinship synthesis*, and generate facial expression videos of children using the expression video of their parents. Kinship synthesis has several benefits. First of all, since we synthesize videos, the hereditary patterns inherited from parent to child can be observed by humans. Observed patterns may even be useful for genetic research. Secondly, there is only one kinship video database (UvA-NEMO Smile Database [8]) available for automatic kinship analysis, thus, our models can be used to create synthetic kinship videos for further research. Lastly, with the help of our model, people will be able to preview how their (probable) future child may look like, as well as seeing his/her smile as a video. Therefore, if a child, whose appearance and expression dynamics are unknown, has been missing for years, generated videos of him/her (based on expressions of the parents) would be better references for the search compared to pictures drawn by forensic artists.

This study is the very first exploration of synthesizing facial images and expression videos for a kin relationship. By transforming temporal dynamics and appearance of a given subject, we generate a video of his/her probable children. Furthermore, we show that the synthesized samples can be used to improve the state of the art in kinship verification.

II. RELATED WORK

Most of the studies that analyze kinship from images using machine learning and computer vision aim to solve kinship verification problem. In their pioneering study, Fang et al. [9] employ facial features such as skin color, position and shape of face parts, and histogram of gradients for kinship verification. Following that study, a number of feature representations for this task are proposed/evaluated such as DAISY descriptors [10], Spatial Pyramid LEarning-based (SPLE) descriptors [11], Gabor-based Gradient Orientation Pyramid (GGOP) [12], Self Similarity Representation (SSR) [13], semantic-related attributes [14], SIFT flow based genetic Fisher vector feature (SF-GFVF) [15], etc. Moreover, a prototype-based discriminative feature learning (PDFL) method has been proposed [16], and a gated autoencoder is trained to characterize the similarity between faces of parents and children for kinship verification [17]. Metric learning has also been adopted for kinship verification problem in various studies [1], [18], [19], [2]. Furthermore, a genetic similarity

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.14

measure between child-parent pairs is learned in an ensemble learning framework [20].

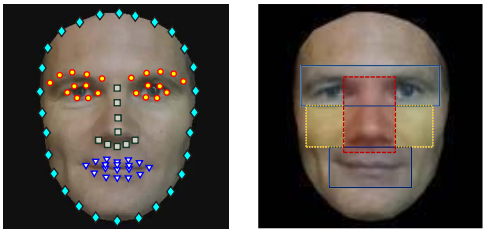
Beside one-to-one kinship verification, a number of studies focus on verification or recognition of kin relations in family images [5], [21], [22], [23], [3]. They predict whether a face image has kin relation with multiple family members [21], classify given a query face image which family it belongs to [22], [23], perform tri-subject kinship verification using the core parts of a family including mother-father pair to verify the kinship of child [3], and recognize the exact type of kin relation in family photos [5]. Recently, kinship verification has also been approached using a pair of videos rather than images, and it is shown that the use of expression dynamics beside the appearance information improves the verification accuracy [4]. However, no study thus far focuses on the synthesis of kin images or videos of a given subject.

In terms of image synthesis, convolutional neural networks have been found to be quite successful for a number of different tasks. For instance, in [24] a deep fully convolutional neural network architecture, SegNet, for semantic pixel-wise segmentation has been proposed. It consists of an encoder network and a corresponding decoder network followed by a pixel-wise classification layer. Decoder network maps the low resolution encoder feature maps to full input resolution feature maps for pixel-wise classification, where the output of the network is the segmented input image. Similarly, [25] uses a fully convolutional encoder-decoder network for contour detection. In [26], a generative up-convolutional neural network has been proposed to re-generate images of objects for a given object style, viewpoint, and color.

In [27], a very deep fully convolutional encoding-decoding framework has been proposed for image restoration. Its encoding network acts as a feature extractor that preserves the primary components of objects in the image while eliminating the corruptions. Decoding network recovers the details of image contents. The output of the network is the denoised version of the input image. [28] designs a recurrent encoder-decoder network to synthesize rotated views of 3D objects. This model captures long-term dependencies along a sequence of transformations with the help of the recurrent structure. A different encoder-decoder architecture has been proposed in [29] to modify facial attributes such as including glasses or a hat on a given face image.

III. METHOD

In this paper, we propose to model relations of facial appearance and dynamics between smiles of parent-child pairs, and combine them to synthesize a smile of the probable/future child of a given subject. To generate such smile videos, we use a single smile video of reference subjects as input (parent) data. To train our models, smile videos of parent-child pairs are used. Our method requires complete smiles that are composed of three phases, i.e., the onset (neutral to expressive), apex, and offset (expressive to neutral), respectively. We focus on (enjoyment) smile since it is one of the most frequently performed facial expression.



(a) (b)

Fig. 1: (a) Normalized/cropped face image, the tracked landmarks, and (b) the defined patches on eyes & eyebrows, nose, mouth, and cheek regions. Note that the cheek patches are only used for the kinship verification experiments

In this section, details of the proposed method are described. The flow of the method is as follows. Facial landmarks in regions of eyes & eyebrows, mouth and nose are tracked during smile videos. Euclidean distances between all possible pairs of the landmarks in each region are computed to describe regional surface deformations. Using these distances, the most similar frames of parent and child videos are matched. Matched parent-child frames are then fed as input-output pairs to a deep encoder-decoder network to model the relation between facial appearances of parent-child pairs. Another network is designed to learn the mapping between smile dynamics of parent-child pairs based on the extracted distance measures over time. Once both networks are trained, smile dynamics of the most probable child (based on the model) of a given subject (reference parent) is estimated. Afterwards, smile dynamics of the reference parent is transformed to that of the estimated child by re-ordering frames of the parent video. The modified video (smile) has the appearance of the given subject but the temporal dynamics of the estimated child. Finally, smile (video) of the estimated child is obtained by transforming the appearance (of each frame) of the modified video to child's appearance through the deep encoder-decoder network.

A. Facial Landmark Tracking and Alignment

To normalize face images in terms of rotation and scale, and to measure regional deformations in face, we track 77 facial landmarks. To this end, we use a state-of-the-art tracker proposed by Jeni *et al.* [30]. Of the tracked landmarks, 28 are on facial boundary, 22 are on eyes & eyebrows, 9 are on nose, and 18 are on mouth as shown in Fig. 1(a). The tracker employs a combined 3D supervised descent method [31], where the shape model is defined by a 3D mesh and the 3D vertex locations of the mesh [30]. A dense parameterized shape model is registered to an image such that its landmarks correspond to consistent locations on the face.

The tracked 3D coordinates of the facial landmarks $\ell' = \{\ell'^X, \ell'^Y, \ell'^Z\}$ are normalized by removing the global rigid transformations such as translation, rotation and scale. Since the normalized face is frontal with respect to the camera, we ignore the depth dimension (Z) and represent each facial point as $\ell = \{\ell^X, \ell^Y\}$. To shape-normalize facial texture, we warp each face image (using piecewise linear

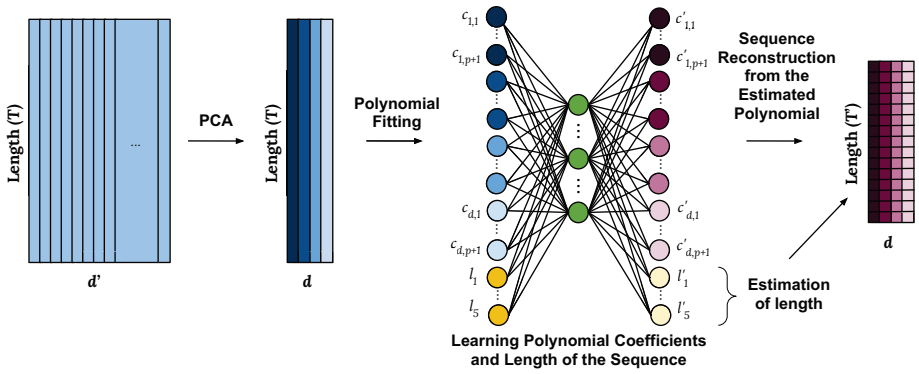


Fig. 2: An illustration of learning temporal dynamics

warping) so as to transform the X and Y coordinates of the detected landmarks ℓ' onto those of normalized landmarks ℓ. Obtained face images are then scaled by setting the inter-ocular distance to 40 pixels, and cropped around the facial boundary as shown in Fig. 1. As a result, each normalized face image (including black pixels around facial boundary) has a resolution of 128×128 pixels.

B. Learning Temporal Dynamics

To model the temporal dynamics of a smile, we first need to effectively describe the change in facial surface deformations. Since previous research shows that facial landmark displacements can successfully describe expression dynamics [8], [32], we use a shape-based representation in our study. To leverage regional properties, a separate descriptor is computed for each of eyes & eyebrows, nose, and mouth regions using the corresponding landmarks (see Fig. 1(a)). Let $\ell_{f,i,t}$ denote the i^{th} landmark (of N_f landmarks) in facial region $f = \{\text{eyes \& eyebrows, nose, mouth}\}$ at frame t of a given smile video. 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:

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

where the length of the feature vector $\mathcal{S}_{f,t}$ is equal to $\binom{N_f}{2}$.

As $\mathcal{S}_{f,t}$ is a frame-based descriptor, temporal dynamics of each facial region during a smile (of T frames) can be represented by a $\binom{N_f}{2}$-dimensional time series with a length of T. N_f equals 22 for $f =$ eyes & eyebrows, 9 for $f =$ nose, and 18 for $f =$ mouth. Since different dimensions (column vectors) of each regional time series $\mathcal{S}_f$ are highly correlated, dimensionality of $\mathcal{S}_f$ is reduced (for each region) to d_f using the Principal Component Analysis (PCA) so as to retain 99% of the variance. The resulting time series (for each region) with reduced dimensionality is hereafter referred to as $\mathcal{R}_f$.

Duration of smiles varies in length (T). Yet, we need to represent dynamics of varying-length smiles by a fixed-length descriptor since we do not employ temporal models. To this end, we fit a separate p^{th}-degree polynomial to each dimension of regional time series $\mathcal{R}_f$. Notice that each column vector (dimension) of $\mathcal{R}_f$ can be considered as $g(t) = y_t$, where $\forall t \in L = \{1,2,\dots T\}$, and polynomials can be fit to these functions. However, to fit better polynomials, we normalize t to have zero mean and unit variance, and obtain $\bar{t}$. By preserving the feature values, our new function becomes $\bar{g}(\bar{t}) = y_{\bar{t}}$. Yet, such a normalization causes the loss of the length information. Thus, to learn the mapping between smile lengths of parents and children, five length-related features are included in our feature set, namely, length of the time series (T), mean value of L (μ_L), standard deviation of L (σ_L), $1-\mu_L$, and $T-\mu_L$. Although one of these features would be sufficient, we estimate a separate length value (T) from each, and use their average as the final estimation to minimize the error. As a result, a $(p+1)\,.\,d_f + 5$ dimensional feature vector is obtained for each of eyes & eyebrows, nose, and mouth regions.

Once the features are computed, the mapping between smile dynamics of parent-child pairs is learned using a neural network with a single hidden layer as illustrated in Fig. 2. Although temporal dynamics of a given time series may be more efficiently learned by deep temporal models such as Recurrent Neural Networks (RNNs), the limited sample size of the video pairs in the UvA-NEMO Smile Database does not allow us to use such models. To train the neural network, we use the feature vectors obtained from parents as inputs and the ones obtained from the corresponding children as targets. We employ stochastic gradient descent (SGD) to train our network with a learning rate of 0.05. During the synthesis phase, we estimate the coefficients of d_f distinct polynomials along with the length (T) of the time series. Using these estimates, regional time series $(\mathcal{R}_f^{\text{child}})$ for the corresponding child can be reconstructed.

C. Learning Appearance

1) Expression Matching: To learn an efficient appearance transformation from parents' face to that of children, we propose to remove the influence of expression differences between input (parent) and target (child) images. To this end, we match the most similar facial expressions of parent-child pairs (in the database) in terms of facial shape. Using the

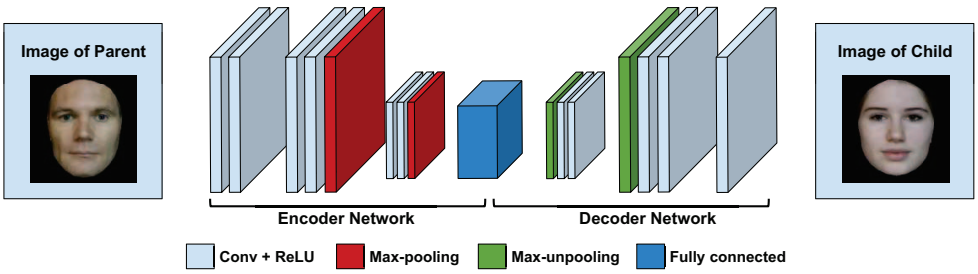


Fig. 3: An illustration of the convolutional encoder-decoder network that models the appearance transformation

regional per-frame shape descriptors $\mathcal{S}_{f,t}$ (see Section III-B), a matching child frame t^* can be obtained for each video frame t of the corresponding parent as:

$$t_f^* = \underset{t' \in \{1,2,3,\ldots,T'\}}{\arg\min} \|\mathcal{S}_{f,t}^{\text{parent}} - \mathcal{S}_{f,t'}^{\text{child}}\| \quad (2)$$

where T' denotes the length (number of frames) of the child's video, and f shows the region that is used for matching. Instead of matching frames based on average similarity of different regions, we obtain a separate set of matched pairs for each of the eyes & eyebrows, nose, and mouth regions. Mouth region is also used to match whole face of parent-child pairs since lip movements define the smile expression as well as influencing the appearance of cheek and chin regions. We match regional patches to better synthesize these regions, and overlay them on the whole face.

2) Model: Once parent-child frames are matched, these image pairs are fed as input-output pairs to a deep convolutional network to model the relation between facial appearances of parent-child pairs as shown in Fig. 3. Our model has an encoder network and a corresponding decoder network. The encoder network contains three convolutional layers followed by a fully connected layer. Each encoder in the encoder network applies convolution operation using a set of filter bank. We employ filters of 3×3 pixels in all convolutional layers. After convolution, rectified linear unit (ReLU) is applied to the output of the convolutional layers in order to add non-linearity to the model. Our encoder network contains two max-pooling layers which are applied after the second and the third convolutional layers. We apply max-pooling with a 2×2 window and stride 2 such that the output of max-pooling layer is downsampled with a factor of 2. Max-pooling summarizes the activated neurons from the previous layer and enables translation invariance over small spatial shifts in the input image. The final layer of the encoding network is the fully connected layer that aims to aggregate information obtained from all neurons from the second max-pooling layer. The decoder network is the symmetric of encoder network such that max-pooling layers are replaced with max-unpooling layers. Note that, similar to the encoder network, convolutional layers are followed by ReLU in the decoder network.

We train four separate networks to learn the appearance transformation of whole face, eyes & eyebrows region, nose region, and mouth region. When we use facial regions to train the network, we crop the corresponding region from the normalized face image (of 128×128 pixels) as shown in Fig. 1(b) and resize it to 64×64 pixels before feeding to our network. Note that the matched-expression pairs of whole face images are determined based on mouth shape ($\mathcal{S}_{\text{mouth}}$). For training, SGD with a fixed learning rate of 0.01 is used, while mean squared error (MSE) is used as the objective function. The encoder and decoder weights are initialized from the uniform distribution over $[-r, r]$ where $r = 1/(W \,.\, H \,.\, U)$, and W is the width and H is the height of the filter. U denotes the number of input planes.

D. Expression Synthesis

This section explains how we use the models of dynamics and appearance to generate a smile video of the estimated child of a given subject. After computing the regional smile dynamics of an estimated child, we transform the regional dynamics of the parent ($\mathcal{R}_f^{\text{parent}}$) to that of the estimated child ($\mathcal{R}_f^{\text{child}}$) by re-ordering the frame sequence of the parent. Let $I_{f,s^{\text{parent}}}^{\text{parent}}$ denote the image sequence of facial region f of the parent, where $s^{\text{parent}} = [1, 2, \ldots, T_{\text{parent}}]$ shows the sequence of frame indices and T_{parent} is the number of frames. Recall that $\mathcal{R}_f$ is a time series of per-frame shape features $\mathcal{R}_{f,t}$ with a reduced dimensionality of d_f (see III-B), where the q^{th} dimension of $\mathcal{R}_{f,t}$ can be shown as $\mathcal{R}_{f,t,q}$. Then, a re-ordered sequence $\hat{s}$ can be obtained ensuring that $\mathcal{R}_{f,\hat{s}}^{\text{parent}} \simeq \mathcal{R}_{f,s^{\text{child}}}^{\text{child}}$ using Algorithm 1. Note that the first dimension of $\mathcal{R}_f$ ($\mathcal{R}_{f,s,q=1}$) can be thought as the amplitude signal of the regional expression, since it explains the majority of the variance of $\mathcal{S}_f$. Thus, if the image sequence of the estimated child displays expressions with higher amplitudes than that of the parent, we reduce the values of $\mathcal{R}_f^{\text{child}}$ such that the regional amplitude of the estimated child can reach only 60-100% of the maximum amplitude of parent's expression. This ratio is defined randomly (see Algorithm 1) to avoid having the same maximum amplitude for smiles of the parent and the estimated child. Length of $\mathcal{R}_f^{\text{child}}$ is accordingly reduced using bi-cubic interpolation to preserve the temporal dynamics such as speed and acceleration of change in $\mathcal{R}_f^{\text{child}}$.

Afterwards, each frame of the re-ordered image sequence $I_{f,\hat{s}}^{\text{parent}}$ of the parent is transformed to that of the estimated child using the learned convolutional model (Section III-C) as visualized in Fig. 4. This procedure is repeated for each of the eyes & eyebrows, nose, and mouth regions, and for

Algorithm 1 Re-ordering the frame sequence of parent so as to display the dynamics of the estimated child

Require: $\mathcal{R}_f^{\text{parent}}$ of size $T_{\text{parent}} \times d_f$
Require: $\mathcal{R}_f^{\text{child}}$ of size $T_{\text{child}} \times d_f$
Require: Explained ratio of $\mathcal{S}_f$'s variance ($\Lambda_{f,q}$) by each dimension $q \in \{1, 2, \ldots, d_f\}$ of $\mathcal{R}_f$ (see Section III-B)
Ensure: $\mathcal{R}_{f,\hat{s}}^{\text{parent}} \simeq \mathcal{R}_{f,s^{\text{child}}}^{\text{child}}$

$m_{\text{parent}} \leftarrow \max(\mathcal{R}_{f,s^{\text{parent}},1}^{\text{parent}})$
$m_{\text{child}} \leftarrow \max(\mathcal{R}_{f,s^{\text{child}},1}^{\text{child}})$
if $m_{\text{child}} > m_{\text{parent}}$ **then**
  $\text{rate} \leftarrow \frac{m_{\text{parent}}}{m_{\text{child}}} \times \text{random}([0.6 \quad 1], \text{uniform})$
  $\mathcal{R}_f^{\text{child}} \leftarrow \text{rate} \times \mathcal{R}_f^{\text{child}}$
  $T_{\text{child}} \leftarrow \lfloor \text{rate} \times T_{\text{child}} \rceil$
  $\mathcal{R}_f^{\text{child}} \leftarrow \text{resize}(\mathcal{R}_f^{\text{child}} \text{ s.t. } T_{\text{child}} \times d_f)$
end if
for $i = 1 \rightarrow T_{\text{child}}$ **do**
  $s_i \leftarrow \underset{j \in \{1,2,\ldots,T_{\text{parent}}\}}{\arg\min} \sum_{k=1}^{d_f} (\mathcal{R}_{f,i,k}^{\text{child}} - \mathcal{R}_{f,j,k}^{\text{parent}})^2 \cdot \Lambda_{f,k}$
end for
$\hat{s} \leftarrow s$

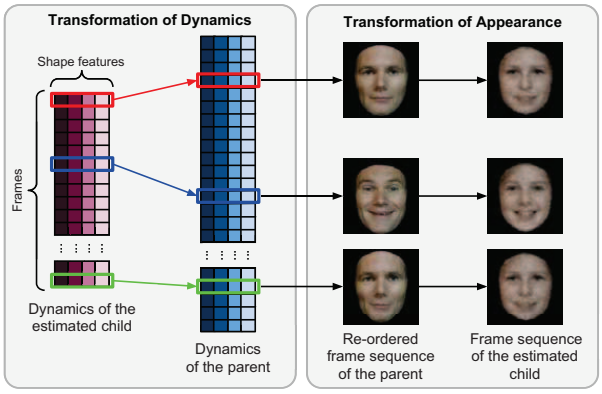


Fig. 4: Generation of the image sequence (whole face) of the estimated child

the whole face. Once regional/whole-facial image sequences of the estimated child are generated, we overlay regional patches on the whole face image at each frame. In order to have a smooth transition between regional textures, alpha blending is used on/around the boundaries of regions. In this way, the smile video of the estimated child is obtained.

IV. Database

In order to synthesize videos of children from videos of the corresponding parents, we employ the kinship set [4] of the UvA-NEMO Smile Database [8]. The kinship dataset has spontaneous and posed enjoyment smiles of the subject pairs who have kin relationships. Ages of subjects vary from 8 to 74 years. Videos have a resolution of 1920×1080 pixels at a rate of 50 frames per second. In our experiments, spontaneous video pairs of Mother-Daughter (M-D), Mother-Son (M-S), Father-Daughter (F-D), and Father-Son (F-S) relationships are used. Each of the subjects in the database has one or two spontaneous enjoyment smiles. By using different video combinations of each kin relation, 159 pairs of spontaneous smile videos are obtained. Note that we also employ the matched frames of posed smile pairs to model the facial appearance but the corresponding posed videos are not used in the test/evaluation stage. The number of subject pairs, spontaneous video pairs, and spontaneous parent videos for each kin relationship are given in Table I.

TABLE I: Distribution of subject and (spontaneous) video pairs in the the kinship database

Relation	Pairs		Parent Videos
	Subject	Video	
Mother-Daughter	16	57	29
Mother-Son	12	36	21
Father-Daughter	9	28	16
Father-Son	12	38	21
All	49	159	87

V. Experiments & Results

Our method aims to synthesize smiles of the most probable children (rather than actual ones) of given subjects. Based on the fact that even the appearances of siblings, except maternal twins, are different, we cannot directly compare synthesized and real children to evaluate our method. Thus, for a quantitative assessment, we use the estimated smiles to train a spatio-temporal kinship verification system, and evaluate our method based on the obtained results. To this end, we use a state-of-the-art method proposed by Dibeklioğlu *et al.* [4]. The method [4] extracts Completed Local Binary Patterns from Three Orthogonal Planes (CLBP-TOP) features [33] from the regions eyes & eyebrows, cheeks, and mouth to describe regional appearance over time. Regional features are concatenated as an appearance feature vector. To represent temporal dynamics of smiles, a set of statistical descriptors are extracted from the displacement signals of eyelids & eyebrows, cheeks, and lip corners, and combined in a dynamics feature vector. After a feature selection step, the temporal appearance and dynamics are separately modeled by SVMs. The final verification result is obtained through a decision level fusion. In the current study, we slightly modify this method by extracting CLBP-TOP features from the regions of eyes & eyebrows, nose, and mouth & cheeks (see Fig. 1(b)). Additionally, we extract dynamics features from the shape-based time series $\mathcal{R}_f$ for regions of eyes & eyebrows, nose, and mouth. Other details are kept same with those of the original method [4].

Kinship set of the UvA-NEMO Smile Database and the generated smiles are used in our experiments. While kinship pairs are used as positive samples, randomly selected pairs that do not have a kin relation are used as negative samples.

TABLE II: Accuracy (%) of using real and synthesized temporal appearance in kinship verification

Training Set	Test Set	
	Real	Synthesized
Real	63.14	61.89
Synthesized	59.37	65.41
Real + Synthesized	67.17	69.18

A separate verification model is trained for each of the M-D, M-S, F-D, and F-S relations. Each experiment is repeated 10 times so as to use a different random set of negative samples each time. Average (over repeated experiments) of the obtained mean (over different relations) correct verification rates are reported. Both kinship verification and synthesis experiments are conducted using a two-level leave-two-pair-out cross-validation scheme. Each time two test pairs are separated, the system is trained and parameters are optimized using leave-two-pair-out cross-validation on the remaining subject pairs.

For the synthesis of whole-facial and regional appearance of the estimated children, we train separate appearance transformation models for each kin relationship, i.e., M-D, M-S, F-D, F-S. To model temporal dynamics, d_f is chosen as 4 for each facial region so as to retain 99% of the variance. Degree of the polynomial fitting (for temporal dynamics) is set to 5 since our preliminary experiments show that polynomial degrees lower than 5 are limited to capture subtle patterns of dynamics while higher degrees are quite sensitive to noise, and could easily generate infeasible smile signals with continuous exponential increase. Dynamics network is trained using all kin relationships due to the limited number of video pairs. To have a similar quality for the pairs of real parent and estimated child during the verification experiments on the synthesized videos, we train a convolutional encoder-decoder network for reconstructing the input face images. Frames of the parent videos are modified by this network. In the remainder of this section, the results of our experiments will be presented.

A. Assessment of the Synthesized Temporal Appearance

In this experiment, we only use the spatio-temporal features (CLBP) extracted from the regions of eyes & eyebrows, nose, and mouth (over videos) for kinship verification. To evaluate the quality of the synthesized videos, we train three different kinship verification models, i.e. using real videos, using synthesized videos, and with their combined set. Each of the trained models are then tested on the real and synthesized samples. As shown in Table II, the temporal appearance features extracted from the synthesized videos achieve an accuracy of 65.41% when the verification model is trained on the synthesized set, which is about 2.3% (absolute) higher than that of the real video pairs when the model is learned on real data. Besides, only 3.5% accuracy decrease is observed for the synthesized videos if the system is trained on the real video pairs. All these results clearly suggest the reliability of our proposed method. Our visual analysis also confirms the realistic appearance of the synthesized face images (see Fig. 5(a)). Furthermore, training the system by using real and synthesized videos together, increases the accuracy for both real (4%) and synthesized pairs (3.8%). Under this setting, synthesized videos perform 2% better than real ones. These findings show that indeed the obtained synthetic data can be used to train a more accurate kinship verification system.

TABLE III: Accuracy (%) of using real and synthesized temporal dynamics in kinship verification

Training Set	Test Set	
	Real	Synthesized
Real	64.91	65.41
Synthesized	63.40	68.43
Real + Synthesized	69.18	70.44

TABLE IV: Accuracy (%) of the combined use of temporal appearance and dynamics of real and synthesized videos in kinship verification

Training Set	Test Set	
	Real	Synthesized
Real	73.71	72.70
Synthesized	71.45	77.74
Real + Synthesized	78.49	80.25

B. Assessment of the Synthesized Dynamics

Similar to the previous experiment, we conduct cross-database experiments using real and synthesized video pairs to evaluate the reliability of the estimated facial dynamics. To this end, we only use dynamics features in the verification model. As shown in Table III, estimated smile dynamics performs slightly (1%) worse than the dynamics of real smiles when the system is trained with real videos. Once we use synthesized dynamics along with the real ones to train the model, a 4.8% accuracy increase obtained for real pairs compared to the model trained using solely real samples. Moreover, synthetic samples perform better than real pairs when the model is learned on the combined data. As in the previous experiment, these findings show the efficacy of our method as well as indicating the importance of using synthetic data in addition to real samples during the training of kinship verification models.

C. Combining Temporal Appearance and Dynamics

To assess the full performance of the synthesized videos in kinship verification, we include both temporal appearance and dynamics features in the verification system. As shown in Table IV, when the system is trained solely on real samples, the accuracy for real samples reaches 73.7% where the accuracy for synthetic videos is only 1% less. Verification accuracy for real pairs are enhanced by 4.8% (absolute) by including the synthesized samples in the training set. Moreover, synthesized videos perform better than the real

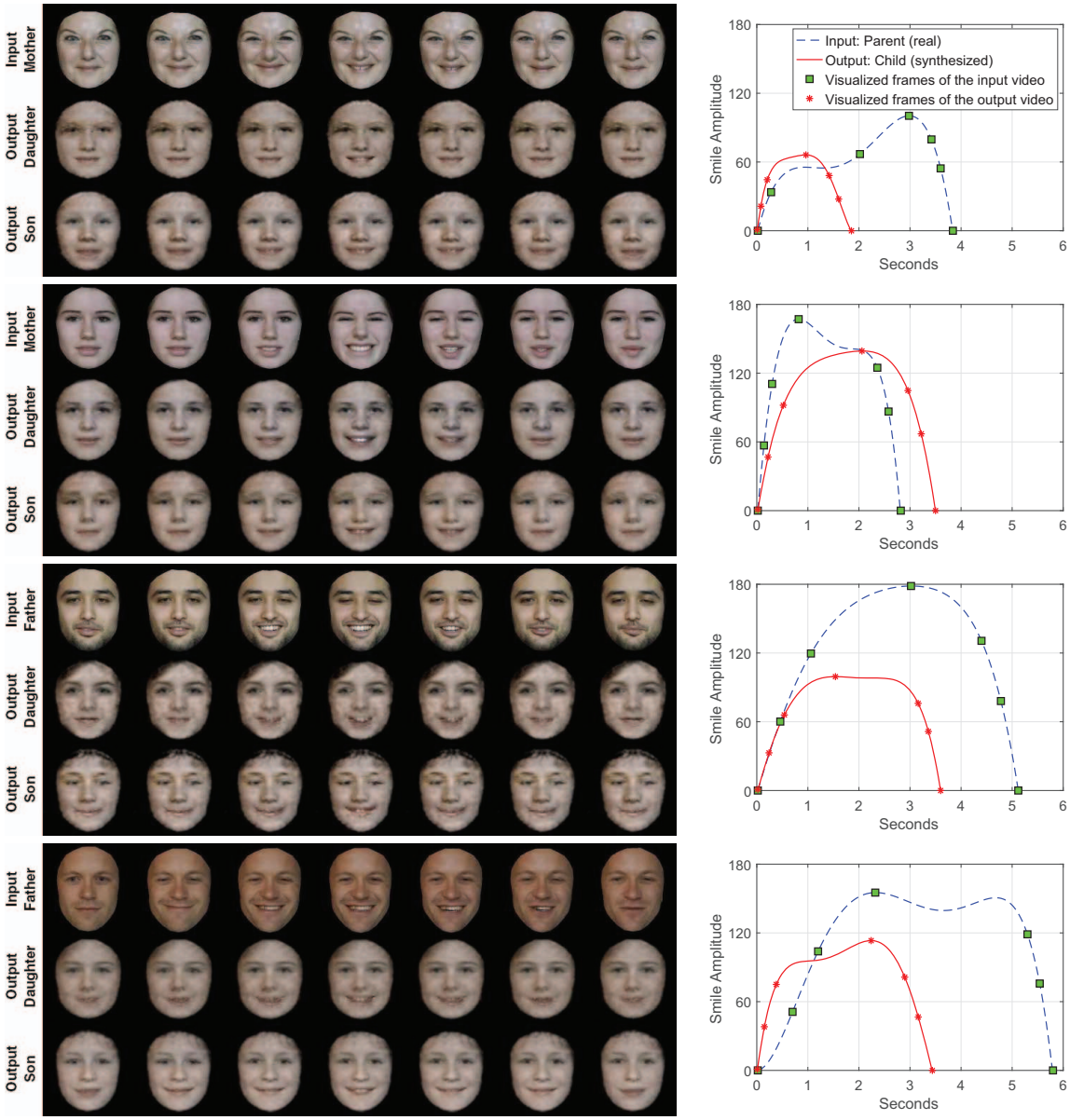


Fig. 5: Samples of input (real) and output (synthesized) videos: (a) Key frames and (b) amplitude signals. Note that the smile amplitude is defined as the first dimension of $\mathcal{R}_{f=\text{mouth}}$

pairs under combined training. This finding suggests that the generated videos of children may be more similar to the parents than the real ones. Next, we visually analyze the obtained videos to validate their quality. As shown in Fig. 5, obtained facial images look quite realistic, and the estimated smile dynamics are meaningful. Thus, we can claim that the proposed method works effectively and it is able generate smile videos of probable children of given parents. Visual demonstration of the synthesis pipeline, and samples of input/output videos can be viewed online[1].

[1] http://visionlab.tudelft.nl/kinship-synthesis

D. Assessment of Different Facial Regions

In this experiment, we evaluate the reliability of different facial regions in terms of temporal appearance and dynamics features. To this end, we trained kinship verification models using the combined set of real and synthetic data. Fig. 6 shows the correct verification rates of using dynamics and temporal appearance of different regions. Results reveal that, mouth (& cheek) region leads to a better verification compared to other regions. This can be explained by the fact that mouth (& cheek) region comprises a large facial area that displays distinct appearance patterns of kinship as well

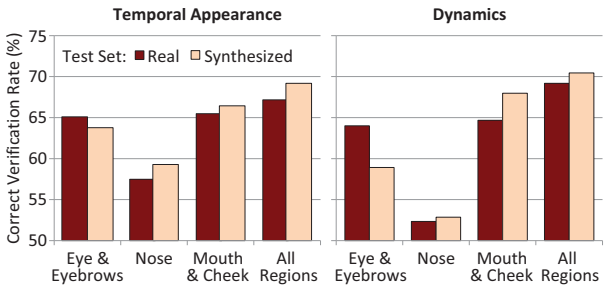


Fig. 6: Accuracy of using temporal appearance and dynamics of different facial regions in kinship verification

as providing most of the dynamic information during a smile. Since nose region is not much affected by smile expression, the use of nose dynamics performs only slightly better than random prediction. Synthesized appearance and dynamics of all regions, except eyes & eyebrows region, perform better than those of real pairs. We can deduce from this finding that our method cannot model the appearance and dynamics of eyes & eyebrows region as accurate as those of other regions.

VI. Conclusion

First time in the literature, we have proposed a kinship synthesis framework that is capable of generating smile videos of probable children of given subjects. As well as synthesizing images using a convolutional encoder-decoder architecture, we model temporal dynamics of expressions, and combine them to synthesize videos of estimated children. We have quantitatively evaluated our synthesized videos in a set of kinship verification experiments. Our results suggest that (1) enhancing training set with synthetic data increases the verification performance; and (2) our proposed method can indeed generate realistic child videos that may even be more similar to the corresponding parent than the real child.

As a future work, we aim to evaluate our method on other facial expressions such as disgust and surprise. Due to data limitations, our models rely solely on the data of a single parent for the synthesis of the probable child. In case of having sufficient data, a further research direction would be to change our network architecture such that the appearance and dynamics of the estimated child are learned from the videos of both mother and father.

References

[1] 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*, vol. 36, no. 2, pp. 331–345, 2014.

[2] H. Yan, "Kinship verification using neighborhood repulsed correlation metric learning," *Image and Vision Computing*, 2016.

[3] X. Qin, X. Tan, and S. Chen, "Tri-subject kinship verification: Understanding the core of a family," *IEEE Trans. on Multimedia*, vol. 17, no. 10, pp. 1855–1867, 2015.

[4] H. Dibeklioğlu, A. Ali Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," in *ICCV*, 2013, pp. 1497–1504.

[5] Y. Guo, H. Dibeklioğlu, and L. van der Maaten, "Graph-based kinship recognition," in *ICPR*, 2014, pp. 4287–4292.

[6] I. Eibl-Eibesfeldt, *Human Ethology*. New York: Aldine de Gruyter, 1989.

[7] G. Peleg, G. Katzir, O. Peleg, M. Kamara, L. Brodsky, H. Hel-Or, D. Keren, and E. Nevo, "Hereditary family signature of facial expression," *PNAS*, vol. 103, no. 43, pp. 15 921–15 926, 2006.

[8] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *ECCV*, 2012, pp. 525–538.

[9] R. Fang, K. D. Tang, N. Snavely, and T. Chen, "Towards computational models of kinship verification," in *ICIP*, 2010, pp. 1577–1580.

[10] G. Guo and X. Wang, "Kinship measurement on salient facial features," *IEEE Trans. on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2322–2325, 2012.

[11] 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*, 2011, pp. 953–956.

[12] 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*, 2012, pp. 725–728.

[13] N. Kohli, R. Singh, and M. Vatsa, "Self-similarity representation of weber faces for kinship classification," in *BTAS*, 2012, pp. 245–250.

[14] S. Xia, M. Shao, and Y. Fu, "Toward kinship verification using visual attributes," in *ICPR*, 2012, pp. 549–552.

[15] A. Puthenputhussery, Q. Liu, and C. Liu, "SIFT flow based genetic fisher vector feature for kinship verification," in *ICIP*, 2016.

[16] H. Yan, J. Lu, and X. Zhou, "Prototype-based discriminative feature learning for kinship verification," *IEEE Trans. on Cybernetics*, vol. 45, no. 11, pp. 2535–2545, 2015.

[17] A. Dehghan, E. G. Ortiz, R. Villegas, and M. Shah, "Who do I look like? Determining parent-offspring resemblance via gated autoencoders," in *CVPR*, 2014, pp. 1757–1764.

[18] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *ACCV*, 2014, pp. 252–267.

[19] H. Yan, J. Lu, W. Deng, and X. Zhou, "Discriminative multimetric learning for kinship verification," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 7, pp. 1169–1178, 2014.

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

[21] M. Ghahramani, W.-Y. Yau, and E. K. Teoh, "Family verification based on similarity of individual family members facial segments," *Machine Vision and Applications*, vol. 25, no. 4, pp. 919–930, 2014.

[22] R. Fang, A. C. Gallagher, T. Chen, and A. Loui, "Kinship classification by modeling facial feature heredity," in *ICIP*, 2013, pp. 2983–2987.

[23] J. P. Robinson, M. Shao, Y. Wu, and Y. Fu, "Family in the wild (FIW): A large-scale kinship recognition database," *arXiv preprint arXiv:1604.02182*, 2016.

[24] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[25] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *CVPR*, 2016.

[26] A. Dosovitskiy, J. Springenberg, M. Tatarchenko, and T. Brox, "Learning to generate chairs, tables and cars with convolutional networks." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2016.

[27] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *NIPS*, 2016, pp. 2802–2810.

[28] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *NIPS*, 2015, pp. 1099–1107.

[29] A. Ghodrati, X. Jia, M. Pedersoli, and T. Tuytelaars, "Towards automatic image editing: Learning to see another you," *arXiv preprint arXiv:1511.08446*, 2015.

[30] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *IEEE AFGR*, 2015.

[31] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[32] H. Dibeklioğlu, F. Alnajar, A. Ali Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Trans. on Image Processing*, vol. 24, no. 6, pp. 1928–1943, 2015.

[33] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *ICCV Workshops*, 2011, pp. 868–875.