Journal of computer science and technology: Instruction for authors. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

Is it easy to recognize baby's age and gender?

Abstract Face analysis tasks, e.g., estimating gender or age from a face image, have been attracting increasing interest in recent years. However, most of existing works focus mainly on analyzing an adult' face and ignore an interesting question: is it easy to estimate gender and age from a baby's face? In this paper, we explore this interesting problem. We first collect a new face image dataset for our research, named BabyFace, which contains 15,528 images from 5,872 babies younger than two years old. Besides gender, each face image is annotated with age in months from 0 to 24. In addition, we propose new age estimation and gender recognition methods. In particular, we introduce the attention mechanism module to solve the age estimation problem on the BabyFace dataset. Inspired by the age estimation method, the gender estimation also uses a two stream structure. We extensively evaluate the performance of the proposed methods agains state-of-the-art methods on BabyFace. Our age estimation model achieves very appealing performance with an estimation error of less than 2 months. The proposed gender estimation method obtains the best accuracy among all compared methods. To the best of our knowledge, we are the first to study age and gender estimation from a baby's face image, which is complementary to existing adult gender and age estimation and can shed some light on exploring baby face analysis.

Keywords Age estimation, gender recognition, baby face, CNNs, attention mechanism

1 Introduction

Face attributes such as age and gender play fundamental roles in social interaction. Analyzing face attributes have been attracting increasing interest in the past due to a wide range of applications, including cross-age face verification, finding missing children, and human-computer interaction. In the literature, a large number of age and gender estimation methods have been proposed [1, 2, 3, 4]. In particular, deep convolutional neural networks (CNNs) based methods continuously refresh previous records on the benchmark datasets (e.g., Adience faces [5] and the Labeled Faces in the Wild (LFW) [6]).

However, most of the existing works and datasets [7, 8, 9, 10, 4] focus on analyzing adult's age and gender, while ignore an interesting question: is it easy to analyze gender and age from baby facial images and how can we do it? There are two main reasons for the lack of research on baby face analysis. The first reason is that the community does not realize the potential applications. In fact, there are many applications of analyzing the facial attributes of babies, such as advertising marketing for parents, intelligent family child care [11] and scientific parenting [12]. As we all know, 0-2 years old is a golden period for the development of a baby and for laying a solid foundation for their lifelong physical and mental health. In addition, due to the support of national policies and people's growing attention to the growth and development of a baby, the parenting market has been expanding and parenting initiatives within electronic media are expanding [13]. Accurate recognition of a baby's age and gender is of great significance to improve scientific parenting. The second reason may be attributed to that existing datasets contain only a small amount or no baby face images at all [8, 9]. This may be traced back to the acddditional challenge of obtaining datasets for babies with accurate gender and age labels. All these real needs have brought motivation to the study of recognizing baby's face attributes.

To explore the aforementioned interesting question, in this paper, we first collect a baby face dataset with gender and age labels for studying age and gender estimation. The collected dataset, named Babyface, contains more than 15,500 face images from 5,872 babies

Regular Paper

less than two years old. Besides gender, each face image is annotated with age in months from 0 to 24. In addition, we propose new methods to estimate the age and the gender of babies using the newly proposed dataset. For age estimation, we used SSR-Net [14] as the backbone network. We introduce the attention mechanism module to solve the age estimation problem on the BabyFace dataset. Inspired by the age estimation method, the gender estimation also uses a two stream structure. Besides conventional regularization, we also implement image augmentation such as random erasing [15] and mixup [16] to prevent a model from overfitting. Our new methods show some promising results. Our age estimation model achieves very appealing performance with an estimation error of less than 2 months. And our gender estimation method obtains the best accuracy rate, i.e., 78.3%. To better analyze the differences on age and gender estimation between adults and babies and whether the analysis methods for adults are suitable for baby, we also conduct a comprehensive comparision of state-of-the-art methods on age and gender estimation on the BabyFace dataset.

Our experimental results show that estimating the age of a baby is not as difficult as everyone thinks, but it is extremely hard to estimate the gender of a baby compared with an adult. Our research is complementary to existing adult face age and gender estimation and can shed some light on exploring baby face analysis. Our work will attract the scientific community to conduct research with baby' faces in a manner that is comparable to a large extent literature that has heavily relied on adult faces.

Our key contributions are summarized in the following:

- We collect a baby face dataset, which contains 15,528 images from 5,872 babies with age ranging from 0 to 24 months. To the best of our knowledge, this dataset is the first and the largest baby face dataset.
- We propose new methods using two stream structure to predict the age and the gender of a baby, which show some promising results.
- We also conduct a comprehensive comparision of state-of-the-art methods on age and gender estimation on the BabyFace dataset. Our experiments show that estimating a baby' age is not as difficult as adult's imagination. The average error of evaluating a baby'age is less than two months. However, the gender estimation a baby is very difficult because gender features such as beard, eyebrows, and pores are not visible in a baby' face.

2 Related Work

2.1 Existing Age and Gender Datasets

Large-scale datasets have contributed greatly to the development of machine learning and computer vision [17]. For gender and age estimation, publicly available benchmarks also greatly facilitate the research. An overview of the databases of face, age and gender images that have been made publicly available is provided in Table 1. We conduct a comprehensive survey of existing datasets for the age and gender estimation tasks in the following.

FG-NET Aging [7]: The dataset is widely used for age estimation. It contains 1,002 facial images from 82 people and most people are white people. Everyone has more than 10 photos taken at different ages. In FG-NET, images vary greatly in lighting conditions, poses, and expressions.

MORPH [8]: This dataset is the most popular for age estimation. MORPH contains 55,134 photos from

Table 1. Existing datasets of face images with gender and age labels.

Datasets	FG-NET Aging	MORPH	CACD	Adience	IMDB-WIKI	VGGFace2
Images	1002	55134	163446	26580	524230	3.31M
Size	-	400×480	-	-	256×256	224×224
Subject	82	13618	2000	2284	20284	9131
Age	0-69	16-77	16-62	8 groups	0-100	-34-
Gender-label	No	No	No	Yes	Yes	Yes
Gender	both	both	both	both	both	both
Year	2004	2006	2014	2014	2015	2018

13,618 people with age ranging from 16 to 77. Each face photo is accompanied with an age tag. The race in MORPH is very uneven, with more than 96% of facial images coming from Africans or Europeans.

CACD [18]: This is a large-scale data set containing about 160,000 facial images of 2,000 celebrities. The dataset has been divided into three subsets: the training set, the test set, and the validation set. And the celebrities in the three subsets are 1,800, 120, and 80, respectively.

VGGFace2 [10]: The dataset contains 3.31 million images of 9,131 subjects, with an average of 362.6 images per subject. Images are downloaded from Google Image Search and vary widely in pose, age, lighting, race, and occupation (such as actors, athletes, and politicians).

IMDB-WIKI [9]: This is the largest publicly available dataset for estimating the age of people in the natural environment, including more than 500,000 images with precise age labels ranging from 0 to 100 years old. For the IMDB-WIKI dataset, these images are captured from IMDB and Wikipedia, with IMDB containing 460,723 images of 20,284 celebrities and Wikipedia containing 62,328 images. Since the images of the IMDB-WIKI dataset are obtained directly from the website, the IMDB-WIKI dataset contains many low-quality images such as blurred images, non-human images, multi-person images, blank images, and so on. In addition, many age labels are not precise.

Overall, the above existing datasets contain either only a small amount or no baby data at all. In this study, in order to verify how difficult it is for a baby's gender and age estimation, we try to address this critical issue by collecting a novel face image dataset with age ranging from 0 to 24 months.

2.2 Age and Gender Estimation

2.2.1 Age Estimation

Recently, the age and gender estimation has received widespread attention, which provides a direct and fastest way to obtain implicit and critical social information. CNN-based methods have been widely used for age estimation due to their superior performance over existing methods. Yi et al. [19] introduce a multi-task learning method with a relatively shallow CNN based age and gender estimation method. Rothe et al. [20] propose DEX: Deep EXpectation of Apparent Age From a Single Image method for Age classification using the ensemble of 20 networks on the cropped faces of IMDB-Wiki benchmark. Wang et al. [21] train a deeper CNN for extracting features from different layers for age estimation on FG-NET and MORPH. Levi et al. validate CNN's performance on unconstrained facial images [1]. In [22], Niu et al. propose to formulate age estimation as an ordinal regression problem with the use of multiple output CNN. A recent and detailed

survey of the Age, Gender classification can be found in the work of [23] All of these methods mentioned above have been verified effectively on adult datasets for age classification which may not suitable for baby face images in practical applications.

2.2.2 *Gender Estimation*

Although more and more researchers have found that gender classification plays an important role in our daily lives, few approaches based on machine learning have been proposed. Next, we briefly review and summarize related methods. Verma and Vig [24] show that CNN is more robust and performs much better than previous classifiers like SVM and Random Forest etc. Inspired by the dropout technique in training phase, Eidinger et al. [5] train a SVM with random dropout of some features and achieved promising results on their relatively small Adience dataset. In [1], Levi and Hassner show that by learning representations using deep convolutional neural networks, a significant increase in performance can be obtained on the gender estimation task. Instead of training on entire images, Mansanet et al. [25] train with local patches and reported better accuracies than holistic image based networks of similar depth. Although these methods have shown many advantages, it is easy to find that the experimental dataset is constrained, so these methods are not suitable for practical applications, including unconstrained image tasks and datasets containing all baby pictures. As in [1], gender estimation mistakes also frequently occur for images of babies or toddlers where obvious gender attributes are not yet visible. The performance of gender classification is low among the baby subgroup in [2].

3 BabyFace Collection

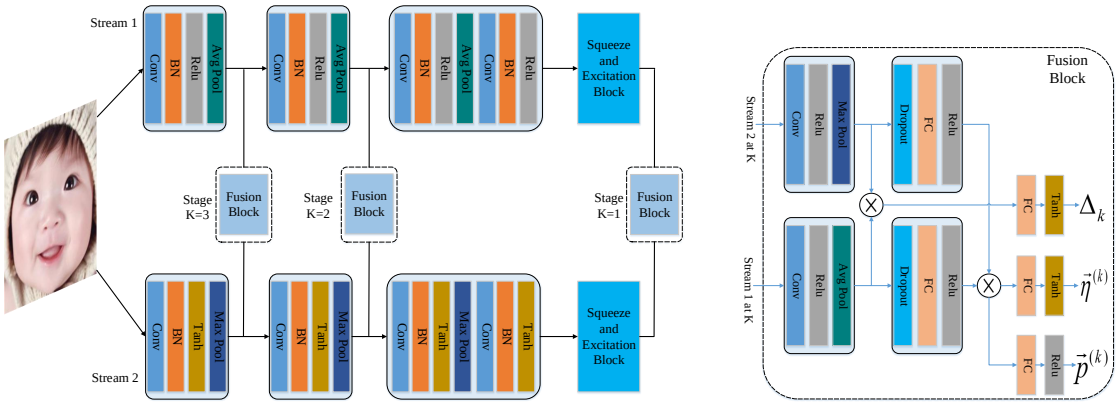
3.1 Data Pre-processing

Our baby face images are generated from both static images and video sequences obtained by smartphones and other mobile devices uploaded by parents. For the original images and the original videos data, we first perform face detection, then perform face cropping, and finally perform picture similarity detection. A detailed description of the pre-processing of the BabyFace dataset can be found below.

3.1.1 *Image Pre-processing*

For image processing, we first use the Dlib [26] visual library and the opencv visual library to perform face detection and cropping on the original image. During the face detection, we adopt the following strategy. First of all, if a face appears, the face of the picture will be cropped and saved. Second, if no face is detected during the detection, rotate the picture 270 degrees clockwise and 90 degrees each time for a total of 3 rotations. If a face appears during the three rotation detection process, then crop and save the face image in the picture. Last if there are two or more faces detected in the picture, we will assume that this image will have an adult face or a face that is not a human face but is misidentified as a human face. Then we will discard such photos.

It is important to note that the area of the original picture that the baby's face is not very large. At this point, the picture is too redundant. If it is used directly for training, the model converges slowly, resulting in poor test results, etc. In order to reduce the large amount of non-face information in the image, therefore, after using the above Dlib face detection strategy, when cropping the face, we crop the face area according to a specific artificial strategy and save it as a face dataset. The main purpose is to obtain a noise-



(a) Structure of SSR-SE (b) Fusion Block in SSR-SE

Fig.1. 1(a) The network structure of the SSR-SE with three stages (K = 3). The size of pooling is fixed at 2×2 for all stages. 1(b) The detailed structure of the fusion block in SSR-SE.

free and good-quality baby face image dataset in order to obtain a better model during training process and a better accuracy during test process. We then crop the original image according to the new picture size, and finally normalize the cropped image(the normalized size is: 256 ×256).

3.1.2 Video Pre-processing

We segment the original video data set, take an image every 30 frames, and then perform the same process as the static image data pre-processing on the images from the video frames, detecting, rotating, and finally cropping the baby's face picture. It should be noted that because the pictures obtained by intercepting video frames may have great similarities, many pictures are redundant, so the only different operation different from the static image is that after the picture is cropped and saved, we need to perform picture similarity matching operations to filter image. We use SSIM [27] to perform similarity matching, and specify to delete images with similarity greater than 90%, and finally obtain 10425 images from the original video data.

3.2 Data Annotation

After pre-processing, the baby face images have to be labeled in a file corresponding to the original image information. The original image information is: serial number, baby age, baby birthday, picture description, baby gender, photo name (or video name), we need to extract the baby age, baby gender, photo or video name in excel.

First, we need to convert all the image suffixes into a unified jpg format. The second step is to read the contents of the excel form to get useful information. Since our image comes from two parts, the last step is divided into two parts: the first part is to use the excel of the picture to get the picture name and determine whether the picture exists. If it exists, the label will be written. The format is baby age, baby gender, the relative path of the picture. The video data is a picture taken every 30 frames, a video may have multiple corresponding pictures, so the naming rule of the divided frame picture is the video name plus the number of frames, which is mainly used for the label. As we all know, a video corresponds to the tags of multiple pictures. According to the excel file of the video, we save the tag information corresponding to the video image

Table 2. Sample number of the BabyFace dataset for all ages (0-24 months).

Age(month)	0	1	2	3	4	5	6	7	8	9	10	11	12
Number	1081	2047	2615	1819	1729	1416	716	1074	649	455	306	426	198
Age(month)	13	14	15	16	17	18	19	20	21	22	23	24	
Number	147	50	74	98	84	130	95	84	87	41	32	75	

as the baby's age, baby's gender, and the relative path of the picture.

From the real label of the user-uploaded image, we select the information of babies from 0 to 2 years old, of which the boys are about 46 percent. The age distribution of the BabyFace dataset is shown in Table 2 and gender information can be seen from Table 3. Compared the age and gender attributes with the existing datasets as listed in Table 1, the BabyFace dataset has been built to bridge the gap between the need of accurate estimation of gender and age of baby and the lack of appropriate datasets for researching.

Table 3. Sample number of gender information for the BabyFace dataset.

Images		Subject	
Girl	Boy	Girl	Boy
8139	7389	3175	2697

4 Methodology

In this section, we describe our contributions in designing new network architectures to estimate the age and the gender of BabyFace.

4.1 Network Architecture for Age Estimation

We use SSR-Net [14] as the backbone network, because the model is compact with only 0.32 MB memory overhead and achieves the state-of-the-art performance on adult dataset. We add the attention mechanism module to the SSR-Net. For the attention mechanism module, we use Squeeze-and-Excitation (SE) block [28]. The overall age estimation pipeline SSR-Net + SE (SSR-SE) is depicted in Figure 1. SE block is used for refining CNN features and highlighting salient aging region, which can adaptively recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels. The detailed structure can be seen in Figure 2, γ is a scaling parameter, which is 16 in this paper. The purpose of this parameter is to reduce the number of channels and thus reduce the computation. C represents the number of channels, and H, W represents the height and width of the feature map input from the previous layer, respectively.

The SE module first performs a Squeeze operation on the feature maps obtained by the convolution to obtain channel-level global features, here we use global average pooling as the Squeeze operation. Then performs an Excitation operation on the global features, two fully connected layers form a bottleneck structure to model the correlation between channels, and output the same number of weights as the input features. As shown in Figure 2, we first reduce the feature dimension to 1/16 of the input, and then activate it through ReLu and then up-sample to the original dimension through a fully connected layer, which learns the relationship between each channel, and also obtains the weight of different channels, and finally multiplies the original feature map to get the final feature. In essence, the SE module performs attention or gating operations on the channel dimension. This attention mechanism allows the model to pay more attention to the channel features with the most information, and suppress those unimportant channel features.

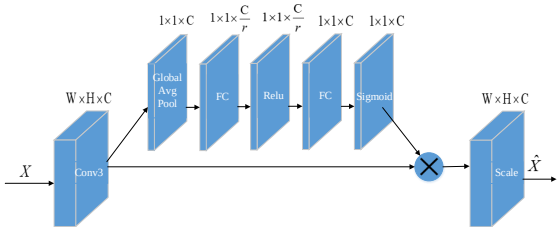


Fig.2. Squeeze-and-Excitation(SE) block, γ is a scaling parameter.

In Figure 1, for both streams, the basic building block is composed of 3 × 3 convolution, batch normalization, non-linear activation and 2×2 pooling. However, different types of activation functions (ReLU versus Tanh) and pooling (average versus maximum) are adopted for each stream to make them heterogeneous. Specific experimental details are consistent with the original paper [14].

4.2 Network Architecture for Gender Estimation

Figure 3 illustrates the overall network structure of the proposed gender estimation approach, the two stream + Squeeze-and-Excitation (SE) block + augmentation (TSSEAug), inspired by the age estimation method, the gender estimation also uses a two stream structure, but because gender is a binary classification problem, the network structure is relatively simpler than the age estimation. For both streams, the basic building block is composed of 3 × 3 convolution, nonlinear activation and 2×2 pooling. However, each stream uses a different type of activation function (ReLU and Tanh) and pooling (average and maximum) to make it heterogeneous. As a result, they can explore different capabilities, and their fusion can improve performance. Then followed by a Squeeze-and-Excitation (SE) block for refining CNN features and highlighting salient expressional region, which adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. Features from both streams after SE block are fused and followed by batch normalization, an average pooling layer, then feed into fully connected layers, followed by a softmax layer to obtain the final possibilities of different gender.

Initially, we implement conventional regularization and image augmentation such as random erasing [15] and mixup [16] to prevent the model from overfitting. During the training phase of the model, any image preprocessed with random erasing will either be kept unchanged or randomly have a rectangle region of an arbitrary size assigned with arbitrary pixel values. By simply introducing a reasonable noise to the training sample, this method enhances the model to become less prone to overfitting.

After the training samples is pre-processed with random erasing, an extra measurement is taken to further guarantee the generalization of the same model. In essence, rather than handling each images independently, mixup groups pairs of samples and their labels together before training the CNN. The method is proven to enhance robustness, stability and accuracy of models on popular datasets like ImageNet, CIFAR-10 and CIFAF-100 dataset, resulting in state-of-the-art architecture. Despite its powerful capability, mixup can be easily implemented. To put it briefly, mixup creates virtual training samples through the formula:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \tag{2}$$

where x_i and x_j are raw input vectors; y_i and y_j are one-hot encoding labels; (x_i, y_i) and (x_j, y_j) are two randomly examples from BabyFace training batch, and $\lambda \in [0, 1]$.

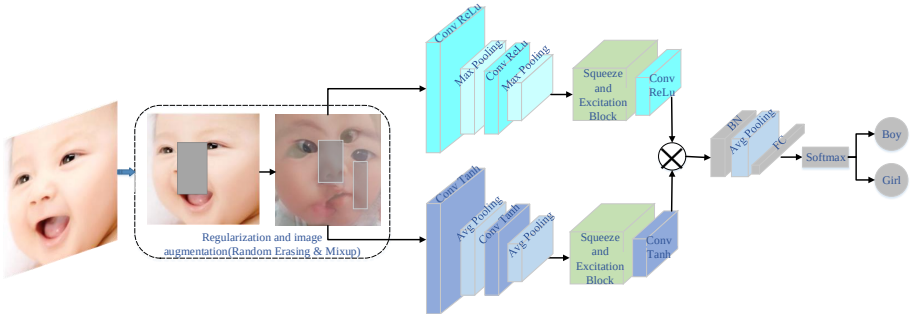


Fig.3. Overall network structure of the proposed gender estimation approach.

5 Experiments on BabyFace

In this section, we describe the experimental settings and demonstrate the effectiveness of the proposed methods by comparing them with state-of-the-art methods.

5.1 Experimental Setup

We set up two major experiments: one experiment is evaluating age estimation analysis methods on BabyFace. Another experiment is gender estimation. All training and testing processes are performed on NVIDIA GeForce GTX 1080Ti 16G GPUs. We developed our models using the open-source deep learning framework Keras. On an Ubuntu Linux system equipped with NVIDIA GPUs, training a model takes 3-4 hours depending on the architecture of age or gender estimation.

5.2 Implementation and Training Details

5.2.1 Age Estimation

For age estimation, we also do experiments on MobileNet, DenseNet for comparison our proposed SSR-SE method. In all experiments, 80% of images are used as the training set while the remaining 20% acted as the validation set. For training, common data augmentation tricks including zooming, shifting, shearing, and flipping are randomly activated. The Adam method [29] is used for optimizing the network parameters with 90 epochs. The learning rate is 0.002 initially and reduced by a factor 0.1 every 30 epochs. The batch size is 128. We report the mean absolute error (MAE) on the BabyFace dataset. MAE represents the average of the absolute errors between the predicted age and the ground truth over all test samples.

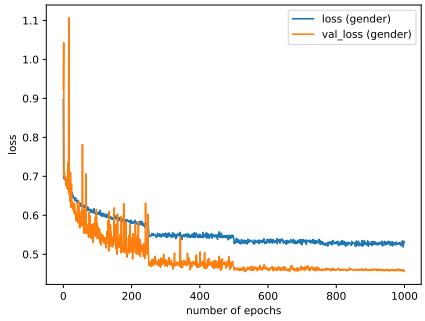
$$MAE = \frac{\sum_{i=1}^{n} |x_i - \hat{x}_i|}{n} \tag{3}$$

where n is the number of images, x_i and $\hat{x}_i$ are the ground truth and predicted age of the baby in image i, respectively. The experimental results are shown in Figure 5.

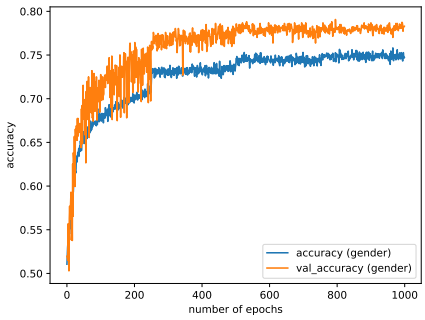
5.2.2 Gender Estimation

In the gender estimation experiments, for comparison, we run some ablations to analyze our TSSEAug approach, named Two Stream, Two Stream+SE (TSSE), and we also verify one stream experiments. For each experiment, 80% of images are used as the training set while the remaining 20% acted as the validation set. For training, common data augmentation tricks including zooming, shifting, shearing, and flipping are randomly activated. The SGD method is used for optimizing the network parameters with 1500 epochs. The batch size is 32. we also compare HyperFace [30], R_CNN_Gender [30], CNN_AgeGender [1] and SlightHound [31], which are often used in recent

years for adult gender estimation. Experiment settings are all consistent with the original experiment. And all CNN are trained from scratch using the BabyFace dataset. The experimental results are shown in Table 4. Figure 4 shows the loss and accuracy of TSSEAug methods.



(a) Loss (TSSEAug)



(b) Accuracy (TSSEAug)

Fig.4. Training loss and accuracy of TSSEAug method.

5.3 Results Analysis

5.3.1 Age Estimation Results

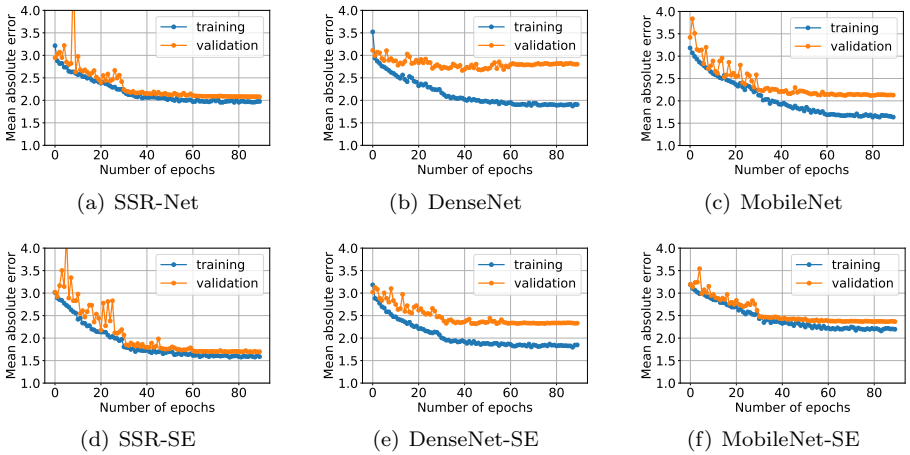
MAE represents the average of the absolute errors between the predicted age and the ground truth over all test samples. Figure 5 demonstrate MAE values with different epoch comparing the training process of SSR-Net, MobileNet, DenseNet and the proposed SSR-SE, MobileNet-SE, DenseNet-SE on the novel BabyFace dataset. In these six sub-figures, the blue curve and orange curve denote the progress of the training error and validation error in MAE, respectively. As stated in [14], if the two curves are close together, the model derived from the training data can be better applied to the validation data and models with this property are less affected by overfitting. From this perspective, in consistency with previous experimental results, SSR-Net outperforms MobileNet and DenseNet. And SSR-Net-SE outperforms MobileNet-SE and DenseNet-SE. This means that the model we train on the training set can be more successfully applied to the validation set.

By comparing the results in Figures 5(a), 5(b), 5(c) with the results in Figures 5(d), 5(e), 5(f), we can clearly see the importance of the SE module refining CNN features and highlighting salient aging region, which adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. Unlike adult age estimation, the unit for estimating the babies? age is in months. For the validation set, MAE $\approx$ 1.7. The result indicates that the model miss a baby's age by a margin of less than 2 months on average. With its compact size and efficiency, SSR-SE is suitable to be employed on mobile or embedded devices for age estimation on baby applications.

5.3.2 Gender Estimation Results

As can be seen from Table 4, the accuracy of our proposed TSSEAug method is the best and the experimental accuracy is significantly higher than the other three one stream network structures, indicating that two stream network architecture could explore different features and their fusion could improve the performance. We have two observations of the gender classification on BabyFace.

Firstly, the TSSEAug network structure obtains the best accuracy rate 78.3%, while TSSE method obtains 72.1%, which shows random erasing and mixup preprocessing and regularization and image augmentation techniques could be applied to the model to bring it out of the current slightly overfitting state and significantly



(a) SSR-Net (b) DenseNet (c) MobileNet

(d) SSR-SE (e) DenseNet-SE (f) MobileNet-SE

Fig.5. Comparision of BabyFace age validation for SSR-Net, DenseNet, MobileNet, SSR-SE, DenseNet-SE, MobileNet-SE.

improves the accuracy by 6.2% of the model. while Two Stream method gets 65.0%, 7.1% lower than the TSSE method, which shows the Squeeze-and-Excitation (SE) block for refining CNN features and highlighting salient aging region adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels.

Table 4. Gender estimation accuracy results on the BabyFace dataset.

Methods	Accuracy(%)
CNN_AgeGender [1]	68.9
R-CNN_Gender [30]	50.1
HyperFace [30]	67.1
Slighthound [31]	51.0
One Stream	58.5
OSSE (One Stream+SE)	68.1
OSSEAug (One Stream+SE+Aug)	66.7
Two Stream	65.0
TSSE (Two Stream+SE)	72.1
TSSEAug (Two Stream+SE+Aug)	**78.3**

Secondly, the state-of-the-art gender estimation methods on the BabyFace dataset can not perform as effectively as they perform on other datasets. For example, the R-CNN_Gender achieves 95% and 91% accuracy in the CelebA [32] and LFWA [6] datasets, respectively, while only 50.1% on BabyFace. HyperFace achieves 97% and 94% accuracy in the CelebA and LFWA datasets, respectively, while only 67.1% on BabyFace. Also, the Slighthound method achieves 91% gender recognition on the Adience benchmark, but the performance is only 51.0% on our BabyFace dataset. Even the CNN_AgeGender method, which achieves relatively high accuracy gender estimation on the BabyFace dataset with only 68.9% accuracy rate, while achieving 86.8% accuracy on the Adidence benchmark.

The significant decreasing of performances of the state-of-the-art methods has shown that the gender estimation for babies are very different from the same task for adults. This result is in good agreement with the experiment in [1]. There are frequent false gender estimations on babies or toddlers' images in their gender estimation experiments in [1]. Because the gender characteristics are not obvious on these babies or toddlers. In terms of gender estimation, the performance of the R-CNN_Gender (50.1%) and the Slighthound (51.0%) are only a slightly better than random guess (50%). This is mainly because adult gender estimation is relatively easy due to some physiological characteristics. However, facial development and hair style during infancy are in a gender-insensitive nature. This also proves that the lack of datasets for babies in previous datasets has resulted in a lack of generalization of gender estimation methods.

6 Conclusion

Age and gender estimation is a difficult problem for many researchers who are looking for a solution, because its application in the real world is very valuable. However, most of existing works focus mainly on analyzing an adult' face. We collect a baby face dataset, named BabyFace, which contains 15,528 images from 5,872 babies with age ranging from 0 to 24 months. To the best of our knowledge, it is the largest dataset for baby's age and gender estimation. In order to faciliate further research, we also benchmark algorithms on this novel dataset. While our age estimation model miss a baby's age by a margin of less than 2 months, the proposed gender estimation structure obtaines the best accuracy rate, i.e., 78.3%. In addition, we provide a comparision of age and gender estimation by extensively evaluate the performance of state-of-the-art methods on baby's faces. Our experiments show that estimating a baby'age is not as difficult as adult's imagination. However, the gender estimation for babies are very different because gender features such as beard eyebrows, pores are not visible in a baby' face. Our research shed light on that recognizing a baby's gender and age is non-trivial and new methods are necessary to be developed to approach baby face recognition. We hope that our research will encourage more research works on the effect of real-world baby age and gender estimation and turning attention from adult's faces into an interesting focus on analyzing the baby's faces.

References

[1] Levi G, Hassner T. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42. 1, 3, 4, 8, 10

[2] Das A, Dantcheva A, Bremond F. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0. 1, 4

[3] Hosseini S, Lee S H, Kwon H J, Koo H I, Cho N I. Age and gender classification using wide convolutional neural network and gabor filter. In *2018 International Workshop on Advanced Image Technology (IWAIT)*, 2018, pp. 1–3. 1

[4] Agbo-Ajala O, Viriri S. Age group and gender classification of unconstrained faces. In *International Symposium on Visual Computing*, 2019, pp. 418–429. 1

[5] Eidinger E, Enbar R, Hassner T. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 2014, 9(12):2170–2179. 1, 4

[6] Huang G B, Mattar M, Berg T, Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 1, 10

[7] Panis G, Lanitis A, Tsapatsoulis N, Cootes T F. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 2016, 5(2):37–46. 1, 2

[8] Ricanek K, Tesafaye T. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, pp. 341–345. 1, 2

[9] Rothe R, Timofte R, Gool L. Dex: Deep expectation of apparent age from a single image[c]. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 01 2015, pp. 10–15. 1, 3

[10] Cao Q, Li S, Xie W, Parkhi O M, Zisserman A. Vggface2: A dataset for recognising faces across pose and age. 2018. 1, 3

[11] Phillips D, McCartney K, Scarr S. Child-care quality and children's social development. *Developmental psychology*, 1987, 23(4):537. 1

[12] Letourneau N. *Scientific parenting: What science reveals about parental influence*. Dundurn, 2013. 1

[13] Simpson A R. The role of the mass media in parenting education. 1997. 1

[14] Yang T Y, Huang Y H, Lin Y Y, Hsiu P C, Chuang Y Y. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, 2018, p. 7. 2, 6, 7, 9

[15] Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 2, 7

[16] Zhang H, Cissé M, Dauphin Y N, Lopez-Paz D. mixup: beyond empirical risk minimization. corr abs/1710.09412 (2017). *arXiv preprint arXiv:1710.09412*, 2017. 2, 7

[17] Deng J, Dong W, Socher R, Li L J, Li K, Li F F. Imagenet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision I& Pattern Recognition*, 2009. 2

[18] Chen B C, Chen C S, Hsu W H. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, 2014, pp. 768–783. 3

[19] Yi D, Lei Z, Li S Z. Age estimation by multi-scale convolutional network. In *Asian conference on computer vision*, 2014, pp. 144–158. 3

[20] Rothe R, Timofte R, Van Gool L. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 2018, 126(2-4):144–157. 3

[21] Wang X, Guo R, Kambhamettu C. Deeply-learned feature for age estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 534–541. 3

[22] Niu Z, Zhou M, Wang L, Gao X, Hua G. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4920–4928. 3

[23] Lapuschkin S, Binder A, Muller K R, Samek W. Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1629–1638. 4

[24] Verma A, Vig L. Using convolutional neural networks to discover cogntively validated features for gender classification. In *2014 International Conference on Soft Computing and Machine Intelligence*, 2014, pp. 33–37. 4

[25] Mansanet J, Albiol A, Paredes R. Local deep neural networks for gender recognition. *Pattern Recognition Letters*, 2016, 70:80–86. 4

[26] King D E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009, 10(Jul):1755–1758. 4

[27] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004, 13(4):600–612. 5

[28] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 6

[29] Kingma D P, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[30] Ranjan R, Patel V, Chellappa R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(1):121–135. 8, 10

[31] Dehghan A, Ortiz E G, Shu G, Masood S Z. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*, 2017. 8, 10

[32] Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738. 10