Vision, Modeling, and Visualization (2018)
F. Beck, C. Dachsbacher, and F. Sadlo (Eds.)

Automatic Infant Face Verification via Convolutional Neural Networks

L. Wöhler[†], H. Zhang[†], G. Albuquerque and M. Magnor

Computer Graphics Lab, TU Braunschweig, Germany

Abstract

In this paper, we investigate how convolutional neural networks (CNN) can learn to solve the verification task for faces of young children. One of the main issues of automatic face verification approaches is how to deal with facial changes resulting from aging. Since the facial shape and features change drastically during early childhood, the recognition of children can be challenging even for human observers. Therefore, we design CNNs that take two infant photographs as input and verify whether they belong to the same child. To specifically train our CNNs to recognize young children, we collect a new infant face dataset including 4,528 face images of 42 subjects in the age range of 0 to 6 years. Our results show an accuracy of up to 85 percent for face verification using our dataset with no overlapping subjects between the training and test data, and 72 percent in the FG-NET dataset for the age range from 0 to 4 years.

Categories and Subject Descriptors (according to ACM CCS): I.5.4 [Pattern Recognition]: Application—Computer Vision

1. Introduction

From automatic tagging of friends on pictures for social networks or the recognition of wanted criminals on surveillance cameras to automated searching for missing children and their identification - in many situations it is necessary to verify whether two images show the same person or not. Due to this strong demand for face recognition and verification applications, many different solutions have been proposed during the last years. Especially the development of new deep learning tools and algorithms has brought the accuracy of automatic face recognition systems to a new level [SCWT14], displaying nearly human-like performance on face recognition benchmarks like LFW [HRBLm], MegaFace [KSMB15], FGNET [fgn] and AgeDB [MPS*17].

However, if it is necessary to verify a person after time has passed and the person has aged, robust age-invariant face recognition and verification is still unreached. Major facial changes include growing of the face, and changes to the shape and appearance which need to be counterbalanced prior or during the recognition process. While aging might already pose problems when recognizing adult subjects, this task can be even more difficult for children and infants. For very young children, facial aging is a complex process that involves substantive facial growing and shape changes. This makes the verification of children which have aged only a few years challenging even for human subjects. Fig. 1 shows an example of major facial changes during early childhood by using a child from the FGNET dataset at the age of 1, 3, and 5, respectively.

Moreover, there is relative few data available to build reasonable models considering children in early childhood. Most publicly available aging datasets are limited in age range and focus on adults or have only a few samples per age, reducing their applicability for deep learning approaches. To the best of our knowledge, the only public domain aging dataset which includes images of children at different ages is the FGNET dataset containing a small number of images from subjects with 12 age-separated images per subject, including 269 images of 75 subjects in the age span of 0 to 6 years. Even considering data augmentation techniques, this amount of data is not sufficient to train a deep neural network.

In this paper we specifically focus on this challenging case and investigate whether a classification network can be used for age-invariant infant facial verification, using facial features learned from a dataset with face images of young children. The problem of age-invariant face recognition was described in previous work using either vision approaches [PTJ10] or deep learning [WLQ16]. As no sufficient dataset for training a deep neural network on infant images was available, previous work mainly focused on developing age-invariant features. For our CNN, we adopt the DeepID2 [SCWT14] network to extract facial features which are directly fed into a classification network to decide whether two presented images show the same child or not. For training, we create a novel dataset which contains 4528 face images of 42 children in the age range of 0 to 6 years and for each child include images at different ages. We split the dataset into a non-overlapping

[†] These authors contributed equally to this work

DOI: 10.2312/vmv.20181253

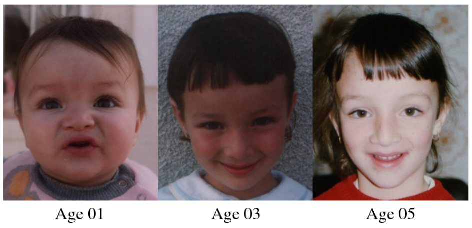

Figure 1: *Cross-age face images for one of the subjects in the FGNET dataset [fgn] in the ages of 1, 3, and 5 years respectively.*

test and training set for our CNN. For classification, we investigate two different classification architectures based on the very successful ResNet [HZRS16] and GoogLeNet [SLJ*15] architectures. We train the combination of feature extraction and classification network end-to-end.

The main contributions of this paper are:

- A novel network structure combining feature extraction based on DeepID2 with a classification part which is trained end-to-end on our new infant face dataset.
- Investigation on two different classification approaches based on state-of-the-art classification architectures.
- Overall our proposed network trained on infant faces is competitive with state-of-the-art systems in the proposed age group of 0 to 6 years. Our network yields an accuracy of 85.3% on the test set of our child face dataset, and an accuracy of 72.6% on the child images of FGNET dataset for the age range from 0 to 4 years.

2. Related Work

Many approaches related to face aging have been proposed in the past, including age synthesis [RC06, GMP*06, FGH10, SMZ*07], prediction [SSSB07], as well as age-invariant [LXZ*16, WLQ16] or age-restricted [TSS12,BBVS16] classification. In this section we focus on approaches using convolutional neural networks that are more closely related to our paper. We begin with a short overview on traditional face classification and the approaches we used in our architecture and continue with neural network methods that focus on age invariant classification.

In recent years deep learning and especially convolutional neural networks have significantly improved the quality of face verification applications and systems. Initially, most methods focused on recognition tasks without considering cross-age applications. Among these deep learning works, a discriminative deep metric learning method [HLT14] was presented for face verification that aimed to find a Mahalanobis distance metric to maximize the interclass variations and minimize the intra-class variations. Taigman et al. [TYRW14] introduced a multi-stage method also called DeepFace that aligns faces to a general 3D shape model and trains a multi-class Siamese Network [CHL05] to optimizes the Euclidean distance between two facial features. FaceNet [SKP15] minimizes the deep triple facial metrics by learning the distance between the positive pairs and negative pairs. In addition, using 128 bytes per face the performance of the method is cost efficient. An innovative loss function called center-loss was proposed in [WZLQ16], that efficiently increases inter-class dispersion and intra-class compactness. In [SWT14, SCWT14] facial feature extractors based on CNNs, named DeepID and DeepID2 respectively, were introduced. Especially DeepID2 aimed to enlarge the inter-personal variances extracted from different identities and reduced the intra-personal variances extracted from the same identity. While all of these approaches greatly contributed to face recognition, we additionally focus on age-invariant recognition for children.

Recently, deep learning models were also used for age-invariant recognition tasks. Usually, approaches for this task either extract information from face images to build age-invariant face features or build age-invariant face models for the matching decision. Focusing on age-invariant features, Wen et al. [WLQ16] proposed a latent factor guided CNN framework to directly learn age-invariant deep face features. The authors analyzed the results for different age groups. While they achieved good performance, the results reported for young children from 0 to 4 years were worse than for older age groups. Other work focused on deep learning age-invariant features by using an age-estimation step to remove aging factors from the extracted facial features [ZDH17] or utilized multi-task learning to improve extracted features [WZK*17]. In contrast to the development of age-invariant features, we focus on investigating whether a previously established face feature extractor based on DeepID2 can also be used for age-invariant recognition of infants. Specifically, we combine DeepID2 features with a classification network and train end-to-end on a dataset containing facial images of young children.

Other age-invariant face recognition frameworks focused on face modeling or synthesis for target ages. In [ABD17] the generation of faces at different ages was done via a generative adversarial network. While the quality of the generated faces is overall convincing, the faces cannot easily be used for feature extraction as they still differ from real faces. As a different modeling approach, Temporal Non-Volume Preserving transformations were introduced [DQL*17], which performed well on cross-age verification tasks. Liu et. al [LXZ*16] proposed to combine age-invariant face modeling and feature matching to solve face verification for large age gaps. Face images for 4 different age groups are modeled for each input image resulting in 4 synthesized image pairs and one original pair that are fed into parallel feature extraction CNNs. In contrast to this work, we only focus on the age category of children and aim to robustly identify children for small and large age gaps. Therefore, our input images would already lie in the same age category and would not benefit from the proposed face synthesis. Our proposed feature extraction part is similar to the parallel networks including the idea to substract the obtained features from each other. However, instead of only using a Softmax to decide whether the images match or not, we apply a classification network after feature extraction in order to achieve robust results within the same age category.

Overall, our paper focuses on the challenging infant face verification through a deep learning method. We directly use the im-

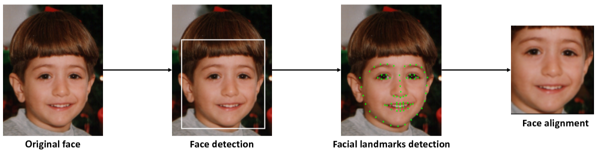

Figure 2: *An example of the face preprocessing pipeline performed on a child from the FGNET dataset [fgn]. First, face detection is performed, then 68 facial landmarks are detected, which are used to align the face by similarity transformations. Ultimately, the face is cropped based on the aligned face region. The final face image as used in the dataset is shown on the right.*

ages as input for our feature extraction CNN and therefore prevent the need of face synthesis or a hand-crafted feature descriptor. Instead we investigate whether CNN features can be directly combined with CNN classification networks to solve the age-invariant face verification task for young children.

3. Infant Face Dataset

In order to investigate whether a CNN can be trained to identify young children at different ages we create a new dataset. As face features vary, especially during early childhood, we only include infants of the age 0 to 6 in our dataset. Overall our infant dataset consists of 4528 face images with a size of 128 × 128 pixels featuring 42 children, including 24 (57%) girls and 18 (43%) boys as well as 2 pairs of sisters and a pair of brother and sister. For all children, images at different ages are included in the dataset.

We first select YouTube videos that show infants at different stages of their development. Our face preprocessing pipeline is inspired by DeepFace [TYRW14], however we use different techniques for each preprocessing step. To collect face information from the chosen videos we use the *Histogram of Oriented Gradients* algorithm [DT05] and employ it to detect face contours in every frame. While this already provides facial regions which could be cropped from the image and used to train CNNs, further data normalization can improve the training process. Therefore, we align the face images to increase the training performance as proposed by Sun et al. [TYRW14]. Since both eyes of the face should be on the same horizontal line it is necessary to extract facial landmarks within the previously detected facial region. This is done using the *Ensemble of Regression Trees* [KS14] algorithm, which detects 68 facial feature points. Afterwards the image is aligned using similarity transformations and cropped based on the area which contains the facial landmarks. Images that cannot be aligned are removed from the dataset. An overview of the preprocessing steps is presented in Fig. 2. For both the HOG and Ensemble of Regression Trees algorithm we use the implementation of the Dlib library [Kin09].

Once the cropped face images are computed, we manually clean the data to remove duplicates and images with insufficient resolution. Finally, each image in the database is labeled with an ID for the represented child and an index. However, the data in our new infant face database is still unbalanced. To balance the data, we perform upsampling to augment the number of face images for children with less than 100 face images to reach 100 face images.

We split our dataset into a training set containing around 3500 images and a test set containing around 1200 images. The training and test split is non-overlapping, therefore each child is exclusively either in the training or test set. We randomly chose 31 children for the training set and use the images of the other 11 children in the test set. For both sets, we randomly generate positive and negative face pairs. The number of positive and negative face pairs is the same in order to avoid training problems due to imbalanced data. Overall, the training set contains 128,000 pairs, while test set contains 32,000 pairs.

4. Network Architecture

We use CNNs for the infant face verification task by presenting two face images to the network and deciding whether these belong to the same child. Our CNN architecture consists of two parts as shown in Fig. 3: In the first part we perform feature extraction by applying DeepID2 [SCWT14] in a Siamese network style [CHL05], in the second part we perform similarity learning on the features as shown in Fig. 5. The input to the network is a pair of RGB images of size 128 × 128 pixels and the output is a probability score indicating whether the same child is presented on the images. Overall, the feature extraction network learns high-level discriminative infant face features, which are afterwards fed to a classification network that predicts the similarity of both learned features respectively.

4.1. Facial Features Extraction

Our network structure for infant facial feature learning is adapted from DeepID2 [SCWT14], which performs deep face features extraction and has shown outstanding performance for the face verification task. Instead of treating the output as a one dimensional vector of facial features which would be obtained by applying a fully-connected layer, we compute a three dimensional feature matrix by summing up the last convolutional layer and max pooling layer. We chose this option, since we want to keep the features' local information when we input them into the following classification part. An overview of the feature extraction network's architecture is given in Figure 4. As we want to use two images, we use two branches each containing the feature extraction network and apply weight sharing similar to the Siamese network [CHL05]. In the end the obtained features are subtracted from each other before passing them to the classification part. An additional experiment has shown that using the difference instead of a concatenation delivers slightly better results for our architecture. Subtracting the feature vectors instead of concatenating them may avoid a possible a local minimum during training.

4.2. Similarity Learning

After we obtain infant face features from the feature extraction network, we need to perform similarity learning to decide whether or

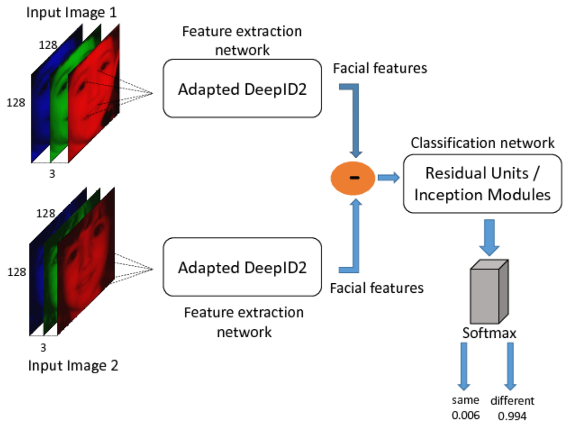

Figure 3: *Overview of the architecture for the whole network which receives two RGB images as input and predicts the probability that the same child is depicted in both images. The first part of the CNN is used to extract facial features from an input image by adapting DeepID2 and using it in two branches of the network. Afterwards, the obtained features from each branch are subtracted from each other before passing them to the classification part of the network. The final output of the network are two probability scores for match and non-match.*

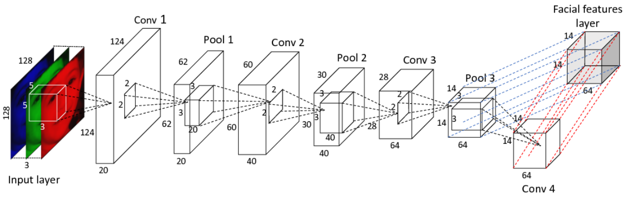

Figure 4: *The architecture for the first part of our network. This part of the CNN is used to extract facial features from an input image. We adapt DeepID2 [SCWT14] to output 3-dimensional features instead of 1-dimensional features. In the last layer, the outputs of Convolutional layer 4 (Conv4) and Max-pooling layer 3 (Pool3) are summed up to form a representative feature cuboid.*

not the same child is present in the input images. While other works suggested techniques like the Mahalanobis distance [HLT14], the Cosine Similarity [XLW*17] or a Joint Bayesian model [SCWT14, CCW*12] to solve the binary classification of matching or non-matching image pairs, we choose a CNN. CNNs for classification tasks have been applied very successfully in other areas, moreover, using a second CNN directly after the feature extraction network enables us to learn end-to-end. Before handing the three dimensional features to the classification network, we first subtract both vectors from each other to obtain the difference between the features of both face images.

For our investigation, we compare two networks using different classification architectures. First, we will discuss a classification based on *ResNet* [HZRS16], which enables efficient training for deep network structures. Afterwards, we employ *Inception Modules* as proposed in the *GoogLeNet* architecture [SLJ*15], which efficiently extracts pixel information using multi-scale convolutional kernels with large receptive fields. Both CNNs take the feature difference as input which has the spatial size of 14 × 14 × 64. An overview of the network structure is given in Fig. 5.

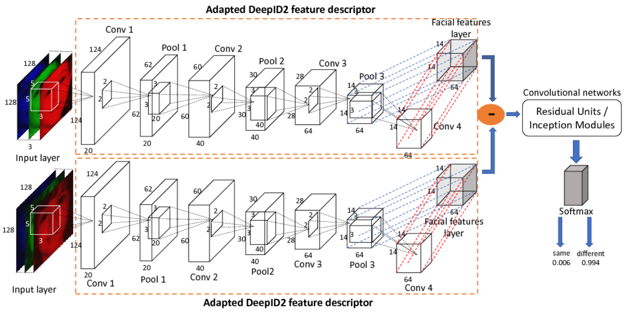

Figure 5: *The combination of the added feature extractor network with a classification network. We test two kinds of classification networks either containing several Residual Units or Inception Modules. The latter classifies three dimensional features and learns the similarity between both extracted face feature vectors.*

4.2.1. Network Architecture with Residual Units

We use *Residual Units* [HZRS16] which allows very deep network structures without performance degradation. The Residual Units enable the network to learn abstract and representative features and therefore produce a good classification performance. The proposed architecture of the residual classification network includes 18 convolutional layers and 7 Residual Units. We applied Parametric Rectified Linear Units (PReLU) [HZRS15] instead of Rectified Linear Units (ReLU) [NH10] as the activation functions. In contrast to ReLUs, PReLUs are able to retain negative values which might exist in the feature difference. The final classification is done using a Softmax layer. The full architecture details of the Residual Units network is described in Table 1.

4.2.2. Network Architecture with Inception Modules

In contrast to the idea of using several Residual Units to increase the depth of the network, *Inception Modules* [SLJ*15] enable larger receptive fields by using several multi-scale kernels in one module. This way information are based on a larger data region and more precise, however, the computation time and complexity is increased. Table 2 details the classification network architecture using Inception Modules. First, we want to feed the feature difference into an Inception Module with output sizes 14 × 14 × 512. For a smoother transition of the initial feature size of 14 × 14 × 64 to the target depth of 512, we employ two convolutional layers with a depth of 128 and 256 respectively, while keeping the height and width fixed. The convolutions in each Inception Module use filter sizes of 1 × 1, 3 × 3 and 5 × 5. Within the Inception Modules, convolutions with a kernel size of 1 × 1 extract information across all pixels and reduce the dimensionality of the output. Locating them before the 3×3 and 5×5 convolutional layers significantly decreases computational cost. Afterwards, the convolutional layers with filter size of 3×3 and 5×5 cover larger pixel regions and extract their features. At the end of the classification architecture, we implement two fully-connected layers after the average pooling layer to get more abstract features. Finally, a Softmax is used to compute the probability of a match between the input images.

4.3. Training

We train our network consisting of feature extraction and classification end-to-end. For the loss calculation of the feature extraction network, we first convert the three dimensional feature representation into a one dimensional representation by concatenating the features into a column vector. This way, we are able to utilize the *Contrastive Loss* [HCL06]:

$$L_c = \frac{1}{2N} \sum_{n=1}^{N} yd^2 + (1-y)\max(t-d,0)^2 \tag{1}$$

where N is the batch size which was set to 64 in our experiments, d indicates the euclidean distance of two samples' features and $y \in \{0,1\}$ denotes whether the pair matches. The t is a given threshold value, which is set to 1 in our experiments as proposed by Hadsell et al. [HCL06]. This loss efficiently reduces intra-personal variations while enlarging inter-personal variations. In combination with the larger feature vector this enables a more discriminative identity component.

Both variations of the classification network use the Softmax loss. The Softmax loss function is often used in CNNs for classification tasks to maximize the inter-class variations. For the Residual Units network the Softmax loss is computed at the end of the network and uses the same weight as the Contrastive Loss. The Inception Modules network uses an additional Softmax loss, since the Inception Modules network produces especially discriminate features in the middle layers. The additional Softmax loss is applied after the Inception (7d) layer for intermediate supervision. Overall the loss of our network can be denoted as

$$L = \alpha L_c + \beta L_{S1} + \gamma L_{S2} \tag{2}$$

where L_c is the Contrastive Loss from Eq. 1. Both L_{S1} and L_{S2} are Softmax losses, however, L_{S2} is used only in the Interception network after the Inception (7d) layer. For all our experiments we use $\alpha = 1$, $\beta = 1$ and $\gamma = 0.3$ as weights for the losses.

We use the *Caffe* [JSD*14] framework to implement our CNNs and use the same parameters for both versions. Before training, we perform data normalization by subtracting the mean color values as present in the training set. During training we create batches of images of size 64 and optimize using *Stochastic Gradient Decent* with an initial learning rate of 0.001, which was decreased using an inverse strategy. All network weights were initialized using the Xavier method [GB10]. To avoid the problem of vanishing gradients Batch Normalization [IS15] was applied. Furthermore, we prevent overfitting using Dropout [SHK*14]. Overall, the training process takes about 15 epochs for the Residual Units network and 10 epochs for the Inception Modules network. Training on a GeForce Titan X took about 15 minutes per epoch leading to a total training time of about 5 hours.

5. Experiments

In this section we present different experiments to investigate the effectiveness of our proposed network architectures on our infant face dataset and the infant subset of the FGNET dataset.

5.1. Experiments on our Infant Face Dataset

We first conduct experiments on the test set of our infant face dataset, to investigate the trainings result for both of the classification models. The achieved performance for both CNN models on our new infant face database is reported in Table 3. The proposed network combination of the adapted DeepID2 feature descriptor with *Inception Modules* achieves a verification accuracy of 85.3% and slightly outperforms the combination with *Residual Units* which achieves an accuracy of 84.01%. We additionally include experiments using the Siamese network of the Caffe framework with an Euclidean and Mahalanobis distance metric for feature matching, and a simplified architecture composed by the DeepID2-based feature extraction and a fully-connected layer with softmax, as baseline. The Mahalanobis metric achieves an accuracy of 77.9%, outperforming the Euclidean metric with an accuracy of 69.4%. The DeepID2-based architecture with a fully-connected layer achieves an accuracy of 78.8%. While the Siamese and fully-connected layer networks perform reasonable, our proposed networks outperforms them by a significant amount. Our results show that the classification of features which are not specifically designed to be age-invariant, can still offer promising results for age-invariant face verification of young children and infants.

5.2. Experiments on the Infant Subset of FGNET

We perform additional experiments on the FGNET [fgn] dataset to compare our method to a state-of-the-art CNN-based age-invariant face recognition proposed by Wen et al. [WLQ16]. They also present their results for very young children for the FGNET dataset, more specifically, for the age range from 0 to 4 years. As one of the public domain face aging datasets, the FGNET consists of 1002 face images from 82 different persons in age range of 0 to 69 years, but it only contains 193 face images for the age group of 0 to 4 years. Before we conduct the experiment, we perform the same preprocessing as for our own dataset as detailed in Section 3. Specifically, we use the HOG algorithm [DT05] to detect face regions and the algorithm Ensemble of Regression Trees [KS14] to detect 68-points facial landmarks used to align the faces. We balance the data by upsampling face images to achieve the same of amount of images on each subject in order to randomly select face pairs with equal possibility.

Wen et al. [WLQ16] proposes a deep learning architecture to learn age-invariant features and achieves outstanding performance on the face recognition task. The authors trained their network(LF-CNN) on large scale web-collected face datasets. In contrast to our infant child dataset, these datasets do not focus on images of the same person at different and young ages. We present the results of the investigated network architectures as well as the Rank-1 identification accuracy as stated by LF-CNN in Table 4. Between our proposed architectures, the network architecture of our adapted DeepID2 features combined with an classification part containing

Type	**Output Size**	**Kernel Size**	**Stride**	**Padding**
Convolution (5)	14 × 14 × 128	3	1	1
Convolution (6)	12 × 12 × 256	3	1	
3 Residual Units (7)	12 × 12 × 256	3	1	1
Convolution (7)	10 × 10 × 512	3	1	
Max Pooling (7)	5 × 5 × 512	2	2	
4 Residual Units (8)	5 × 5 × 512	3	1	1
Convolution (8)	4 × 4 × 1024	2	1	
Max Pooling (8)	2 × 2 × 1024	2	2	
Fully-Connected (9)	1 × 1 × 1024			
Fully-Connected (10)	1 × 1 × 4096			
Fully-Connected (11)	1 × 1 × 2			

Table 1: *The architecture details of the classification network using Residual Units. Further parameters considering the detailed layout of the Residual Units are chosen as suggested by the ResNet [HZRS16] architecture.*

Type	**Output Size**	**Kernel Size**	**Stride**
Convolution (5)	14 × 14 × 128	3 × 3	1
Convolution (6)	14 × 14 × 256	3 × 3	1
Inception (7a)	14 × 14 × 512		
Inception (7b)	14 × 14 × 512		
Inception (7c)	14 × 14 × 512		
Inception (7d)	14 × 14 × 528		
Inception (7e)	14 × 14 × 832		
Max Pool	7 × 7 × 832	3 × 3	2
Inception (8a)	7 × 7 × 832		
Inception (8b)	7 × 7 × 1024		
Average Pooling	1 × 1 × 1024	7 × 7 × 1	
Dropout 40%	1 × 1 × 1024		
Fully-Connected	1 × 1 × 2048		
Fully-Connected	1 × 1 × 2		
Softmax	1 × 1 × 2		

Table 2: *The architecture details of the classification network using Inception Modules. The number of kernels for each Inception Module are chosen as suggested by the GoogLeNet [SLJ*15] architecture.*

Network Architecture	Acc.
Siamese network+Euclidean metric	0.694
Siamese network+Mahalanobis metric	0.779
DeepID2 + Fully-Connected + Softmax	0.788
Ours (Residual Units)	0.841
Ours (Inception Modules)	**0.853**

Table 3: *Overview of the accuracy of the proposed CNNs. We also include results for a Siamese network using Caffe's standard implementation, and a simplified architecture based on DeepID2 with a fully-connected layer and softmax, as baseline.*

Network Architecture	True-positive Rate	Acc.
Siamese+Euclidean metric	0.532	0.583
LF-CNNs [WLQ16] (Identification)	0.601	-
Ours (Residual Units)	0.665	0.694
Ours (Inception Modules)	**0.709**	**0.726**

Table 4: *The results of the infant face verification using our network architectures on the infant subset of the FGNET dataset. Here, we include the true-positive rate and compare it with the Rank-1 identification rate for LF-CNNs [WLQ16]. The accuracy values (Acc.) indicate the verification accuracy.*

Inception Modules again achieves the best verification accuracy with 72.6% while the combination with Residual Units achieves 69.4% accuracy. In LF-CNN no accuracy was stated for the investigated age group. We used the true-positive rate to compare our networks with the Rank-1 identification result stated in LF-CNN. While the true-positive rate indicates the percentage of matches that were correctly identified as images of the same child, the Rank-1 identification rate denotes that the highest rated retrieved match is also the correct match. Using our networks, the best true-positive rate is also achieved by the Inception Modules network at 70.9% followed by the Residual Units network with 66.5%. Both networks outperform the Rank-1 identification rate of LF-CNN which is 60.1% on the infant age group of 0 to 4 years in the FGNET dataset, further supporting the idea of training standard facial feature extractors on infant child datasets. As the code of the LF-CNN network was not publicly available, we were not able to train their network on our infant child dataset and better distinguish the influence of using our dataset and our network architecture. Fig. 6 shows examples of image pairs that were not classified correctly by our network, illustrating false positives as well as false negatives. Most failure examples might be difficult to classify even for human observers due to the drastic changes of the infants facial shape.

We perform an additional Experiment on the FGNET dataset to further investigate the effects of our network using *Inception Modules* on age-invariant child verification. For this experiment we extract all images from out test set for a target age and compute the true-positive rates when testing with other age groups. For example, we chose each child at the age of 1 and pair it with all children at age 2 which yields a true-positive rate of 93%. In contrast, choosing a child at the age of 1 and forming pairs with all children at age 6 showed a true-positive rate of 84%. Table 5 shows the true-positive

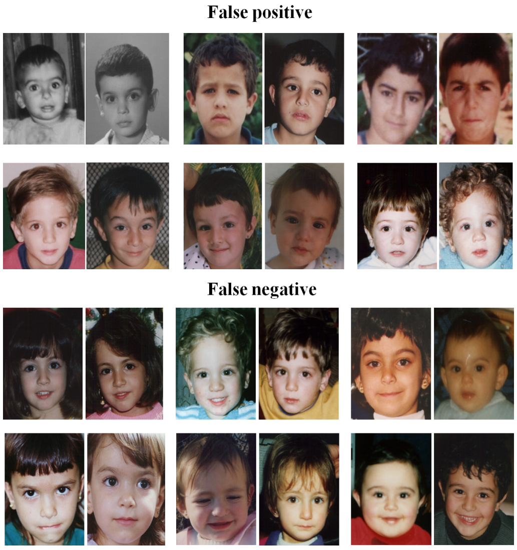

Figure 6: *A demonstration for failure cases on the infant subset of the FGNET dataset. The subjects in this sub-dataset are in age range of 0 to 4 years and show drastic changes in their facial features.*

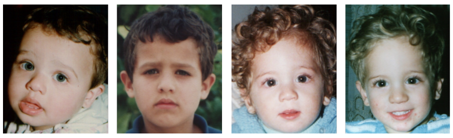

Figure 7: *An example for matched pairs that were correctly identified by our network. It is clearly visible how the facial features have changed and that verifying the child's identity is more difficult for the larger age gap. Left: a child at 1 year and 6 years. Right: a child at 1 year and 2 years.*

Chosen Age	Tested Age	True-positive Rate
1 year old	2 years old	0.93
1 year old	3 years old	0.86
1 year old	4 years old	0.84
1 year old	5 years old	0.93
1 year old	6 years old	0.84

Table 5: *The table shows the true-positive rates when forming pairs of children at specific ages to investigate the robustness of age-invariant verification. While the true-positive rate is generally higher for smaller age gaps, our approach still performs well when we use children at the age of 1 and children at the age of 6 yielding a true-positive rate of 84%.*

rate between images of 1 year old children and 2 to 6 years old ones in our test set. An example of two correctly verified pairs is given Fig. 7. As expected, with increasing age difference the true-positive rate becomes worse. However, the performance degradation is still very reasonable considering the amount of changes to facial features in the first years.

6. Conclusions

In this paper, we investigated how the task of age-invariant infant face verification can be solved using adapted DeepID2 features [SCWT14] combined with popular classification network architectures. In contrast to previous work focusing on the creation of age-invariant feature extraction or generation, we showed that classic deep feature descriptors can be adapted for age-invariant infant verification when trained on an appropriate dataset. We focused on children faces whose features change drastically in the early years as these are especially challenging for the face verification task. Since no existing dataset provided enough infant face images to train a CNN, we first prepared a dataset consisting of 4,528 face images of 42 children in age range of 0 to 6 years. Therefore, our dataset entails a vast number of examples for facial features changing due to aging effects in the early childhood.

We tested classification networks based on *Residual Units* [HZRS16] and *Inception Modules* [SLJ*15] as both architectures have shown outstanding performance at classification tasks. For both tests, we first use adapted DeepID2 features in two branches as proposed by Siamese networks to extract two feature matrices from two presented input images. Afterwards the difference of the features is computed and passed to the classification part of the network. The classification network produces a probability as output which indicates whether the two input images show the same child or not. We train both of our combined networks end-to-end on the train subset of our dataset, achieving an accuracy of 85. Both network combinations showed promising results on our test dataset and on the infant subset of the FGNET [fgn] dataset. For our test dataset, we achieve the best accuracy (85,3%) with the *Inception Modules* network, slightly outperforming the *Residual Units* network (84,1% accuracy). Overall, the proposed network outperforms previous work when applied to the children subset of the FGNET database containing children from 0 to 4 years. The combinations of our adapted DeepID2 features with a classifier based on *Inception Modules* achieves an accuracy of 72.6%on this subset. In the future, we intend to investigate face verification of older children and will evaluate the possibilities of making our infant dataset available for future research.

7. Acknowledgments

The authors gratefully acknowledge funding by the L3S Research Center, Hanover, Germany. We also want to thank the anonymous reviewers for their many valuable suggestions.

References

[ABD17] Antipov G., Baccouche M., Dugelay J.-L.: Face aging with conditional generative adversarial networks. In *Proc. IEEE International Conference on Image Processing (ICIP)* (2017), pp. 2089–2093. 2

[BBVS16] Bharadwaj S., Bhatt H. S., Vatsa M., Singh R.: Domain specific learning for newborn face recognition. *IEEE Transactions on Information Forensics and Security 11*, 7 (2016), 1630–1641. 2

[CCW*12] Chen D., Cao X., Wang L., Wen F., Sun J.: Bayesian face revisited: A joint formulation. In *Proc. European Conference on Computer Vision* (2012), Springer, pp. 566–579. 4

[CHL05] CHOPRA S., HADSELL R., LECUN Y.: Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2005), vol. 1, pp. 539–546. 2, 3

[DQL*17] DUONG C. N., QUACH K. G., LUU K., LE T. H. N., SAVVIDES M.: Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 3755–3763. 2

[DT05] DALAL N., TRIGGS B.: Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2005), vol. 1, pp. 886–893. 3, 5

[FGH10] FU Y., GUO G., HUANG T. S.: Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence 32*, 11 (2010), 1955–1976. 2

[fgn] Fgnet aging database. http://www.fgnet.rsunit.com/. 1, 2, 3, 5, 7

[GB10] GLOROT X., BENGIO Y.: Understanding the difficulty of training deep feedforward neural networks. In *Proc. International Conference on Artificial Intelligence and Statistics* (2010), pp. 249–256. 5

[GMP*06] GOLOVINSKIY A., MATUSIK W., PFISTER H., RUSINKIEWICZ S., FUNKHOUSER T.: A statistical model for synthesis of detailed facial geometry. *ACM Transactions on Graphics (TOG) 25*, 3 (2006), 1025–1034. 2

[HCL06] HADSELL R., CHOPRA S., LECUN Y.: Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2006), vol. 2, pp. 1735–1742. 5

[HLT14] HU J., LU J., TAN Y.-P.: Discriminative deep metric learning for face verification in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)* (2014), pp. 1875–1882. 2, 4

[HRBLm] HUANG G. B., RAMESH M., BERG T., LEARNED-MILLER E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 1

[HZRS15] HE K., ZHANG X., REN S., SUN J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE International conference on computer vision* (2015), pp. 1026–1034. 4

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778. 2, 4, 6, 7

[IS15] IOFFE S., SZEGEDY C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning* (2015), pp. 448–456. 5

[JSD*14] JIA Y., SHELHAMER E., DONAHUE J., KARAYEV S., LONG J., GIRSHICK R., GUADARRAMA S., DARRELL T.: Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM International conference on Multimedia* (2014), ACM, pp. 675–678. 5

[Kin09] KING D. E.: Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research 10*, Jul (2009), 1755–1758. 3

[KS14] KAZEMI V., SULLIVAN J.: One millisecond face alignment with an ensemble of regression trees. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1867–1874. 3, 5

[KSMB15] KEMELMACHER-SHLIZERMAN I., SEITZ S. M., MILLER D., BROSSARD E.: The megaface benchmark: 1 million faces for recognition at scale. *CoRR abs/1512.00596* (2015). 1

[LXZ*16] LIU L., XIONG C., ZHANG H., NIU Z., WANG M., YAN S.: Deep Aging Face Verification with Large Gaps. In *IEEE Transactions on Multimedia* (2016), pp. 64–75. 2

[MPS*17] MOSCHOGLOU S., PAPAIOANNOU A., SAGONAS C., DENG J., KOTSIA I., ZAFEIRIOU S.: Agedb: the first manually collected, in-the-wild age database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). 1

[NH10] NAIR V., HINTON G. E.: Rectified linear units improve restricted boltzmann machines. In *Proc. International conference on machine learning (ICML)* (2010), pp. 807–814. 4

[PTJ10] PARK U., TONG Y., JAIN A. K.: Age-invariant face recognition. In *IEEE transactions on pattern analysis and machine intelligence* (May 2010), pp. 947–954. 1

[RC06] RAMANATHAN N., CHELLAPPA R.: Modeling age progression in young faces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2006), vol. 1, pp. 387–394. 2

[SCWT14] SUN Y., CHEN Y., WANG X., TANG X.: Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems* (2014), pp. 1988–1996. 1, 2, 3, 4, 7

[SHK*14] SRIVASTAVA N., HINTON G. E., KRIZHEVSKY A., SUTSKEVER I., SALAKHUTDINOV R.: Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research 15*, 1 (2014), 1929–1958. 5

[SKP15] SCHROFF F., KALENICHENKO D., PHILBIN J.: Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 815–823. 2

[SLJ*15] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1–9. 2, 4, 6, 7

[SMZ*07] SUO J., MIN F., ZHU S., SHAN S., CHEN X.: A multi-resolution dynamic model for face aging simulation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2007), pp. 1–8. 2

[SSSB07] SCHERBAUM K., SUNKEL M., SEIDEL H.-P., BLANZ V.: Prediction of individual non-linear aging trajectories of faces. In *Computer Graphics Forum* (2007), vol. 26, Wiley Online Library, pp. 285–294. 2

[SWT14] SUN Y., WANG X., TANG X.: Deep learning face representation from predicting 10,000 classes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1891–1898. 2

[TSS12] TIWARI S., SINGH A., SINGH S. K.: Intelligent method for face recognition of infant. *International Journal of Computer Applications 52*, 4 (2012). 2

[TYRW14] TAIGMAN Y., YANG M., RANZATO M., WOLF L.: Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1701–1708. 2, 3

[WLQ16] WEN Y., LI Z., QIAO Y.: Latent factor guided convolutional neural networks for age-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4893–4901. 1, 2, 5, 6

[WZK*17] WANG X., ZHOU Y., KONG D., CURREY J., LI D., ZHOU J.: Unleash the black magic in age: a multi-task deep neural network approach for cross-age face verification. In *Proc. IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (2017), pp. 596–603. 2

[WZLQ16] WEN Y., ZHANG K., LI Z., QIAO Y.: A discriminative feature learning approach for deep face recognition. In *Proc. European Conference on Computer Vision* (2016), pp. 499–515. 2

[XLW*17] XIAO T., LI S., WANG B., LIN L., WANG X.: Joint detection and identification feature learning for person search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 3376–3385. 4

[ZDH17] ZHENG T., DENG W., HU J.: Age estimation guided convolutional neural network for age-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 12–16. 2