sensors

MDPI

Article

Age Estimation of Faces in Videos Using Head Pose Estimation and Convolutional Neural Networks

Beichen Zhang * and Yue Bao

Visual Media Laboratory, Department of Information Science, Tokyo City University, Tokyo 1588557, Japan; bao@g.tcu.ac.jp
* Correspondence: g1991804@tcu.ac.jp

Abstract: Age estimation from human faces is an important yet challenging task in computer vision because of the large differences between physical age and apparent age. Due to the differences including races, genders, and other factors, the performance of a learning method for this task strongly depends on the training data. Although many inspiring works have focused on the age estimation of a single human face through deep learning, the existing methods still have lower performance when dealing with faces in videos because of the differences in head pose between frames, which can lead to greatly different results. In this paper, a combined system of age estimation and head pose estimation is proposed to improve the performance of age estimation from faces in videos. We use deep regression forests (DRFs) to estimate the age of facial images, while a multiloss convolutional neural network is also utilized to estimate the head pose. Accordingly, we estimate the age of faces only for head poses within a set degree threshold to enable value refinement. First, we divided the images in the Cross-Age Celebrity Dataset (CACD) and the Asian Face Age Dataset (AFAD) according to the estimated head pose degrees and generated separate age estimates for images with different poses. The experimental results showed that the accuracy of age estimation from frontal facial images was better than that for faces at different angles, thus demonstrating the effect of head pose on age estimation. Further experiments were conducted on several videos to estimate the age of the same person with his or her face at different angles, and the results show that our proposed combined system can provide more precise and reliable age estimates than a system without head pose estimation.

Keywords: age estimation; deep learning; CNN; head pose estimation

Citation: Zhang, B.; Bao, Y. Age Estimation of Faces in Videos Using Head Pose Estimation and Convolutional Neural Networks. *Sensors* **2022**, *22*, 4171. https://doi.org/10.3390/s22114171

Academic Editor: Qingming Huang

Received: 25 April 2022
Accepted: 27 May 2022
Published: 31 May 2022

1. Introduction

Age estimation from a facial image has become an important yet challenging problem in many applications, such as human–computer interaction [1–3], identification [4], security [5], and precision advertising [6].

Although there have been a great deal of studies on the age estimation issue, the performance of age estimation from facial images is still a huge gap with real-life application demands in terms of both accuracy and stability. The reasons that make age estimation such a challenging problem come from two groups: Objective conditions of the external environment that include illumination, distance, pose, perspective, and expression [7]; physiological conditions of intrinsic features that include ethnicity, gender, and health status [8]. Previous studies carried out a lot of work on the external conditions since the intrinsic features can not be normalized and are therefore difficult to solve. The intrinsic facial features are always inhomogeneous for several reasons: (1) even people of the same age can exhibit enormous variation in facial appearance (Figure 1); and (2) in different periods of age, human faces change differently. For example, children usually have a fast speed of bone growth; on the other hand, adults' faces change very slowly [9] (Figure 2). Consequently, it is a difficult problem to make an estimator which could predict real age of human face images accurately from such widely diverse appearance factors.

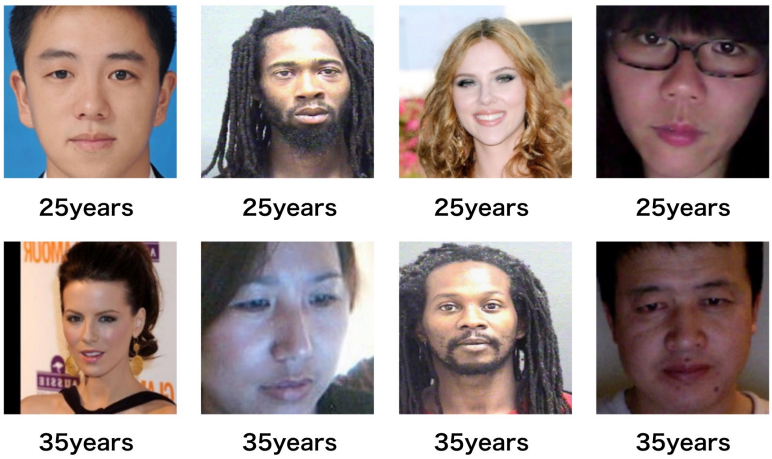

Figure 1. Differences between different people of the same age.

Figure 2. Changes in facial appearance from childhood to adulthood.

In recent years, deep learning has led to impressive works on various computer vision tasks, including age estimation [10–12]. However, all these works have used datasets including only frontal facial images, which cannot adequately reflect the conditions of real-life applications. Different from most facial images in datasets, the head pose may vary greatly in videos or webcam streams, leading to intolerable errors in the estimated age.

In this work, a combined system of age estimation and head pose estimation is proposed to solve the problem of age estimation from faces in videos or webcam streams. First, we use deep regression forests (DRFs) [11] to estimate the age of facial images, which can achieve high precision for frontal facial images. Meanwhile, a multiloss convolutional neural network (CNN) is also utilized to estimate the head pose [13]. Then, we can use the trained system to estimate age and head pose from several videos frame by frame. When using the trained mapping between age and head pose, we set a degree threshold for the head pose and perform age estimation only for frames where the head pose is within this threshold to enable value refinement of the age estimated from the video.

Experiments were conducted in two phases. First, a multiloss CNN was trained on the 300W-LP dataset [14] for head pose estimation. We also divided the Cross-Age Celebrity Dataset (CACD) [15] and the Asian Face Age Dataset (AFAD) [12] based on the estimated head pose angles and trained DRFs separately on the subsets of frontal and nonfrontal images. The results showed that the accuracy of age estimation from frontal facial images was better than that for faces at different angles. Then, we tested the trained models on several videos to estimate the age of the same person with his or her face at different angles. The experimental results demonstrate that our proposed system with a head pose angle constraint achieves a standard deviation of the estimation errors for videos that is smaller than what can be achieved when performing age estimation alone. The results show that our approach improves the precision and reliability of age estimation for faces in videos compared to traditional methods.

This paper consists of five sections. Section 2 introduces some related works of age estimation, Section 3 presents the proposed method with the whole architecture and several details, the experimental results for the proposed method are discussed in Section 4, and a conclusion and a discussion of future work can be found in Section 5.

2. Related Work

Deep learning methods are used in age estimation because of their great success in many computer vision tasks. Similar to [16], Yi et al. [17] used CNN for age estimation with the features extracted from different regions of the face, and introduced the mean squared loss as the measurement criterion. Niu. Z et al. [12] noticed the continuous feature of age and trained an ordinal CNN; multiple binary outputs were also used for better performance. Another use of continuous information comes from [10] with multiple binary neural networks; the multiple outputs were aggregated as final result. Ref. [18] used softmax fuction in another way, in which softmax outputs of each neuron were used as weights of age, and a weighted average value was calculated instead of using the softmax classification result directly; the experiment results showed better performance. Multi-task learning methods were used for age estimation in [19,20]; several other facial features were jointly learned and enhanced the performance of each task. Deep regression forests (DRFs) [11] used random regression forests coupled with CNN and obtained better performance.

For age estimation from faces in videos, the most closely related work is the deep age estimation model [21], in which Ji et al. used a CNN with an attention mechanism. Facial features were extracted by CNN then aggregated from features vectors to a single feature by an attention block. They trained the model using a new loss function, leading to better precision and stability across every frame for age estimation.

Another work for age estimation where static and dynamic features can be learned from expressions of face simultaneously in videos is called the spatially-indexed attention model (SIAM) [22]. In this model, Ji Pei et al. employed CNNs to extract the latent appearance features from each frame and then used recurrent networks to process all the features to simulate time dynamics. Furthermore, they used a specifically designed spatially indexed attention mechanism, and all the accentuated facial areas in each frame could be extracted by convolutional layers. A time attention layer was also used to allocate attention weights at each frame. This method focuses on both frames and face areas with important information, resulting in better performance. The relevance between spatial facial areas and time frames, as well as age estimation, can also be revealed.

However, Ji et al. used continuous frames as input data, rather than using a single image, in order to guarantee stability, which increased the computation and made the network more complicated. In addition, to train this model, a new dataset must be collected with labels' annotation, causing more time consumption. The SIAM method has limitations in terms of which types of facial expression images it can consider; specifically, only smile and disgust databases were used in experiments. Therefore, we want to propose a new approach that can be trained with all types of facial images from existing databases using a single image as input to make the age estimation easy and feasible.

3. Proposed Method

In this section, each step of the system flow will be explained in detail.

3.1. Datasets

In these experiments, we used two datasets containing different racial groups for age estimation training and one dataset for head pose estimation. Figure 3 depicts exemplar images from each dataset for age estimation.

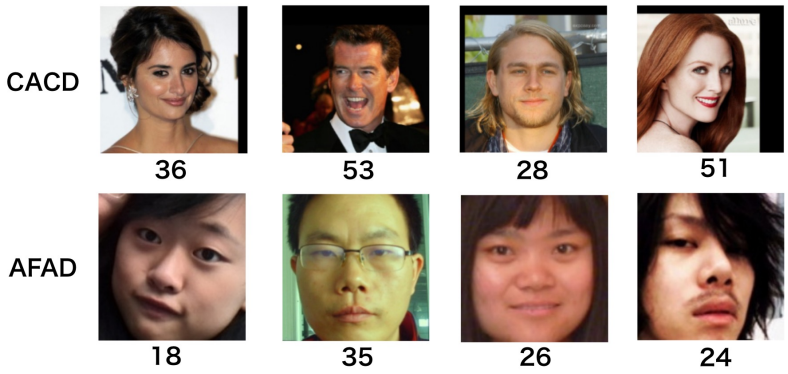

Figure 3. Examples from Cross-Age Celebrity Dataset [15] and Asian Face Age Dataset [12]. The number below each image is the ground truth age of the subject.

Cross-Age Celebrity Dataset: The CACD dataset, released in 2014 by the University of Maryland Computer Science Department [15], is a large-scale dataset for face recognition and retrieval across ages. It contains 163,446 images of 2000 celebrities. Images were collected by search engines using keywords of the celebrity's name and a year (2004–2013). The age of the celebrity on the image can be estimated by simply subtracting the year of birth from the year the photo was taken. There are training, validation, and testing parts of the dataset and the training part is very noisy. Therefore, in our experiments, we only used a cleaned subset which was hand-selected to 18,171 photographs. For evaluation, the dataset was randomly divided into 85% for training and 15% for testing.

Asian Face Age Dataset: AFAD [12] released this in 2016 for age recognition, containing 164,432 images of faces with accurate age and gender labels. As the Asian Face Age Dataset (AFAD), all of the images from the dataset are Asian faces. AFAD was built by collecting selfie photos from the Renren social network (RSN). Not only do a large number of Asian students from middle school to graduate school use RSN frequently, but plenty of graduated students also use RSN to contact their classmates. Therefore, the ages of RSN users span a wide range from 15 to more than 40 years old. We used a subset of AFAD with about 60k images of people from 18 to 39, and the subset was balanced for training. For evaluation, the dataset was randomly divided into 85% for training and 15% for testing.

300W across Large Poses: The 300W dataset [14] uses 68 landmarks to standardize multiple alignment databases, including AFW [23], LFPW [24], HELEN [25], IBUG [26], and XM2VTS [27]. A face profiling method was applied to 300W, and about 60 thousand pose data were extracted (about 1800 in IBUG, 5200 in AFW, 16,000 in LFPW, and 37,000 in HELEN; the data from XM2VTS were not in use). All the data were flipped, and therefore multiplied to 120 thousand. The resulting dataset is called 300W-LP (300W across Large Poses).

3.2. Face Alignment

Detection performance will be changed as well as the surroundings of the face change. The different types of face alignments could result in additional performance changes. An ideal facial image should have similar size, with front view, face centering, and the face alignment normalized with fixed location and cleaned background. Therefore, we chose the multi-task cascaded convolutional network (MTCNN) [28] face detector to obtain the face from an image. In order to minimize the impact of surrounding pixels, we resized all the images to 256×256 and made a random crop to 224×224. The crop process makes the face randomly located at a different position in the image, regardless of the originating data. This approach can improve the robustness of our model to figure out the problem of various scenes with different face alignments. The 224×224 pixel image also fits the input size of VGG-16 network.

3.3. Age Estimation

Figure 4 shows a diagram of a DRF [11]. A CNN combined with deep regression forests is introduced in this work and estimates the real age from facial image. The model is trained on facial image datasets with known ages and face landmarks as labels. The training process in this paper begins with the pretrained weights from the ImageNet dataset, as with the same model used in [29]. Then, the CNN is fine-tuned on the two target datasets used for age estimation. The fine-tuning process makes the CNN obtain the features, distribution, and bias of each dataset and optimizes the performance.

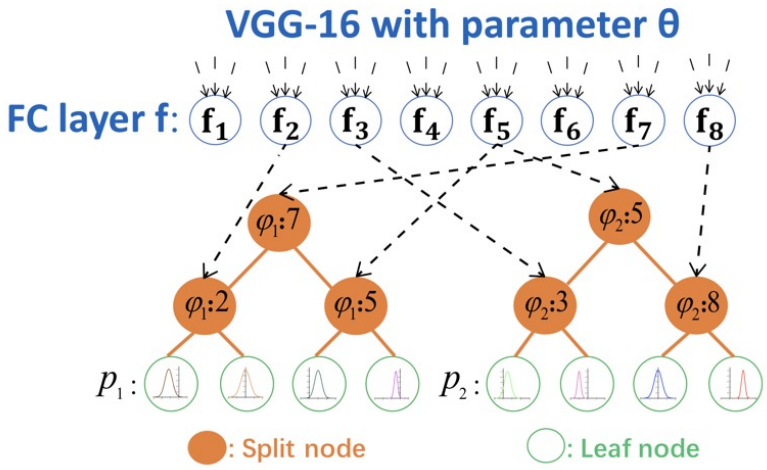

Figure 4. Illustration of deep regression forests.

The upper blue circles represent the output neurons from the CNN defined by the function **f** with parameter θ. All these neurons come from the last fully-connected (FC) layer of VGG-16. The middle orange circles represent the split nodes and the bottom green circles represent the leaf nodes of deep regression forests. φ_1 and φ_2 represent the index functions of each tree. The black dashed arrows point out the correspondence from the split nodes of each tree to the neurons of VGG-16 FC layer. Each neuron may correspond to the split nodes of different trees. Each tree has its own distribution π for its leaf node (represented by the distribution curves on the leaf nodes). The final output for the whole forest can be calculated as the mix of the predictions of the individual trees. The parameters $f(\cdot;\Theta)$ and π will be trained simultaneously end-to-end.

VGG-16: The VGG-16 CNN architecture was selected since, first, the architecture of VGG-16 is deep, representing high performance, but also manageable, representing expandability; secondly, Russakovsky et al. have achieved impressive work [30] with the VGG-16 model on the ImageNet challenge; and thirdly, pretrained models for classification of VGG-16 are publicly available, which can accelerate the process of training. VGG-16 network is much deeper than previous architectures, for example, AlexNet [31], specifically, consisting of 13 convolutional layers and 3 FC layers. It is characterized by a number of 3×3 filters with convolution kernel and the stride of filters are set to 1; within comparison, AlexNet has much larger filters (up to 11×11), and the stride of filters are set to 4. Therefore, each convolution filter from VGG-16 has simpler geometry, but the increased depth also allows much more complexity.

Deep Regression Tree: DRFs are combination of several deep regression trees. For each tree, three are input–output pairs $\{x_i, y_i\}_{n=1}^{N}$, in which $x_i \in \mathbb{R}^{D_x}$ and $y_i \in \mathbb{R}$. A deep regression tree model describes the mapping relationships from input to output over CNNs connect with a regression tree. A deep regression tree $\mathcal{T}$ has a number of split nodes $\mathcal{N}$ and leaf nodes $\mathcal{L}$. To be specific, an input x_i will be passed to the left or right node relative to one node, which will be decided by each split node $n \in \mathcal{N}$; while for a leaf node $\ell \in \mathcal{L}$,

it can be described by a Gaussian distribution, where $p_\ell(y_i)$ represents the mean and μ_l represents the variance σ_l^2 of Gaussian distribution.

Split Node: A split node is associated with a splitting function $S_n(x_i;\Theta) : x_i \rightarrow [0,1]$, which is parameterized by Θ—the parameters of CNNs. Normally, this splitting function is defined as $s_n(x_i;\Theta) = \sigma(f_{\varphi(n)}(x_i;\Theta))$, where $\sigma(\cdot)$ represents the sigmoid function, $\varphi(\cdot)$ represents an index function to point out the $\varphi(n)$ element of $f(x_i;\Theta)$ consistent with the split node n, and $f(x_i;\Theta)$ are the learned deep features. Figure 4 illustrates the simple diagram of the DRFs, where φ_1 and φ_2 represent the index function of each tree. For a given x_i, the probability of reaching the leaf node l can be calculated as

$$\omega_\ell(x_i|\Theta) = \prod_{n\in\mathcal{N}} s_n(x_i;\Theta)^{[\ell\in\mathcal{L}_{n_l}]}(1 - s_n(x_i;\Theta))^{[\ell\in\mathcal{L}_{n_r}]} \quad (1)$$

Here, $\mathcal{L}_{n_l}$ and $\mathcal{L}_{n_r}$ are the sets of leaf nodes belonging to the subtrees $\mathcal{T}_{n_l}$ and $\mathcal{T}_{n_r}$. Subtree $\mathcal{T}_{n_l}$ means that the root of the tree is the left children n_l of node n and $\mathcal{T}_{n_l}$ means that the root of the tree is the right children n_r of node n.

Leaf Node: Consider a tree $\mathcal{T}$; for each input x_i, an $\ell \in \mathcal{L}$ leaf node represents a predictive distribution on y_i, denoted by $p_\ell(y_i)$. Specifically, there we assumed that $p_\ell(y_i)$ is in obedience to the Gaussian distribution: $\mathcal{N}(y_i|\mu_l, \sigma_l^2)$. Therefore, the final distribution with the conditional probability of y_i on x_i can be calculated by averaging the probability of the route to each leaf node:

$$p_\mathcal{T}(y_i|x_i;\Theta,\pi) = \sum_{\ell\in\mathcal{L}} \omega_\ell(x_i|\Theta)p_\ell(y_i) \quad (2)$$

where Θ are the parameters from CNNs and π are the distribution parameters $\{\mu_l, \sigma_{l2}\}$. This distribution can be considered as a mixed distribution, in which $\omega_\ell(x_i|\Theta)$ are the mixing coefficients and $p_\ell(y_i)$ represents the Gaussian distributions at the $\ell^t h$ leaf node. The π has different value for each tree; therefore, π_k is used with the corresponding index in the subsequent part.

Deep Regression Forests: Deep regression forests are combinations of several deep regression trees, $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$; the final output distribution of prediction can be calculated by an input x_i, as the average of all trees:

$$p_\mathcal{F}(y_i|x_i,\Theta,\Pi) = \frac{1}{N}\sum_{n=1}^{N} p_{\mathcal{T}_n}(y_i|x_i,\Theta,\pi_n) \quad (3)$$

where N represents the total number of trees and $\Pi = \{\pi_1, \ldots, \pi_N\}$. $p_\mathcal{F}(y_i|x_i,\Theta,\Pi)$ represents the possibility when the ith input yields output of y_i.

3.4. Head Pose Estimation

In most works on predicting head pose using convolutional networks, the easiest way is using a mean squared error loss, and the output angles of head pose have been regressed directly. However, this approach fails to meet adequate performance requirements on the dataset we wish to use for age estimation.

Therefore, we adopted Ruiz's method [13], in which deep multiloss CNNs are trained for head pose estimation with satisfactory accuracy. The ResNet50 network [32] was introduced for head pose estimation and three losses are used for three angles separately. There are two parts of each loss: the mean squared error regressed directly and the cross-entropy loss from classification of pose. There are three FC layers being used for three angles and sharing the previous parts of the network. By adopting additional cross-entropy losses from classification, we constructed three signals to be backpropagated to improve the learning process. The predictions of three output angles were computed as the final head pose results. The details of the architecture are shown in Figure 5.

Mean Absolute Error: To evaluate the performance of different age estimation algorithms, as a criterion of measurement for age estimation algorithms, mean absolute error

(MAE) metric is used for the estimation. By calculating the average absolute error between the precited age and the ground truth age, the defining equation of MAE is

$$MAE = \frac{1}{K}\sum_{i=1}^{K}|\tilde{x}_i - x_i|$$

where K is the number of samples, x_i is the ground truth age of the i-th sample, and $\tilde{x}_i$ is the predicted age of the i-th sample. A small MAE represents great performance of age estimation.

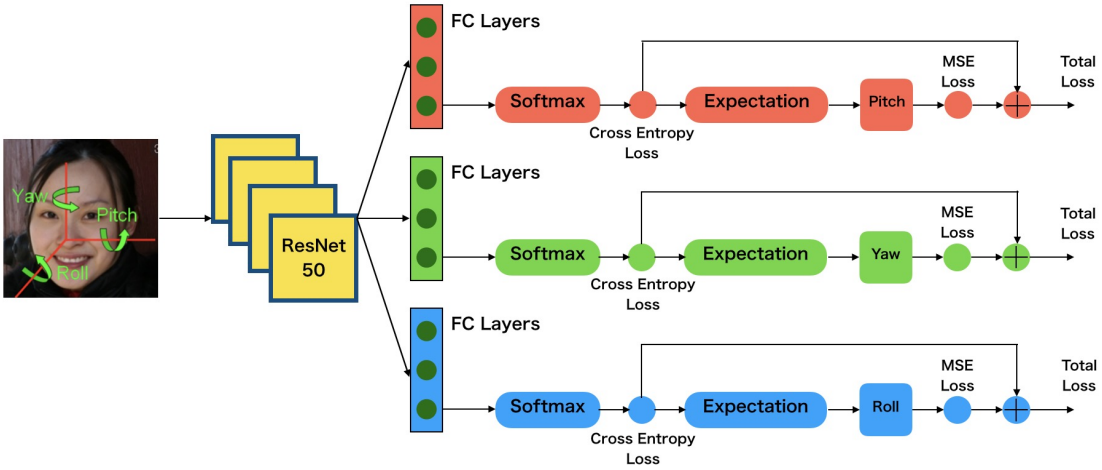

Figure 5. CNN with combined mean squared error and cross-entropy losses.

4. Experiments

In the following section, the implementation details of experiments are presented along with their quantitative and qualitative results. It concludes with a discussion on the findings.

4.1. Implementation Details

For each experiment, we used the existing weights for VGG16 as the initial value from ImageNet. The training parameters of the neural network are listed as follows: the batch size of training data is 64, the ratio of dropout layer is set to 0.5, the stochastic gradient descent (SGD) is used as gradient descent method, the learning rate is set to 0.2 as an initial value, and reduces by half per 5k iterations. The training parameters of the regression forests are listed as follows: the number of trees is set to 4, the depth of each tree is set to 5, the number of output unit is set to 64, the value of leaf node will be updated per 10 iterations, and the prediction result from leaf nodes will be updated per 30 iterations. This model is then fine-tuned with CACD and AFAD for age estimation. ResNet50 was trained on the 300W-LP dataset for head pose estimation, and the training parameters of the ResNet50 are listed as follows: the Adam optimization is used as gradient descent method, and the learning rate is set to 10^{-5} with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

During the training phase, the training data are split as follows: 80% for training and 20% for validation. The training process will be aborted early when the model has been overfit on the validation set. The models were trained on Nvidia GTX 1080 GPUs.

4.2. Results and Comparison

First, a multiloss CNN was trained on the 300W-LP dataset for head pose estimation. Subsequently, we divided the images into AFAD and CACD based on the estimated head pose angles and trained DRFs separately on the subsets of frontal and nonfrontal images. Then, we trained DRFs on several subsets of CACD and AFAD with different threshold. Finally, we tested the models on two facial video datasets to estimate the age of the same person with his or her face at different angles and compared the results with those of

previous methods. The same network structure and training strategy were used to ensure fair comparisons.

4.2.1. Head Pose Estimation

We trained the adopted multiloss CNN on the 300W-LP dataset in order to make the head pose estimation for age estimation. To verify the performance of the head pose estimation method, we tested the model on a subset of 300W-LP called AFLW2000 [14] which have images cropping around the face area with small size. The AFLW2000 dataset have marked ground truth landmarks; therefore we compared our method with it and other methods, such as commonly used detectors FAN [33] and Dlib [34]. The quantitative results can be seen in Table 1. Although our method is not the best, it is better than traditional detectors and is suitable for our combined system.

Table 1. Mean average error of Euler angles across different methods on the AFLW2000 dataset.

Methods	Yaw	Pitch	Roll	Average
Dlib [34]	23.153	13.633	10.545	15.777
Fan [33]	6.358	12.277	8.714	9.116
CPAM [35]	1.479	1.804	1.869	1.697
Multiloss CNN	**6.470**	**6.559**	**5.436**	**6.155**
Ground truth landmarks	5.924	11.756	8.271	8.651

4.2.2. Testing on Facial Image Datasets

In this section, the performance of DRFs for age estimation based on frontal and nonfrontal facial images is presented. The frequently used AFAD and CACD datasets, representing Asians and Europeans, respectively, were used in this experiment. We used the trained multiloss CNN to estimate the head poses in both datasets. For each facial image, three rotational angles were estimated, one on each axis. We set 30 degrees as the threshold for the sum of the three angles, and images with head pose angle estimates summing to more than 30 degrees were defined as nonfrontal images. Figure 6 depicts exemplar images of nonfrontal facial images from the datasets.

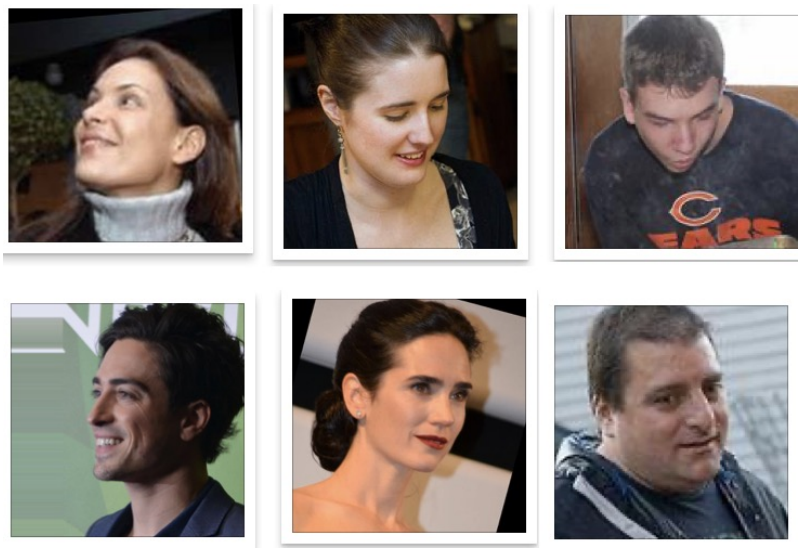

Figure 6. Examples of nonfrontal facial images.

Based on the estimated angles, AFAD was divided into frontal and nonfrontal subsets consisting of 53,983 and 5361 images, and CACD was divided into frontal and nonfrontal subsets consisting of 15,145 and 3026 images, respectively. Both subsets were randomly split into training/test (85%/15%) sets, and the training process was repeated five times with different random separation; the final outcome is the average of five times' outputs.

The quantitative results are summarized in Table 2. The results show that the accuracy of age estimation from frontal facial images is significantly better than that for nonfrontal images.

Table 2. Performance (MAE) comparison on the frontal and nonfrontal subsets of AFAD [12] and CACD.

Subset	AFAD	CACD
Frontal	3.73	4.59
Nonfrontal	4.97	5.65

4.2.3. Testing with Different Threshold

In this section, we separate the CACD and AFAD dataset into several subsets with different thresholds of the head pose degrees. Based on the estimated angles, we set the threshold from 50 degrees to 10 degrees with step of 10 degrees; when the threshold becomes more strict, the number of samples of head pose degree within the threshold become smaller. The correspondence between the threshold and the number of samples, as well as the performance of age estimation, are summarized in Table 3. When the threshold is smaller than 30 degrees, the number of samples reduces rapidly but the performance of age estimation is almost unchanged. Therefore, we chose 30 degrees as the threshold to obtain the best trade-off between performance and number of samples.

Table 3. Performance (MAE) and image numbers comparison on the AFAD [12] and CACD with different threshold.

Threshold (degree)	AFAD		CACD	
	MAE	Number	MAE	Number
50	3.97	59,173	4.87	18,023
40	3.84	57,232	4.70	16,842
30	3.73	53,983	4.59	15,145
20	3.72	36,748	4.58	10,398
10	3.71	18,753	4.58	7569

4.2.4. Testing on Facial Video Datasets

Two new facial video datasets were constructed to evaluate our model in terms of age estimation performance. We collected 18,282 and 18,944 frames from two twelve-minute facial videos of Asian and European subjects, respectively. It should be noted that each facial video dataset was collected from the same person, and these datasets were used only for evaluating the age estimation models; currently, there is no facial video dataset available to be used for training the whole model. We first trained DRFs on AFAD and CACD, representing Asians and Europeans, respectively. Then, we tested the two trained models on the facial video datasets with simultaneous head pose estimation. Examples of the test images are shown in Figure 7. We performed age estimation only for faces with head poses within 30 degrees, and we compared the results with the results for all images without head pose restrictions. Several other models were also trained on AFAD and CACD and then tested on the facial video datasets for more comprehensive comparisons.

We trained a DRF on AFAD and tested the model on the Asian video dataset with head pose restrictions. We also trained a DRF on CACD and tested the model on the European video dataset with head pose restrictions. We compared the results of our method with those of other outstanding age estimation models, and the quantitative results are summarized in Table 4. All models were trained on AFAD and CACD with the same training strategy to ensure fair comparisons. On the task of facial video estimation, our method achieves the best MAE, 5.12, of the Asian facial video dataset and the best MAE, 5.56, of the European facial video dataset. The variance is reduced by 0.62 on the Asian facial video dataset and 1.53 on the European facial video dataset compared to the best

existing method. From the results, although the accuracy and performance are different on different datasets, our proposed method can achieve better MAE and variance compared to other methods.

Figure 7. Examples from the facial video datasets with age and head pose estimates. The numbers represent the predicted age. Green and red colors indicate that the sum of the head pose rotational angles is less than and greater than 30 degrees, respectively.

Table 4. Accuracy (MAE) and variance results for comparison with state-of-the-art methods on the Asian and European facial video datasets.

Method	Asian		European	
	MAE	**Variance**	**MAE**	**Variance**
AlexNet [31]	6.19	6.92	6.93	7.15
DEX [18]	6.72	8.65	7.17	8.22
DRF [11]	5.96	4.12	6.39	5.84
Our method	**5.12**	**3.50**	**5.56**	**4.31**

5. Conclusions

In this paper, a combined system of age estimation and head pose estimation is proposed to solve the problem of age estimation based on faces in videos or webcam streams, where different head poses may lead to intolerable errors on the estimated ages. Experimental results show that with a head pose restriction such that age estimation is performed only for facial images with head poses within a specified degree threshold to ensure value refinement, our method achieves promising improvements in accuracy and stability for age estimation from video.

The main contributions of this paper are as follows: (1) We are the first to couple age estimation and head pose estimation for age estimation in videos; (2) our method shows significantly improved performance in age estimation on facial video datasets compared to other state-of-the-art methods in terms of both accuracy (MAE) and variance.

However, we only tested our method on two datasets and there might be some video that does not contain any frames that meet our frontal view criteria. In future work, we would collect and annotate more facial images from videos and create a new database including more people. The new database could be trained with our method and obtain more reliable and robust results. We would also attempt to calibrate the nonfrontal samples, instead of just not using them, to make our system widely available.

Author Contributions: Conceptualization, Y.B.; methodology, B.Z.; software, B.Z.; validation, B.Z.; data curation, B.Z.; writing—original draft preparation, B.Z.; writing—review and editing, Y.B.; visualization, B.Z.; supervision, Y.B.; project administration, Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Links to datasets used in this paper. CACD: https://bcsiriuschen.github.io/CARC/. AFAD: https://afad-dataset.github.io. 300W-LP: http://www.cbsr.ia.ac.cn/users/xiangyuzhu/projects/3DDFA/main.htm (accessed on 17 April 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DRFs	Deep regression forests
CACD	Cross-Age Celebrity Dataset
AFAD	Asian Face Age Dataset
CNN	Convolutional neural network
SIAM	Spatially-indexed attention model
FC	Fully-connected
MAE	Mean absolute error
SGD	Stochastic gradient descent

References

1. Lanitis, A.; Draganova, C.; Christodoulou, C. Comparing different classifiers for automatic age estimation. *IEEE Trans. Syst. Man Cybern. Part Cybern.* **2004**, *34*, 621–628. [CrossRef] [PubMed]
2. Han, H.; Otto, C.; Jain, A.K. Age estimation from face images: Human vs. machine performance. In Proceedings of the ICB, Madrid, Spain, 4–7 June 2013; pp. 1–8.
3. Geng, X.; Zhou, Z.-H.; Zhang, Y.; Li, G.; Dai, H. Learning from facial aging patterns for automatic age estimation. In Proceedings of the ACM International Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 307–316.
4. Lanitis, A.; Taylor, C.J.; Cootes, T.F. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 442–455. [CrossRef]
5. Song, Z.; Ni, B.; Guo, D.; Sim, T.; Yan, S. Learning universal multi-view age estimator using video context. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 241–248.
6. Shan, C.; Porikli, F.; Xiang, T.; Gong, S. (Eds.) Video Analytics for Business Intelligence. In *Studies in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2012.
7. Han, H.; Otto, C.; Liu, X.; Jain, A.K. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1148–1161. [CrossRef] [PubMed]
8. Guo, G.; Mu, G.; Fu, Y.; Huang, T. Human age estimation using bio-inspired features. In Proceedings of the IEEE CVPR, Miami, FL, USA, 20–25 June 2009; pp. 112–119.
9. Ramanathan, N.; Chellappa, R.; Biswas, S. Age progression in human faces: A survey. *J. Vis. Lang. Comput.* **2009**, *15*, 3349–3361.
10. Chen, S.; Zhang, C.; Dong, M.; Le, J.; Rao, M. Using ranking-CNN for age estimation. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 5183–5192.
11. Shen, W.; Guo, Y.; Wang, Y.; Zhao, K.; Wang, B.; Yuille, A. Deep Regression Forests for Age Estimation. In Proceedings of the IEEE CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2304–2313.
12. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 4920–4928.
13. Ruiz, N.; Chong, E.; Rehg, J.M. Fine-grained head pose estimation without key-points. In Proceedings of the CVPR Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2074–2083.
14. Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; Li, S.Z. Face alignment across large poses: A 3d solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 146–155.
15. Chen, B.; Chen, C.; Hsu, W.H. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimed.* **2015**, *17*, 804–815. [CrossRef]
16. Kwonand, Y.; Lobo, N. Age classification from facial images. In Proceedings of the IEEE CVPR, Seattle, WA, USA, 21–23 June 1994; pp. 762–767.
17. Yi, D.; Lei, Z.; Li, S.Z. Age estimation by multi-scale convolutional network. In Proceedings of the IEEE ICCV, Santiago, Chile, 7–13 December 2015; pp. 144–158.
18. Rothe, R.; Timofte, R.; Gool, L.V. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.* **2016**, *126*, 1–14. [CrossRef]
19. Wang, F.; Han, H.; Shan, S.; Chen, X. Multi-task learning for joint prediction of heterogeneous face attributes. In Proceedings of the IEEE FG, Washington, DC, USA, 30 May–3 June 2017; pp. 173–179.

20. Han, H.; Jain, A.K.; Wang, F.; Shan, S.; Chen, X. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, 40, 2597–2609. [CrossRef] [PubMed]
21. Ji, Z.; Lang, C.; Li, K.; Xing, J. Deep Age Estimation Model Stabilization from Images to Videos. In Proceedings of the International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018.
22. Pei, W.; Dibeklioğlu, H.; Baltrušaitis, T.; Tax, D. Attended End-to-End Architecture for Age Estimation From Facial Expression Videos. *IEEE Trans. Image Process.* **2019**, 29, 1972–1984. [CrossRef] [PubMed]
23. Zhu, X.; Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In Proceedings of the IEEE CVPR, Providence, RI, USA, 16–21 June 2012; pp. 2879–2886.
24. Belhumeur, P.N.; Jacobs, D.W.; Kriegman, D.; Kumar, N. Localizing parts of faces using a consensus of exemplars. In Proceedings of the IEEE CVPR, Colorado Springs, CO, USA, 20–25 June 2011; pp. 545–552.
25. Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; Yin, Q. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In Proceedings of the IEEE ICCVW, Sydney, NSW, Australia, 2–8 December 2013; pp. 386–391.
26. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE ICCVW, Sydney, NSW, Australia, 2–8 December 2013; pp. 397–403.
27. Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G. XM2VTSDB: The extended M2VTS database. In Proceedings of the Second International Conference on Audio and Video-Based Biometric Person Authentication, Washington, DC, USA, 22–24 March 1999; Volume 964, pp. 965–966.
28. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional NETWORKS for large-Scale Image Recognition. CoRR abs/1409.1556. 2014. Available online: https://arxiv.org/abs/1409.1556 (accessed on 30 September 2021).
30. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]
31. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2016**, arXiv:1512.03385.
33. Bulat, A.; Tzimiropoulos, G. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1021–1030.
34. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
35. Singh, T.; Mohadikar, M.; Gite, S.; Patil, S.; Pradhan, B.; Alamri, A. Attention Span Prediction Using Head-Pose Estimation With Deep Neural Networks. *IEEE Access* **2021**, *9*, 142632–142643. [CrossRef]