Image and Vision Computing 79 (2018) 133–143

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Modeling and synthesis of kinship patterns of facial expressions ☆

Itır Önal Ertuğrul[a,*], László A. Jeni[a], Hamdi Dibeklioğlu[b]

[a] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[b] Department of Computer Engineering, Bilkent University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 16 October 2017
Received in revised form 12 May 2018
Accepted 12 September 2018
Available online 22 September 2018

Keywords:
Kinship synthesis
Kinship verification
Temporal analysis
Facial action units
Facial dynamics

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, for the first time in the literature, we propose a kinship synthesis framework, which generates smile and disgust videos of (probable) children from the expression videos (smile and disgust) of parents. While the appearance of a child's expression is learned using a convolutional encoder–decoder network, another neural network models the dynamics of the corresponding expression. The expression 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. Furthermore, we evaluate the expression similarity between input and output frames, and show that the proposed method can fairly retain the expression of input faces while transforming the facial identity.

1. 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* [2-4]; given a pair of face images, they try to identify whether these two have a kin relationship or not. Contrarily, *kinship recognition* studies aim to classify the type of kin relationship such as father–daughter and mother–son [5].

In addition to general appearance of face, style and appearance of expressions can also be inherited. Facial expressions of congenitally blind and deaf children with phocomelia, 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 the naked eye. In this study, for the 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 are only two kinship video/expression databases (UvA-NEMO Smile [8] and UvA-NEMO Disgust databases [9]) 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/disgust expression 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 paper has been recommended for acceptance by Hatice Gunes.

* Corresponding author.
E-mail address: iertugru@andrew.cmu.edu (I. Önal Ertuğrul).

https://doi.org/10.1016/j.imavis.2018.09.012
0262-8856/

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.

We extend our previous study [10] in many ways. Along with an extended literature, (1) we use intensity of facial action units (AUs) instead of facial landmark displacement for both expression matching and learning temporal dynamics, (2) we model the facial appearance in a holistic manner, rather than learning facial regions individually since a set of AUs can effectively describe the whole face during expression matching, (3) we extend our dataset including the UvA-NEMO Disgust Database [9] and generate disgust videos of children in addition to their smile videos, (4) we enhance the reliability of the kinship verification method used in our experiments, and (5) we perform analyses to evaluate the quality of synthesized expressions in terms of occurrence and intensity of AUs.

2. 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. [11] 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 [12], Spatial Pyramid LEarning-based (SPLE) descriptors [13], Gabor-based Gradient Orientation Pyramid (GGOP) [14], Self Similarity Representation (SSR) [15], semantic-related attributes [16], and SIFT flow based genetic Fisher vector feature (SF-GFVF) [17]. Moreover, a prototype-based discriminative feature learning (PDFL) method has been proposed [18], and a gated autoencoder is trained to characterize the similarity between faces of parents and children for kinship verification [19]. Metric learning has also been adopted for kinship verification problem in various studies [1,2,20-24]. Differently, in [25] a hierarchical representation learning framework is presented for kinship verification. In addition, multi-view heterogeneous similarity learning has been proposed to learn and predict gender-unknown kin relations [26]. [27] has constructed a topological cubic feature space by kernelized bidirectional PCA for age-aware facial kinship verification. Furthermore, a genetic similarity measure between child–parent pairs is learned in an ensemble learning framework [28].

Beside one-to-one kinship verification, a number of studies focus on verification or recognition of kin relations in family images [5,3,29-32]. They predict whether a face image has kin relation with multiple family members [29], classify given a query face image which family it belongs to [30,31], perform tri-subject kinship verification using the core parts of a family including mother–father pair to verify the kinship of child [3], recognize the exact type of kin relation in family photos [5], and employ a denoising autoencoder based marginalized metric learning for kinship verification on families in the wild [32]. 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 so 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 [33] 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, in [34], convolutional encoder–decoder architecture is combined with an iterative learning approach for medical image segmentation. Additionally, [35] uses a fully convolutional encoder–decoder network for contour detection. In [36], a generative up-convolutional neural network has been proposed to re-generate images of objects for a given object style, viewpoint, and color.

In [37], 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. [38] 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. Moreover, the model proposed in [39] is a novel recurrent encoder–decoder architecture that estimates facial landmarks from videos for sequential face alignment. A different encoder–decoder architecture has been proposed in [40] to modify facial attributes such as including glasses or a hat on a given face image. In [41], a convolutional encoder–decoder architecture is proposed for one-step time-dependent future video frame prediction.

3. Method

In this paper, we propose to model relations of facial appearance and dynamics between smile/disgust expressions of parent–child pairs, and combine them to synthesize a smile/disgust expression of the probable/future child of a given subject. To generate such videos, we use a single video of reference subjects as input (parent) data. To train our models, smile and disgust videos of parent–child pairs are used. Our method requires complete smile and disgust expressions 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 and disgust since they are frequently performed basic facial expressions [42].

In this section, details of the proposed method are described. The flow of the method is as follows. Dense facial landmarks are tracked during smile and disgust videos and are used to normalize faces. Action unit (AU) occurrences and intensities are estimated for each frame. Using the AU intensities, 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 expression dynamics of parent–child pairs based on the extracted AU intensity values over time. Once both networks are trained, or smile/disgust dynamics of the most probable child (based on the model) of a given subject (reference parent) is estimated. Afterwards, dynamics of the reference parent is transformed to that of the estimated child by re-ordering frames of the parent video. The modified video with smile or disgust expression has the appearance of the given subject but the temporal dynamics of the estimated child. Finally, 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.

3.1. Facial landmark tracking and alignment

To normalize face images in terms of rotation and scale, and to measure regional deformations in face, we track 1024 dense facial landmarks as shown in Fig. 1 (a). To this end, we use a state-of-the-art tracker proposed by Jeni et al. [43]. The tracker employs a combined 3D supervised descent method [44], where the shape

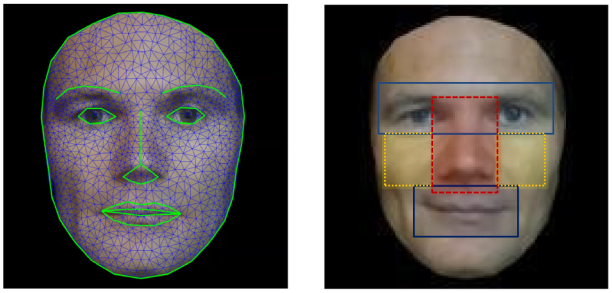

Fig. 1. (a) Normalized/cropped face image, the tracked landmarks, and (b) the defined patches on eyes & eyebrows, nose, mouth, and cheek regions for kinship verification experiments.

model is defined by a 3D mesh and the 3D vertex locations of the mesh [43]. A dense parameterized shape model is registered to an image such that its landmarks correspond to consistent locations on the face. The accuracy and robustness of the method for 3D registration and reconstruction from 2D video was validated in a series of experiments in [43].

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 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.

3.2. Estimating action unit (AU) occurrence and probability

To automatically code 19 facial action units we use the pre-trained large-margin classifier of Girard et al. [45]. Using the tracked landmarks the system registers each video frame to a canonical view with the size of the face normalized to have an inter-ocular distance of 100 pixels. It extracts HOG descriptors [46] around 49 landmarks corresponding to the eyes, eyebrows, nose and lip regions using 64 × 64 pixel patches divided into 16 cells and 8 orientation bins. Features are normalized to have zero mean and unit variance, and two-class linear SVMs are used for each AU. The system provides binary values for action unit occurrences and their corresponding continuous outputs from the SVMs. Action unit probabilities are computed based on the continuous SVM outputs with Platt scaling [47].

3.3. Learning temporal dynamics

While occurrence of an AU or combination of AUs in a frame may be an indicator of an expression, intensity of AUs can reflect the strength of an expression. Many studies [48-52] have used classifier decision values as estimates of expression intensity. Following these studies, we use classifier posterior probabilities of existence of AUs for each frame as AU descriptors, which reflect the AU intensities. Let $\mathcal{A}_{u,t}$ denote the uth action unit probability estimated for frame t. Then, each frame t is represented by a $\mathcal{U}$-dimensional vector $\mathcal{A}_t = \{\mathcal{A}_{u,t} \mid u \in \{1,2,\ldots\mathcal{U}\}\}$, where $\mathcal{U}$ represents the number of AUs, whose probabilities are estimated.

Since $\mathcal{A}_{u,t}$ is a frame-based descriptor, temporal dynamics of each AU during an expression (smile or disgust) of T frames can be represented by a $\mathcal{U}$-dimensional time series with length of T. Note that, a few AUs are observed during smile and disgust expressions and the remaining ones do not carry significant information. Therefore, we reduce the dimensionality of $\mathcal{A}$ to d using PCA by retaining 90% of the variance. The resulting reduced time series (obtained from AU probabilities estimated from whole face) is hereafter referred to as $\mathcal{R}$. The ratio of the explained variance by each dimension (component) $q \in \{1,2,\ldots,d\}$ of $\mathcal{R}$ is Λ_q. Notice that we keep all AUs together to effectively describe facial expressions, and then we apply PCA to reduce the dimensionality.

Duration of expressions varies in length (T). Yet, we need to represent dynamics of varying-length smile/disgust expressions by a fixed-length descriptor since we do not employ temporal models. To this end, we fit a separate pth-degree polynomial to each dimension of time series $\mathcal{R}$ obtained from AU descriptors of whole face. Notice that each column vector (dimension) of $\mathcal{R}$ can be considered as $g(t) = y_t$, where $\forall t \in L = \{1,2,\ldots T\}$, and polynomials can be fit to these functions. Yet, 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 $(p + 1).d + 5$ dimensional feature vector is obtained.

Once the features are computed, the mapping between expression 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 and Disgust databases do 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 distinct polynomials along with the length (T) of the time series. Using these estimates, reduced time series obtained from AU descriptors of whole face ($\mathcal{R}^{\text{child}}$) for the corresponding child can be reconstructed.

3.4. Learning appearance

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 action units.

We use per-frame AU descriptors described in Section 3.3 to obtain a matching child frame t^* for each video frame t of the corresponding parent as follows:

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

where T' denotes the length (number of frames) of the child's video.

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.

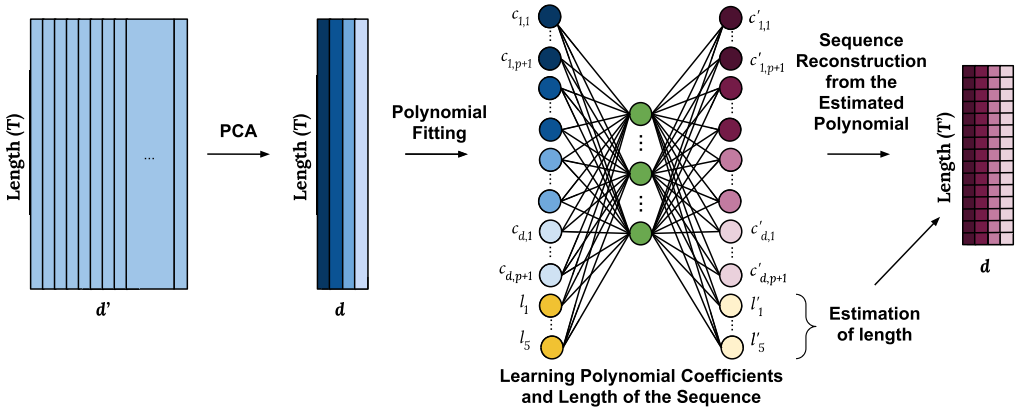

Fig. 2. An illustration of learning temporal dynamics.

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.

A separate appearance model is learned for each of the mother–daughter, mother–son, father–daughter, and father–son relations using smile and disgust expressions together. 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 $r = 1/(W \times H \times U)$, and W is the width and H is the height of the filter. U denotes the number of input planes.

3.5. Expression synthesis

This section explains how we use the models of dynamics and appearance to generate a smile or disgust video of the estimated child of a given subject. After computing the expression dynamics of an estimated child, we transform the dynamics of the parent ($\mathcal{R}^{\text{parent}}$) to that of the estimated child ($\mathcal{R}^{\text{child}}$) by re-ordering the frame sequence of the parent.

Let $I^{\text{parent}}_{s^{\text{parent}}}$ denote the image sequence of face 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 $\mathcal{R}$ is a time series of per-frame AU descriptors $\mathcal{R}_t$ with a reduced dimensionality of d (see Section 3.3), where the qth dimension of $\mathcal{R}_t$ can be shown as $\mathcal{R}_{t,q}$. Then, a re-ordered sequence $\hat{s}$ can be obtained ensuring that $\mathcal{R}^{\text{parent}}_{\hat{s}} \simeq \mathcal{R}^{\text{child}}_{s^{\text{child}}}$ using Algorithm 1. Note that the first dimension of $\mathcal{R}$ ($\mathcal{R}_{s,q=1}$) can be thought as the amplitude signal of the expression, since it explains the majority of the variance of $\mathcal{A}$. Thus, if the image sequence of the estimated child displays expressions with higher amplitudes than that of the parent, we reduce the values of $\mathcal{R}^{\text{child}}$ 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 expressions of the parent and the estimated child. Length of $\mathcal{R}^{\text{child}}$ is accordingly reduced using bicubic interpolation to preserve the temporal dynamics such as speed and acceleration of change in $\mathcal{R}^{\text{child}}$. Afterwards, each frame of the re-ordered image sequence $I^{\text{parent}}_{\hat{s}}$ of the parent is transformed to that of the estimated child using the learned convolutional model (Section 3.4) as visualized in Fig. 4.

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

Require: $\mathcal{R}^{\text{parent}}$ of size $T_{\text{parent}} \times d$
Require: $\mathcal{R}^{\text{child}}$ of size $T_{\text{child}} \times d$
Require: Explained ratio of $\mathcal{A}$'s variance (Λ_q) by each dimension $q \in \{1, 2, \ldots, d\}$ of $\mathcal{R}$ (see Section 3.3)
Ensure: $\mathcal{R}^{\text{parent}}_{\hat{s}} \simeq \mathcal{R}^{\text{child}}_{s^{\text{child}}}$

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

4. 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 [8] and Disgust [9] databases, which are the only available kinship video databases in the literature. In our previous study [10], we have evaluated our model solely on UvA-NEMO Smile database. In the current study, to show the generalizability of the proposed method for other facial expressions, we include an evaluation on UvA-NEMO Disgust database. While some studies do no recognize disgust as a universal expression due to its cultural variability [53],

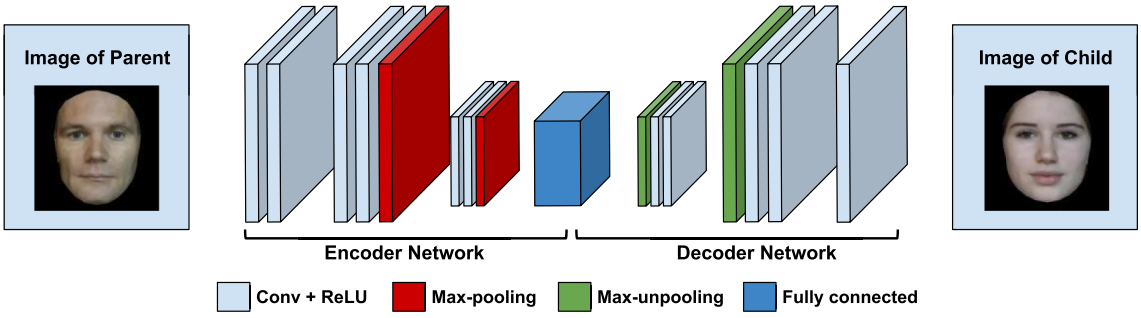

Fig. 3. An illustration of the convolutional encoder–decoder network that models the appearance transformation.

a large number of studies have shown that it is one of the six basic emotions [42,54,55]. Recognizability of an expression across cultures is not a need for synthesizing child videos from parents' videos. Our method relies on the empirical findings that disgust and smile expressions display hereditary patterns [4,7] which are aimed to be modeled in this study.

The kinship dataset has spontaneous and posed enjoyment smiles and posed disgust expressions 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 smile video pairs and posed disgust 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, and one or two disgust expressions. By using different video combinations of each kin relation, 159 pairs of spontaneous smile videos and 151 pairs of posed disgust 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, video and matched frame pairs, and parent videos for each kin relationship are given in Table 1.

5. Experiments & results

Our method aims to synthesize smile and disgust expressions 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 or disgust expressions to train a spatio-temporal kinship verification system, and evaluate our method based on the obtained results. To this end, we employ 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 [56] from the regions eyes & eyebrows, cheeks, and mouth to describe regional appearance over time. Regional features are concatenated as an appearance feature vector. In [4], a set of statistical descriptors are extracted from the displacement signals of eyelids & eyebrows, cheeks, and lip corners to represent temporal dynamics of expressions, and these descriptors are 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

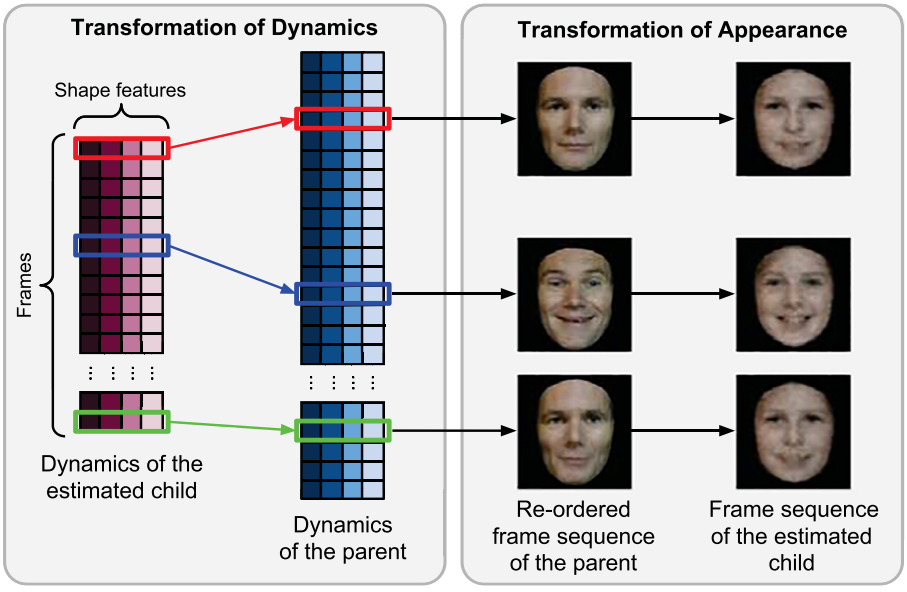

Fig. 4. Generation of the image sequence (whole face) of the estimated child.

Table 1
Distribution of subject, video and matched frame pairs and number of parent videos used in our experiments.

Relation	Smile				Disgust			
	Pairs			Parent videos	Pairs			Parent videos
	Subject	Video	Matched frame		Subject	Video	Matched frame	
Mother–daughter	16	57	6042	29	16	52	3359	28
Mother–son	12	36	4291	21	12	42	2257	22
Father–daughter	9	28	3371	16	7	16	1150	10
Father–son	12	38	4564	21	11	44	2425	21
All	49	159	18,268	87	46	151	9191	81

the regions of eyes & eyebrows, nose, and mouth & cheeks (see Fig. 1 (b)). We also compute LBP features for the first and the last frame of the expression onset (i.e., neutral face and expression peak, respectively). LBP descriptors are extracted from 8×8 non-overlapping (equally-sized) blocks on the face. In our implementation, dynamics features are extracted from the AU-based time series $\mathcal{A}$. A separate classifier (SVM) is modeled for each of these feature sets (CLBP, LBP, and dynamics). A weighted SUM rule is used to fuse the computed posterior probabilities for the target classes of these classifiers. Other details are kept same with those of the original method [4].

Kinship sets of the UvA-NEMO Smile and UvA-NEMO Disgust databases, and the generated smile/disgust expressions 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. 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.

Our experimental protocol uses 3 different test configurations. (1) Actual condition test, in which we train and test the system with either real or synthesized videos to obtain actual verification performance. (2) Cross-condition test, in which we train the system with real videos and test it with synthesized videos or vice versa to analyze whether these real and synthesized sets are related. (3) Combined test, in which we employ combination of real and synthesized videos to train the system. Recall that, one of the main motivations of this work is that, our model can be used to create a synthetic database, which can be used for further research. Therefore, we used obtained synthetic database during training to explore its significance in providing better generalizability and better kinship verification accuracy. 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 facial appearance of the estimated children, we train appearance transformation models for each kin relationship, i.e., M–D, M–S, F–D, F–S. 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 and disgust signals with continuous exponential increase. Dynamics network is trained using all kin relationships for smile and disgust expression separately due to the limited number of video pairs of each kin relationship. In the remainder of this section, the results of our kinship verification and AU similarity experiments will be presented.

5.1. Assessment of the synthesized appearance

In this experiment, we aim to assess the static appearance quality of the synthesized faces. To this end, we solely employ LBP features extracted from the first (neutral face) and the last frame (expressive face) of the onset phase of the expressions (smile and disgust). Using the extracted LBP descriptors, for smile and disgust expressions (separately), 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 2, the static appearance features extracted from the synthesized smile videos achieve an accuracy of 62.60% when the verification model is trained on the synthesized set, which is about 1.4% (absolute) higher than that of the real smile video pairs when the model is learned on real data. Differently, training and testing the model with real disgust videos (58.83%) and with synthesized disgust videos (58.29%) provide nearly the same verification accuracy. Furthermore, if the system is trained on the real video pairs, only 2.5% and 2.9% accuracy decreases are observed for the synthesized smile and disgust videos, respectively. 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)). Moreover, training the system by using real and synthesized smile videos together, increases the accuracy for both real (3.3%) and synthesized pairs (3.2%). Under this setting, synthesized smile videos perform 1.3% better than real ones. Similarly, the synthesized disgust videos perform 0.5% better than the real ones when the model is trained with real and synthesized disgust videos. These findings show that indeed the obtained synthetic data can be used to train a more accurate kinship verification system.

5.2. Assessment of the synthesized dynamics

Similar to the previous experiment, we conduct cross-database experiments using real and synthesized smile and disgust 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 3, for both smile and disgust, estimated expression dynamics performs slightly worse than the dynamics of real expressions when the system is trained with real videos. Once we use synthesized smile and disgust dynamics along with the real ones to train the model, 4.9% and 4.7% accuracy increases are obtained for real pairs compared to the model trained using solely real samples. Moreover, synthetic smile and disgust samples perform better than real ones when the model is learned on the combined data. As in the previous experiment, these findings show the efficacy of our

Table 2
Accuracy (%) of using real and synthesized static appearance in kinship verification.

Training set	Test set			
	Smile		Disgust	
	Real	Synth.	Real	Synth.
Real	61.17	59.06	58.83	55.42
Synth.	58.25	62.60	55.13	58.29
Real + synth.	64.51	65.85	61.93	62.48

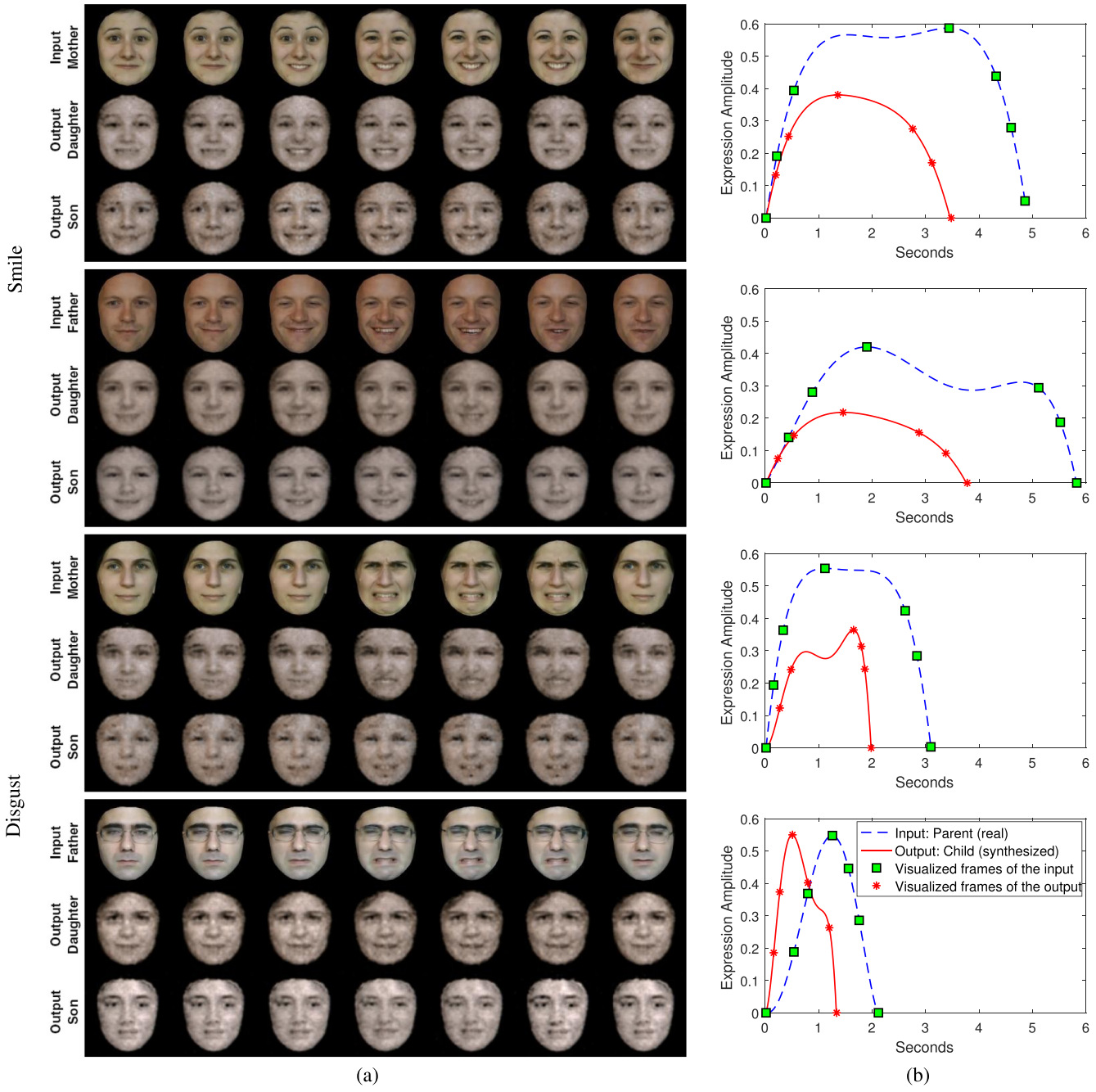

Fig. 5. Samples of input (real) and output (synthesized) videos: (a) Key frames and (b) amplitude signals. Note that the expression amplitude is defined as the first dimension of $\mathcal{R}$.

method as well as indicating the importance of using synthetic data in addition to real samples during the training of kinship verification models.

Table 3
Accuracy (%) of using real and synthesized temporal dynamics in kinship verification.

Training set	Test set			
	Smile		Disgust	
	Real	Synth.	Real	Synth.
Real	68.35	66.92	65.49	64.69
Synth.	66.03	71.01	62.82	68.78
Real + synth.	73.32	72.15	70.28	67.07

5.3. Combining appearance and dynamics

In this experiment, we first use the spatio-temporal features (CLBP-TOP) extracted from the regions of eyes & eyebrows, nose, and mouth (over videos) for kinship verification. We perform experiments using real and synthesized smile and disgust video pairs to evaluate the reliability of the estimated temporal appearances. Results in Table 4 reflect that, when the system is trained with real and synthesized videos, synthesized smile videos perform %2.9 better compared to real smile videos while synthesized disgust videos perform %1.2 better than the real disgust videos, showing that our model is capable of generating child videos which are more similar to the parents than their real children. We believe that, we obtain that result since child videos are synthesized using a single parent video.

Table 4
Accuracy (%) of using real and synthesized temporal appearance in kinship verification.

Training set	Test set			
	Smile		Disgust	
	Real	Synth.	Real	Synth.
Real	68.92	66.06	66.22	63.46
Synth.	63.69	71.95	61.39	67.99
Real + synth.	73.48	76.41	70.54	71.78

Table 5
Accuracy (%) of the combined use of static and temporal appearance and dynamics of real and synthesized videos in kinship verification.

Training set	Test set			
	Smile		Disgust	
	Real	Synth.	Real	Synth.
Real	76.80	74.55	72.92	70.80
Synth.	74.96	80.80	71.80	75.35
Real + synth.	81.71	84.92	76.46	79.01

Moreover, when the system is tested with real videos, training the model with real and synthesized videos leads to the best verification accuracy for smile (73.48%) and disgust (70.54%) videos. These results also show the significance of the synthesized dataset and support the reliability of our method.

In the final verification experiment, we combine static appearance, temporal appearance, and temporal dynamics features and use in the verification system to assess the full performance of the synthesized smile and disgust videos in kinship verification. As shown in Table 5, when the system is trained solely on real samples, the accuracy of employing real samples reaches 76.8% for smile, and 72.92% for disgust where the accuracy for synthetic videos is only 2.2% and 2.1% less for smile and disgust, respectively. Verification accuracy for real pairs are enhanced by 4.9% (absolute) for smile and 3.5% for disgust by including the synthesized samples in the training set. Moreover, synthesized videos perform better than the real 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 and disgust dynamics are meaningful. Thus, we can claim that the proposed method works effectively and it is able to generate smile and disgust videos of probable children of given parents. Our obtained images are not that crispy since the architecture could only learn general facial patterns rather than appearance details. We observe that, when an expression causes a significant change in the facial appearance of parents, the corresponding child frames also display a significant change in the appearance as shown in Fig. 5.

Notice that, we obtain a higher kinship verification accuracy on the smile dataset compared to the disgust dataset, which can be explained by the fact that (1) there are fewer matched frames in the disgust database to train the model, and (2) the variability of the posed disgust expression among individuals is much higher compared to that of smiles.

We compare the kinship verification accuracy values obtained using facial landmarks [10] and AU intensities during expression matching and synthesis of dynamics on smile database in Fig. 6. We perform kinship verification experiments using real frames, synthesized frames or their combination as training set and testing with real or synthesized frames. Moreover, we compare the results obtained with static appearance, temporal dynamics and their combination. Results reflect that, when only static appearance features are employed, we obtain the lowest verification accuracy and using facial landmarks performs better compared to using AU intensity during expression matching and synthesis. On the other hand, when temporal dynamics and combination of appearance and dynamics are employed, using AU intensity outperforms facial landmarks. We can infer that, among all configurations, using AU intensity in the architecture and employing combination of static appearance and temporal dynamics in verification experiments leads to the best accuracy.

Note that, when landmarks are used, expression matching is performed using the local landmarks around mouth, nose and eye & eyebrow regions separately. Since a set of AUs can effectively describe the whole face, employing AU intensities during expression matching leads to a more accurate analysis.

5.4. Analysis of AU occurrences

In order to identify the most frequent AUs appearing in the parents' smile and disgust videos, we compute base rates averaged over all input parent videos. The base rate of an AU equals the ratio of the number of frames containing the corresponding AU to the total number of tracked frames. In Table 6, we highlight four rows containing the AUs with the highest base rates for the smile and disgust datasets, separately. Please see Fig. 7 for the visualization of these AUs.

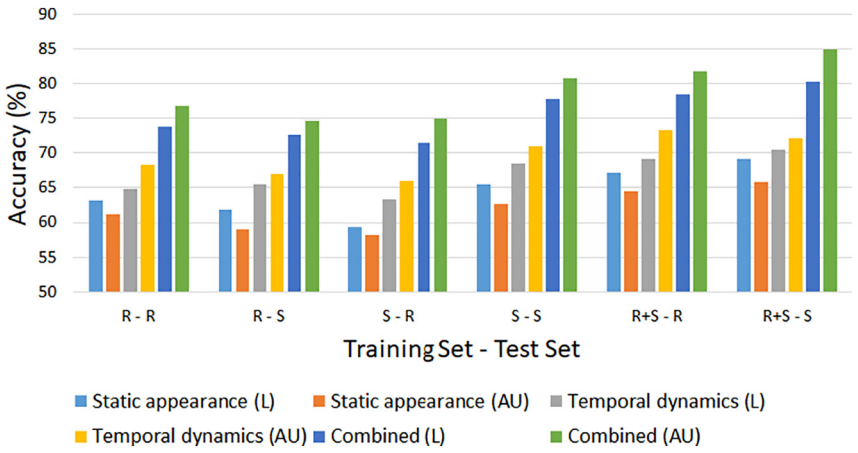

Fig. 6. Comparison of using facial landmarks and AU intensity values in kinship verification experiments performed with smile videos. R denotes real frames, S denotes synthesized frames and R + S denotes combination of real and synthesized frames.

Table 6
Base rates (%) of AU occurrences, f-scores and correlations computed using the AU probabilities of "real parent"–"synthesized child" pairs. AUs with the highest base rates are highlighted for smile and disgust datasets.

AU	Smile			AU	Disgust		
	Base rate (%)	f-Score	Corr		Base rate (%)	f-Score	Corr
1	1.23	0.01	0.01	1	1.34	0.00	0.07
2	22.39	0.05	0.00	2	11.16	0.04	0.08
4	12.71	0.00	−0.01	4	48.38	0.11	0.20
5	7.80	0.00	0.15	5	4.34	0.00	0.32
6	**82.99**	**0.55**	**0.48**	**6**	**59.24**	**0.49**	**0.57**
7	**88.25**	**0.41**	**0.22**	**7**	**73.23**	**0.47**	**0.37**
9	9.89	0.02	0.06	9	41.89	0.14	0.31
10	**71.10**	**0.30**	**0.45**	10	29.07	0.16	0.48
11	53.61	0.26	0.14	11	32.83	0.25	0.40
12	**87.22**	**0.56**	**0.55**	12	48.23	0.44	0.54
14	11.46	0.01	0.05	14	6.27	0.02	0.04
15	6.26	0.00	0.06	15	34.08	0.09	0.17
17	49.23	0.36	0.26	**17**	**71.64**	**0.29**	**0.12**
18	0.90	0.00	0.28	18	4.53	0.08	0.24
19	38.46	0.31	0.09	19	52.73	0.39	0.10
22	2.60	0.01	0.14	22	9.46	0.02	0.04
23	57.78	0.45	0.15	**23**	**61.68**	**0.44**	**0.10**
24	31.97	0.41	0.11	24	39.60	0.30	−0.01
28	28.41	0.21	0.13	28	42.30	0.24	0.07

The results reported in Table 6 show that the most frequent AUs displayed in smile videos are AU6, AU7, AU10, and AU12. This outcome is consistent with the finding that enjoyment smiles contain AU6 and AU12. On the other hand, the most frequent AUs which appear in disgust videos are AU6, AU7, AU17, and AU23, while the disgust expression is expected to contain AU9, AU15, and AU16 (note that AU16 is not tracked by our system). From Table 6 we can see that AU9 and AU15 occurrences are observed with relatively low base rates. Since the disgust videos are posed, the variability of facial surface deformations during these videos is quite high among different parents. Therefore, AU9 and AU15 in the disgust videos are not displayed as frequently as AU6 and AU12 in the smile videos. When we compare the highest base rates of AU occurrences for smile and disgust videos, we can see that the base rates of AU6 and AU12 observed in smile videos are much higher than those of AU7, AU17 and AU23 observed in disgust videos.

5.5. Assessment of AU similarity

We match expressions of kin pairs using AU probabilities since intensities of a set of AUs can effectively describe facial expressions. As a result, the input (parent) and output (child) images of the appearance network should have similar facial expressions. Thus, based on the fact that the appearance network would tend to learn more frequent expressions better, we would expect a higher reliability for the expressions that have higher (occurrence) base rates in the database. Therefore, if a parent (input) frame contains a frequently displayed AU, it should also be observed in the (corresponding) synthesized child frame. In order to evaluate the expression similarity between input–output pairs, we employ the automatically coded probabilities of 19 AUs. For each AU u, we compute the normalized f-scores using the AU occurrence vector (O_u^{parent}) of the frames of parent and AU occurrence vector ($\hat{O}_u^{child}$) of the corresponding frames of generated child. Since the base rate of each AU are different, and AU occurrences are skewed in our dataset, we compute the normalized f-score [57]. The normalized f-score represents the f-score value that would be obtained if the data (AU occurrences) were balanced. After obtaining normalized f-score values for each video, we compute average f-score values over all smile and disgust videos in the dataset, separately (see Table 6).

Estimates of AU occurrences may be noisy. For instance, an AU with a probability of 0.49 is labeled as *not occurred* while the one with a probability of 0.51 is labeled as *occurred*. Therefore, AU probabilities can provide more detailed information compared to the sole use of AU occurrences. Thus, we also compute the Pearson correlation coefficients between AU probabilities of (real) parent and (synthesized) child frame sequences.

From f-score values and correlation coefficients reported in Table 6, we can infer that our model is more successful to learn, represent, and synthesize AU6 and AU12 compared to other AUs for smiles. In other words, for a parent frame containing these AUs, our model is likely to generate a child frame having the same AUs. Notice that, for AU1, AU4, AU5, AU9, AU14, AU15, AU18, and AU22, the obtained f-score values and correlation coefficients are quite low because these AUs are very rare in the input (parent) videos.

While AU2, AU11, AU19, AU24, and AU28 are fairly observed in the dataset, f-score values and correlation coefficients for these AUs are not comparable with those of AU6, AU7, AU10, AU12. This can be explained based on the fact that AU6, AU7, AU10, and AU12 are displayed in most frames and, thus, they can be modeled much more effectively. For AU7, the most frequently displayed AU in the smile videos, we obtain lower f-score compared to AU6 and AU12, meaning

Fig. 7. Action units having the highest base rates in our smile and disgust datasets. Images are from CK+ dataset, ©Jeffrey Cohn.

that our model cannot learn to capture and/or synthesize AU7 as good as AU6 and AU12.

When we analyze the results for the disgust expression, it is seen that the highest f-scores are obtained for AU6, AU7, and AU23 (see Table 6). Notice that, these AUs are among the most frequently observed ones in the disgust dataset. On the other hand, the highest correlations are obtained for AU6, AU10, and AU12. Although the AUs accounting for the disgust expression (AU9 and AU15), are fairly observed in the disgust dataset, their base rates are lower compared to the others so that our model cannot learn to synthesize them well, resulting in lower f-scores and correlations.

From the results in Table 6 we can infer that, our encoder–decoder architecture can effectively model and synthesize AUs that are frequently displayed in the training data.

Moreover, the quality of the synthesized smile frames are better compared to that of disgust in terms of the AU similarity between input–output pairs. This outcome can be explained by higher variability of the posed disgust expression (compared to that of smiles) and by significantly less number of matched frames for disgust (see Table 1 compared to the smiles.

6. Conclusion

We have proposed a kinship synthesis framework that is capable of generating smile and disgust 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 and AU similarity experiments. Our results suggest that (1) the proposed appearance network can conserve the facial expressions (in terms of AUs) of the input frames while transforming the facial identity; (2) enhancing training set with synthetic data increases the kinship verification performance; and (3) 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. 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. In addition, generative models such as Generative Adversarial Networks (GANs) can be used to synthesize child appearance.

References

[1] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, J. Zhou, Neighborhood repulsed metric learning for kinship verification, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 331–345.

[2] H. Yan, Kinship verification using neighborhood repulsed correlation metric learning, Image Vis. Comput. (2016).

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

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

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

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

[7] G. Peleg, G. Katzir, O. Peleg, M. Kamara, L. Brodsky, H. Hel-Or, D. Keren, E. Nevo, Hereditary family signature of facial expression, PNAS 103 (43) (2006) 15921–15926.

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

[9] H. Dibeklioğlu, F. Alnajar, A. Ali Salah, T. Gevers, Combining facial dynamics with appearance for age estimation, IEEE Trans. Image Process. 24 (6) (2015) 1928–1943.

[10] I.Ö. Ertugrul, H. Dibeklioğlu, What will your future child look like? Modeling and synthesis of hereditary patterns of facial dynamics, Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE. 2017, pp. 33–40.

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

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

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

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

[15] N. Kohli, R. Singh, M. Vatsa, Self-similarity representation of Weber faces for kinship classification, BTAS, 2012. pp. 245–250.

[16] S. Xia, M. Shao, Y. Fu, Toward kinship verification using visual attributes, ICPR, 2012. pp. 549–552.

[17] A. Puthenputhussery, Q. Liu, C. Liu, SIFT flow based genetic fisher vector feature for kinship verification, ICIP, 2016.

[18] H. Yan, J. Lu, X. Zhou, Prototype-based discriminative feature learning for kinship verification, IEEE Trans. Cybern. 45 (11) (2015) 2535–2545.

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

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

[21] H. Yan, J. Lu, W. Deng, X. Zhou, Discriminative multimetric learning for kinship verification, IEEE Trans. Inf. Forensics Secur. 9 (7) (2014) 1169–1178.

[22] H. Yan, J. Hu, Video-based kinship verification using distance metric learning, Pattern Recogn. (2017).

[23] H. Yan, Kinship verification using neighborhood repulsed correlation metric learning, Image Vis. Comput. 60 (2017) 91–97.

[24] J. Lu, J. Hu, Y.-P. Tan, Discriminative deep metric learning for face and kinship verification, IEEE Trans. Image Process. 26 (9) (2017) 4269–4282.

[25] N. Kohli, M. Vatsa, R. Singh, A. Noore, A. Majumdar, Hierarchical representation learning for kinship verification, IEEE Trans. Image Process. 26 (1) (2017) 289–302.

[26] X. Qin, D. Liu, D. Wang, Heterogeneous similarity learning for more practical kinship verification, Neural. Process. Lett. (2017) 1–17.

[27] M.M. Dehshibi, J. Shanbehzadeh, Cubic norm and kernel-based bi-directional PCA: toward age-aware facial kinship verification, Vis. Comput. (2017) 1–18.

[28] X. Zhou, Y. Shang, H. Yan, G. Guo, Ensemble similarity learning for kinship verification from facial images in the wild, Inf. Fusion 32 (2016) 40–48.

[29] M. Ghahramani, W.-Y. Yau, E.K. Teoh, Family verification based on similarity of individual family member's facial segments, Mach. Vis. Appl. 25 (4) (2014) 919–930.

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

[31] J.P. Robinson, M. Shao, Y. Wu, Y. Fu, Family in the Wild (FIW): A Large-scale Kinship Recognition Database, arXiv preprint. 2016, arXiv:1604.02182.

[32] S. Wang, J.P. Robinson, Y. Fu, Kinship verification on families in the wild with marginalized denoising metric learning, Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE. 2017, pp. 216–221.

[33] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder–decoder architecture for scene segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2017).

[34] J.U. Kim, H.G. Kim, Y.M. Ro, Iterative deep convolutional encoder–decoder network for medical image segmentation, Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, IEEE. 2017, pp. 685–688.

[35] J. Yang, B. Price, S. Cohen, H. Lee, M.-H. Yang, Object contour detection with a fully convolutional encoder–decoder network, CVPR, 2016.

[36] A. Dosovitskiy, J. Springenberg, M. Tatarchenko, T. Brox, Learning to generate chairs, tables and cars with convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. (2016).

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

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

[39] X. Peng, R.S. Feris, X. Wang, D.N. Metaxas, A recurrent encoder–decoder network for sequential face alignment, ECCV, 2016.

[40] A. Ghodrati, X. Jia, M. Pedersoli, T. Tuytelaars, Towards Automatic Image Editing: Learning to See another You, arXiv preprint. 2015, arXiv:1511.08446.

[41] V. Vukotić, S.-L. Pintea, C. Raymond, G. Gravier, J. Van Gemert, One-Step Time-Dependent Future Video Frame Prediction With a Convolutional Encoder–Decoder Neural Network, arXiv preprint. 2017, arXiv:1702.04125.

[42] P. Ekman, E.R. Sorenson, W.V. Friesen, Pan-cultural elements in facial displays of emotion, Science 164 (3875) (1969) 86–88.

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

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

[45] J.M. Girard, W.-S. Chu, L.A. Jeni, J.F. Cohn, Sayette Group formation task (GFT) spontaneous facial expression database, Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE. 2017, pp. 581–588.
[46] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, CVPR, 1, IEEE. 2005, pp. 886–893.
[47] J. Platt, et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classifiers, 10(3), 1999. pp. 61–74.
[48] M.S. Bartlett, G.C. Littlewort, T. Sejnowski, J. Movellan, A prototype for automatic recognition of spontaneous facial actions, Advances in Neural Information Processing Systems, 2003. pp. 1295–1302.
[49] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE. 2006, pp. 223–230.
[50] S. Koelstra, M. Pantic, Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics, Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, IEEE. 2008, pp. 1–8.
[51] A. Savran, B. Sankur, M.T. Bilge, Regression-based intensity estimation of facial action units, Image Vis. Comput. 30 (10) (2012) 774–784.
[52] K. Shimada, Y. Noguchi, T. Kuria, Fast and robust smile intensity estimation by cascaded support vector machines, Int. J. Comput. Theory Eng. 5 (1) (2013) 24.
[53] R.E. Jack, Culture and facial expressions of emotion, Vis. Cogn. 21 (9–10) (2013) 1248–1286.
[54] M. Biehl, D. Matsumoto, P. Ekman, V. Hearn, K. Heider, T. Kudoh, V. Ton, Matsumoto and Ekman's Japanese and Caucasian facial expressions of emotion (JACFEE): reliability data and cross-national differences, J. Nonverbal Behav. 21 (1) (1997) 3–21.
[55] P. Ekman, W.V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, Universals and cultural differences in the judgments of facial expressions of emotion., J. Pers. Soc. Psychol. 53 (4) (1987) 712.
[56] T. Pfister, X. Li, G. Zhao, M. Pietikainen, Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, ICCV Workshops, 2011. pp. 868–875.
[57] L.A. Jeni, J.F. Cohn, F. De La Torre, Facing imbalanced data—recommendations for the use of performance metrics, Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE. 2013, pp. 245–251.