symmetry

MDPI

Article

Accurate Age Estimation Using Multi-Task Siamese Network-Based Deep Metric Learning for Front Face Images

Yoosoo Jeong, Seungmin Lee, Daejin Park * and Kil Houm Park *

School of Electronics Engineering, Kyungpook National University, Daegu 41566, Korea; ysjung@ee.knu.ac.kr (Y.J.); lsm1106@knu.ac.kr (S.L.)

* Correspondence: boltanut@knu.ac.kr (D.P.); khpark@ee.knu.ac.kr (K.H.P.); Tel.: +82-53-950-5548 (D.P.)

Received: 26 July 2018; Accepted: 4 September 2018; Published: 6 September 2018

Abstract: Recently, there have been many studies on the automatic extraction of facial information using machine learning. Age estimation from front face images is becoming important, with various applications. Our proposed work is based on the binary classifier, which only determines whether two input images are clustered in a similar class, and trains the convolutional neural networks (CNNs) model using the deep metric learning method based on the Siamese network. To converge the results of the training Siamese network, two classes, for which age differences are below a certain level of distance, are considered as the same class, so the ratio of positive database images is increased. The deep metric learning method trains the CNN model to measure similarity based on only age data, but we found that the accumulated gender data can also be used to compare ages. From this experimental fact, we adopted a multi-task learning approach to consider the gender data for more accurate age estimation. In the experiment, we evaluated our approach using MORPH and MegaAge-Asian datasets, and compared gender classification accuracy only using age data from the training images. In addition, from the gender classification, we found that our proposed architecture, which is trained with only age data, performs age comparison by using the self-generated gender feature. The accuracy enhancement by multi-task learning, for the simultaneous consideration of age and gender data, is discussed. Our approach results in the best accuracy among the methods based on deep metric learning on MORPH dataset. Additionally, our method is also the best results compared with the results of the state of art in terms of age estimation on MegaAge Asian and MORPH datasets.

Keywords: convolutional neural network (CNN); deep metric learning; multi-task learning; image classification; age estimation

1. Introduction

The machine learning-based age estimation technique from face images is becoming more and more important, because this is widely used for individual authentication [1], forensic research [2], security control [3], human–computer interaction [3] and social media [4]. Recently, there have been many studies using deep learning based on CNNs [3], such as AlexNet [5], VggNet [6], and Inception [7], with wide use for image classification and image detection. CNN-based learning, as one of the machine learning-based approaches, enables automatic and accurate feature extraction and classification for sample sets; these sets are too large for humans to describe all cases of matching patterns. AlexNet, VggNet, and Inception have been recently used for multi-class classification, and they are widely used as the base models of CNN.

Deep expectation (DEX) [4], an age estimation approach based on CNN models, has been introduced. It uses VggNet to resolve multi-class classification problems for age estimation and

adopts a method to estimate the appropriate age through expectation value calculation for which the trained results in the softmax layer are considered the probability in the corresponding class. Instead of considering the age estimation problem from the perspective of multi-class classification, this approach applies multi-task CNN by considering the age classification problem as a regression-based problem by estimating continuous variables [8].

As another approach, a binary classifier with shallow layers is applied for all classes of age instead of using a CNN model with deep layers. The final age estimation is deducted through the ranking-based comprehensive combination of all results by each binary classifier [9]. This ranking CNN is one of the existing machine learning methods using the cascaded-based combination of the results of binary classifiers.

1.1. Motivation

These approaches aim to estimate absolute age from the input face images directly, but it is not easy to estimate absolute age accurately without any reference data [10]. To overcome this limitation, Abousaleh et al. [11] introduced a new approach, called by comparative region convolutional neural network (CRCNN), which input face images are compared with reference images to determine whether they are older or younger for age estimation. Our study was also inspired by this CRCNN, comparing the age relatively instead of directly estimating absolute age, so we adopted the deep metric learning method to train the logic of comparing age in the CNN model. Deep metric learning reduces the complex classification task to the nearest neighbor problem [10]. In addition, this approach has the advantage that it makes use of relationships using more data.

A Siamese network [12] is widely used as a deep metric learning-based approach. Two input images are applied to two CNN models, then each input image is mapped to a point in multi-dimensional space, where the similarity of the two input images is described as the corresponding distance. These CNN models are trained using the loss function, by which the points are closely clustered in the case of higher similarity. The well-trained Siamese network generates well-clustered data for the training images. The input image can be accurately labeled by selecting the nearest clustered data compared to the features extracted from input images. Here, the nearest neighbor selection process corresponds with our approach of estimating the labels by comparing the input images with the training images.

However, Siamese network-based deep metric learning has the drawback of difficulty in converging the results. When this learning method is applied for age estimation, all remaining classes except the correct class are negative so divergence often occurs in the learning process. Related to this issue, CRCNN trains a Siamese network using loss function to determine whether the age is younger using two images instead of comparing the similarities. Additionally, CRCNN proposes a selection approach for specific images compared with the input images. This could avoid the side effect of continuously learning with negative reference images.

1.2. Contribution

With these motivations, by applying a Siamese network-based deep metric learning for exact age estimation, we propose a method to converge the process learning a Siamese network. Our proposed approach allows a certain level of error tolerance to increase the ratio of positive data, so that it can perform comparisons for all images in the database, still decreasing the possibility of divergence in the training process.

Additionally, the deep metric learning method trains the CNN model to measure similarity based only on age data, but we found that the accumulated gender data can be additionally used to compare the age. From this experimental fact, we adopted a multi-task learning approach to consider the gender data for more accurate age estimation. Multi-task learning is a method to train CNN models simultaneously with multiple tasks to effectively assist in the training. This method enables the CNN models to be trained to simultaneously perform the age estimation tasks and separate tasks to classify

the gender, so that more relationship data can be involved, which is more helpful to increase the performance in terms of accuracy.

The whole process is as follows. We use Inception V3 for CNN model [13], which is pre-trained with ImageNet [14], and perform the feature-embedding by considering the value of the fully connected layer. The loss function is designed to train our architecture to decrease the distance between feature vectors when two images in batch are in the same class, as well as to increase the distance between feature vectors in the case of differences in class for two images. In this step, we allow a certain level of error tolerance for determining whether two images are in the same class. We define the two feature vectors for measuring age similarity and for measuring gender similarity, respectively. Two feature vectors are simultaneously trained to perform the multi-task learning method.

After training step with these conditions, the feature vectors for all training databases are extracted and the distribution of the clustered data with respect to age similarity can be obtained.

In the test step, the featured vector of an input image is selected with the nearest one in the feature space to compare the relative location in the clustered data distribution.

This paper is organized as follows. Section 2 explains in detail our architecture to perform the learning for age estimation. Section 3 shows the experimental results using the proposed approach, and discusses the performance of the proposed models. Section 4 provides the conclusion of this study.

2. Proposed Architecture

The structure of the neural network in our proposed architecture is described in Figure 1, which is a Siamese network [12]. As shown in Figure 1, the structure and weights in these two networks are completely equivalent. The outputs of two CNN model for input images A and B are used in loss function and the relationship is determined according to the design of loss function. These two networks are used to apply the loss function for the inference as a result of two input images A and B.

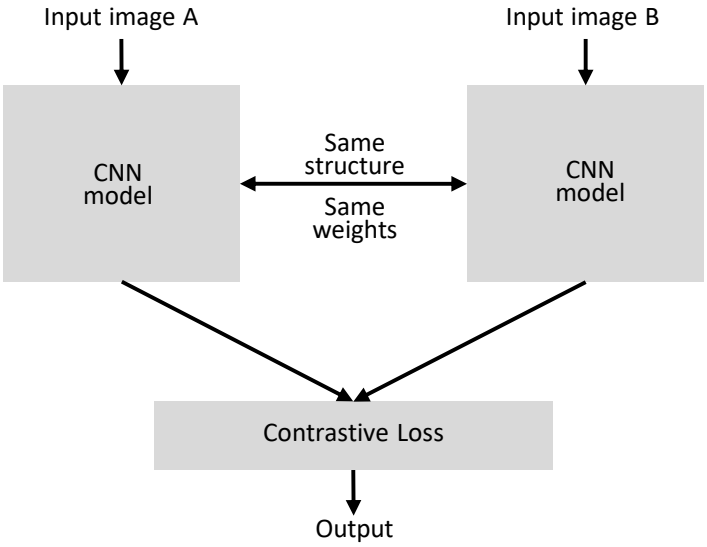

Figure 1. Structure of the Siamese network.

In this paper, instead of using two Siamese network-based CNN models for age comparison from two input images, we apply the contrastive loss function using the inference results for the corresponding images by selecting two images from a batch of training models in a single network.

Figure 2 shows an illustration of the overall algorithm. Inception V3 is used for the construction of the CNN model, but with a fully connected layer, not using a softmax layer. To apply the multi-task learning to estimate age and gender simultaneously, one more fully connected layer is constructed. The first fully connected layer performs age comparison and the second fully connected layer assists age comparison by performing the gender comparison task.

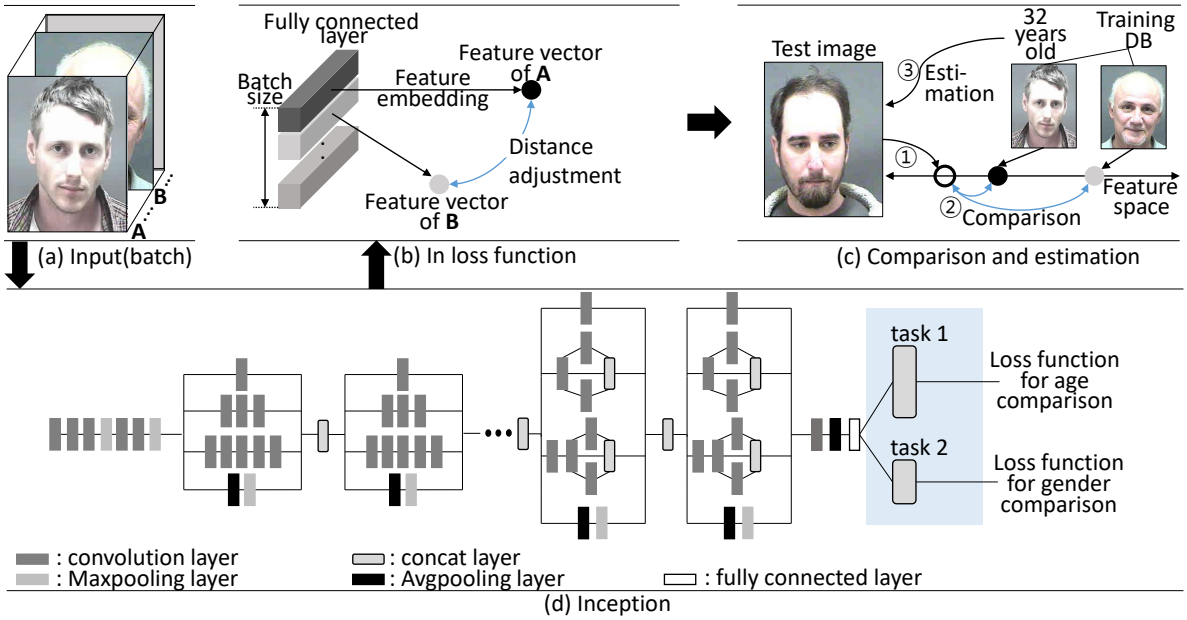

Figure 2. Illustration of the proposed algorithm.

As shown in Figure 2a, two input images are selected from the batch, by considering all selectable combinations. The selected A and B images are mapped to the feature vector which is a final output of the fully connected layers in Figure 2d. In the proposed loss function, the gradient value is propagated into the network to decrease the distance between feature vectors when two images in batch are in same class, as well as to increase the distance between feature vectors in case of different in class for two images, as shown in Figure 2b. Our architecture is trained using the proposed algorithm to determine the similarity between two input images.

In test step, feature vector of test image is compared to feature vectors of all the training database to perform age estimation by selecting the most similar age class as shown in Figure 2c. The detailed process of the proposed algorithm is as follows.

2.1. Inception V3

The proposed algorithm in this paper adopted Inception V3 [13], which is an enhanced version with batch normalization and filter size reduction.

Figure 3 compares module of the Inception model and module of the Inception V3 model. In the Inception model, the filter sizes are 5×5 and 1×1, but the Inception V3 model uses 1×1 and $N \times 1$ filters continuously; as a result, the calculation cost and the number of parameter coefficients are reduced. In this paper, we adopt the Inception V3 model and configure the $(N = 3) \times 1$ filter due to the benefit of the Inception V3 model. To perform Siamese network-based deep metric learning using this Inception V3 model, the final output of fully connected layers is used as the feature vector instead of using the softmax layer, as shown in Figure 2b.

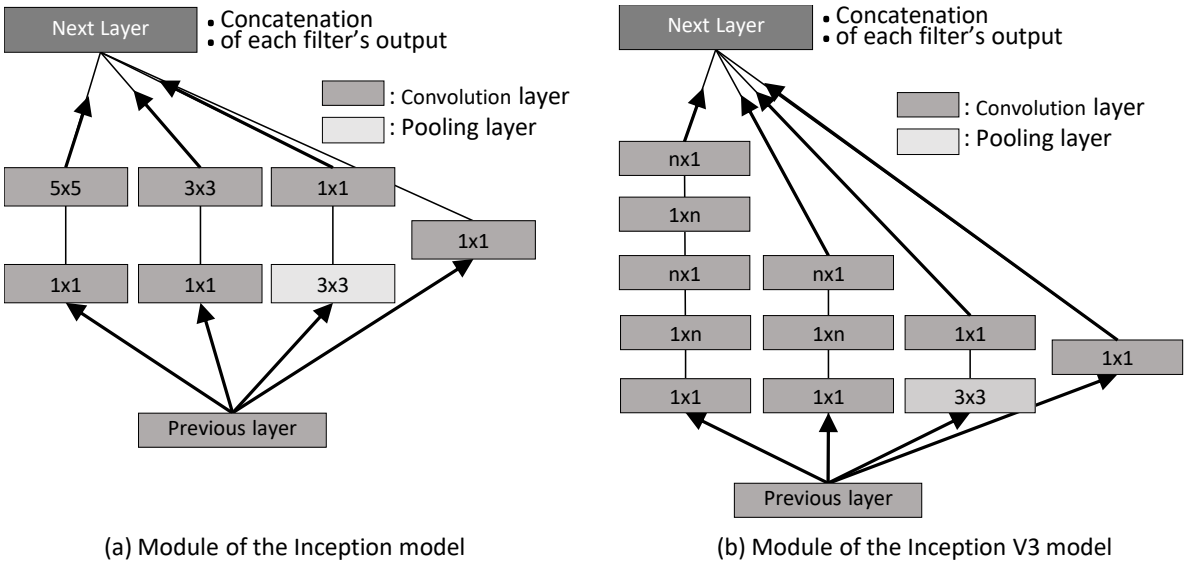

Figure 3. Inception module.

2.2. Selection of Two Images and the Feature-Embedding Process

To implement the Siamese network using a single network, two images are selected from the batch, as shown in Figure 4, and they are used to measure the similarity. The comparison repeats the number of available combinations by selecting two images from the batch. This approach performs the comparisons and trains the model between all images in the batch instead of selecting only specific images, as in the previous CRCNN [11].

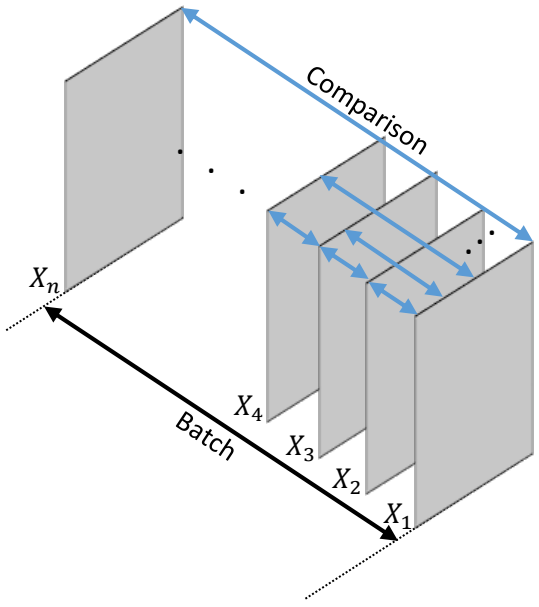

Figure 4. Image selection for comparison in batch.

The two selected images X_i, X_j in the batch, as shown in Figure 5, are mapped and shrunk to the final fully connected layer in N_a dimensions, which is described using Inception V3. The shrunk data are represented with the corresponding features $FV(X_i), FV(X_j)$, in which integer i, j are an index in the batch.

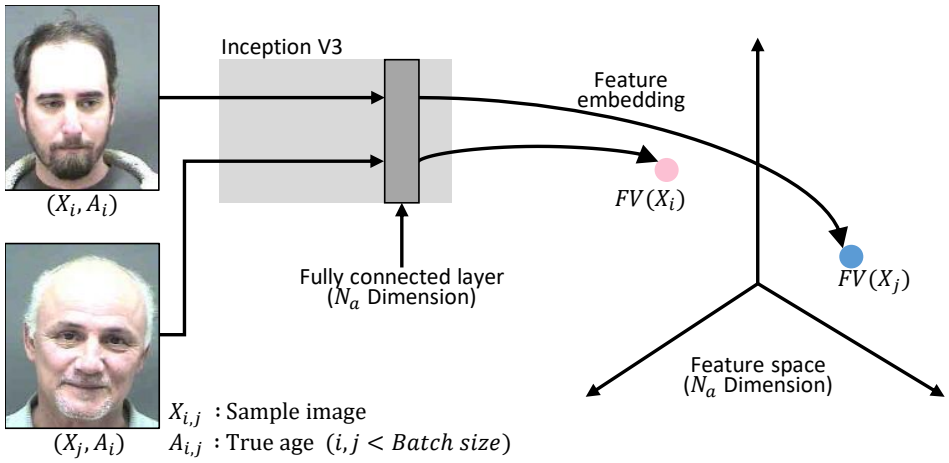

Figure 5. Feature-embedding.

2.3. Distance as Similarity between Two Images

The feature vectors are extracted by the Inception V3-based feature-embedding method, as shown in Figure 5.

The proposed algorithm aims to effectively train the model by mapping the feature vectors into feature space so that similar images are clustered with smaller distance. Therefore, similarity between two images and distance between feature vectors of two images are reciprocal proportion relationship. The distance between the feature vectors is calculated using L^1-norm which calculates the absolute distance of the corresponding value in each dimension with the following equation in terms of the distance D between two feature vectors.

$$D = ||FV(X_i) - FV(X_j)||_{L^1} \tag{1}$$

Some previous approaches [11,13] use the Euclidean distance calculation method called norm2, but the preferred approach in previous studies has been to use norm1 instead of norm2 for Siamese networks [12].

In this paper, we define the distance using L^1-norm and we are able to successfully converge the training result, which was evaluated in the experiment.

2.4. Loss Function for the Training Comparison Task

The feature vector comparison as a representative descriptor for a given image is equivalent to comparing the image itself. Our proposed approach defines the loss function and trains the comparison task of the CNN model so that the extracted features are relatively positioned in the feature space in terms of the similarity of two feature vectors.

The loss function used in this paper is described as follows. The loss function corresponds to the contractive loss function in the Siamese network, which was introduced as a contrastive loss function [12].

$$loss = (1 - \overline{Z})L^-(D) + (\overline{Z})L^+(D) \tag{2}$$

$\overline{Z}$ is a Boolean function that outputs 1 in the case of two similar images; otherwise, it outputs 0. L^- has to satisfy the condition in the manner of a decreasing function, and L^+ of an increasing function, as shown in the following equation.

$$\overline{Z} = \begin{cases} 1, \text{if two images are considered as same class} \\ 0, \text{otherwise} \end{cases} \tag{3}$$

$$L^{-}(x) = 2 \times Qe^{-\frac{2.77}{Q}x}, L^{+}(x) = \frac{2}{Q} \times x^{2} \quad (4)$$

Q is a constant to determine the upper limit of dissimilarity, which is 100 in this paper. Figure 6 is a graph to describe the loss function in terms of the distance between feature vectors. $\overline{Z}$ is 1 in the case of two similar images in the same class, and the L^{+} term remains. The gradient is propagated into the network so that the distance is reduced to minimize the loss in the designed loss function. $\overline{Z}$ is 0 in the case of two images that are considered to be in different classes, and the L^{-} term remains. The gradient is propagated into the network so that the distance is increased for the decreased loss function. With these operations in the network, the weights for feature vector extraction is updated.

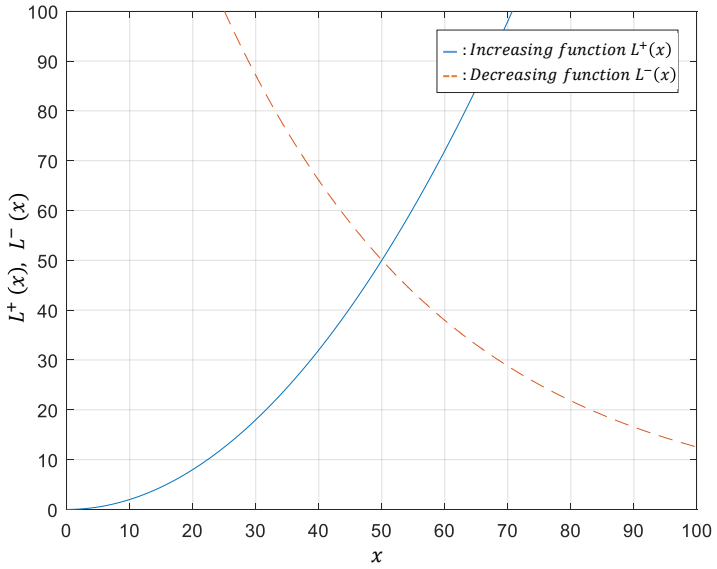

Figure 6. A designed loss function for the proposed algorithm.

Because this designed loss function is used to train the network to determine the distance between feature vectors, there is no inefficiency limiting the basis of the mapping plane. However, unlike the trained database, the proposed method has to search and determine a nearest neighbor from feature vectors. In addition, an approach using this loss function enables the multi-class classification for age estimation of various bands to be simplified as a binary classification problem which only measures the similarity. It mitigates imbalance of the accuracy over all classes, which is caused by the biased training database. However, if this loss function is applied to the binary classifier as it is, the images in the same age class are considered positive, and all other classes are negative; as a result, the trained database becomes imbalanced due to the large number of classes. That is why the Siamese network does not easily converge the training results.

To resolve this issue, CRCNN adopts a technique to select the comparison images in advance to prevent the network from being continuously trained with the negative database. Instead of comparing the similarities in age, it redesigns the loss function to only determine whether the age is younger or older; as a result, it could converge the training results of the Siamese network.

Our approach could succeed in converging the training result by adopting a method to increase the ratio of the positive data, for which the Boolean function $\overline{Z}$ determining age class allows for error tolerance. For example, if three years is allowed as a margin, the loss function considers classes between $N-3$ and $N+3$ years old to be the same class. The proposed technique is helpful to increase the ratio of positive data, so the entire process of training the CNN model is not negatively influenced by the error tolerance.

In fact, our approach loses discrimination by class in the CNN model with the margin-allowed error, but it results in more accurate age estimation by enabling all comparisons for all age ranges. Even though a specific feature vector is involved with the class within a certain range of marginal error tolerance, clustering can be processed further with accuracy of the margin value, by comparing with the feature vector within (margin+1) and −(margin+1) compared to the currently clustered age. The entire clustering procedure using the proposed approach is described in Figure 7.

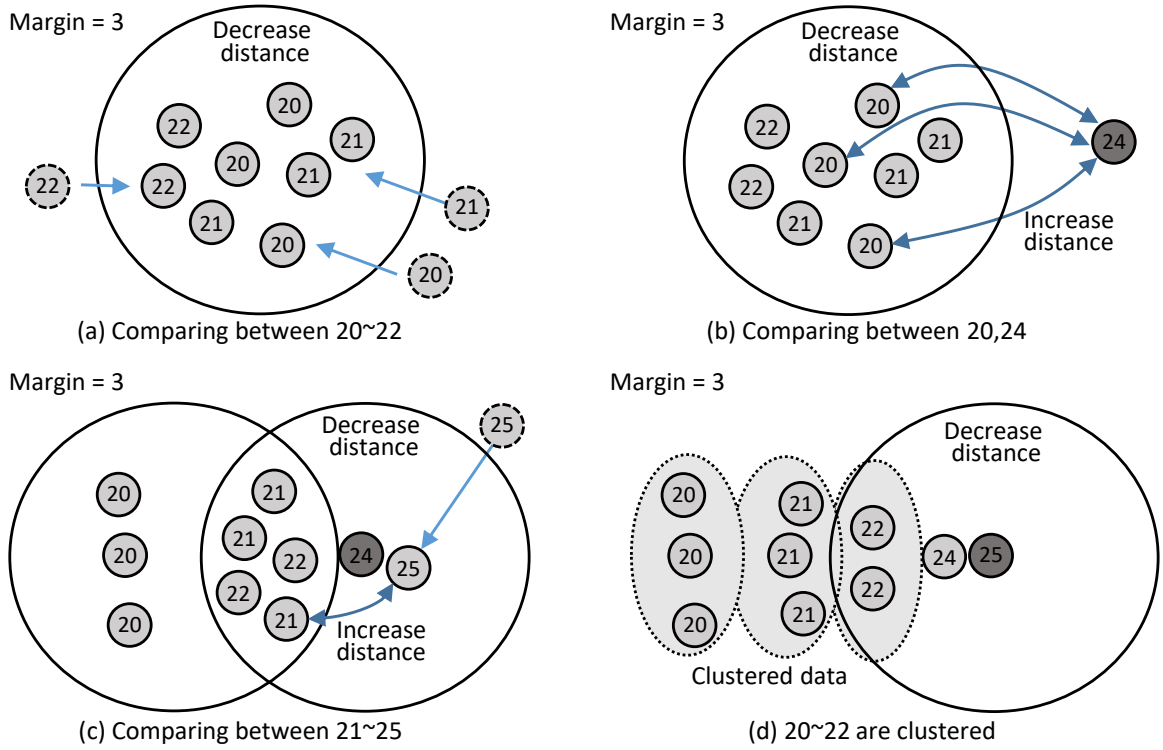

Figure 7. Clustering process allowing marginal error tolerance.

Figure 7 assumes that the margin is defined as 3; the feature vectors of the images are compared and clustered using the proposed loss function. For example, as shown in Figure 7a, if only the feature vectors of the images that are 20–22 years old are compared, then all images are considered similar because the margin is defined as 3, so only the distance decreases, but the clustering does not proceed further. This means that the estimation accuracy is three years. However, as shown in Figure 7b, if the feature vector for an image classified as 24 years old is compared to one classified as 20 years old, the network is trained to increase the distance, so the feature vector of age 24 is clustered to be positioned far away. As shown in Figure 7c, the network is trained so that the feature vector for 21–22 age is clustered to be closely positioned, because ages 20–21 and 24 are within the margin, which can be considered the same class. When a feature vector with 25 is compared, 22 and 25 are considered the same class through the same process, so the network is trained to have a close distance between 22 and 25. As a result, the feature vectors of 20, 21, 22, 24, and 25 are separately clustered, so we can distinguish the age of the images with an accuracy of one year.

2.5. Age Estimation

In the test step using the database trained by the proposed approach, the age estimation process initially involves calculating the feature vectors in N_a dimensions to search for similar images compared to the trained database. Because the CNN model has already been trained to determine the age similarity, the test model comparing the input image is prepared with the clustered feature vectors. The feature vector for the input image is extracted using the same CNN model, then compared with

the clustered data in the test model. The test process involves age estimation performed by calculating mean age of among M_{th} nearest neighborhoods. The distance-based nearest neighborhood search method is also based on L^1-norm which is used in the training process. The entire test process is described in Figure 8.

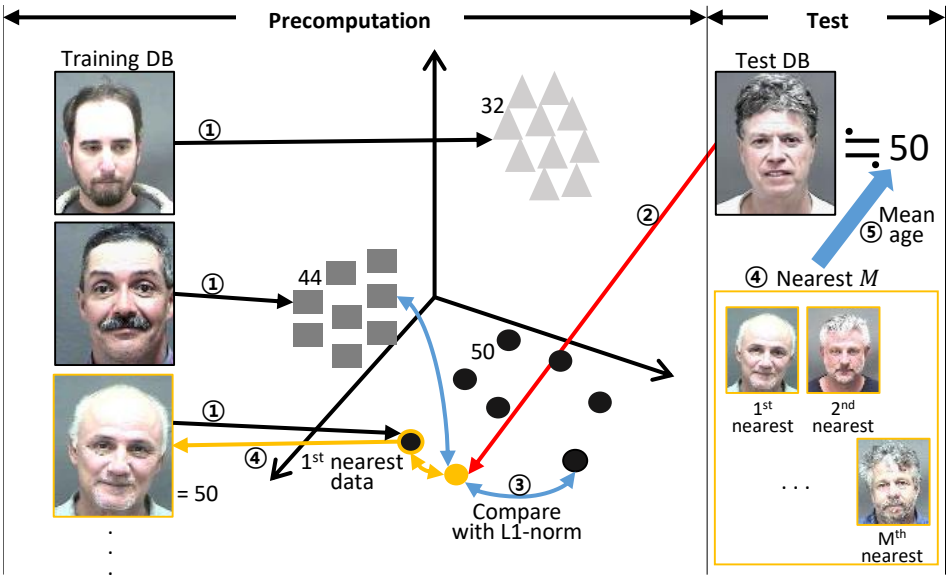

Figure 8. Age estimation process selecting the nearest neighborhood in the feature space.

2.6. Multi-Task Learning for Age and Gender Estimation

The loss function for the proposed method is designed to train the CNN model with age similarity as the relation of classes. Even though the CNN model is trained to determine a similar level using the age data, it can be further trained by clustering the classes closely with similar images using detailed conditions, such as face angle, hair length, and beard. An algorithm that determines age using various conditions, in addition to the absolute age data, is more appropriate. That is why the detailed conditions are automatically configured and applied to the training model by only defining the age-based similarity.

With this concept, we first tried the gender classification by using the model trained by only the age data, and then we measured the accuracy of the gender-matching result. We found that our approach using only age-based estimation was able to classify the gender with 81.23% accuracy compared to the result of gender-based classification. The result is summarized in Table 2. The result gave us the following two insights. First, our approach internally uses gender-based conditions to perform the age estimation. Second, the gender data can be an important clue to estimate age. In fact, the 81.23% accuracy of gender classification based on age data means that the age estimation is tightly coupled with gender.

Based on this speculation, our approach adopts the multi-task learning approach so that it additionally provides gender data to the trained model when comparing age. The multi-task learning simultaneously trains the model to increase the performance in terms of accuracy of age estimation. If the individual tasks have a cross-coupled relationship, the multi-task learning approach enables the model to be trained by selecting commonly important variables in the multiple tasks. Utilizing the capability to train the model considering the relationships between tasks, we were able to assist in the age estimation with gender data, thus training the model to consider age and gender simultaneously.

The multi-task learning technique, which is applied in this paper, is described in Figure 9. A fully connected layer in N_g dimensions is added for the gender comparison used to compare age in Inception V3. We also designed a loss function to train the logic of the gender comparison so that the weights in the layer are updated in a similar way as in the age comparison algorithm. The margin of

comparison in the loss function is 0, and it divides the positive and negative data on the basis of gender. This additional task for gender comparison is temporarily used to assist the data in training the age estimation logic. The finally calculated loss is the sum of the loss by the age estimation and gender classification.

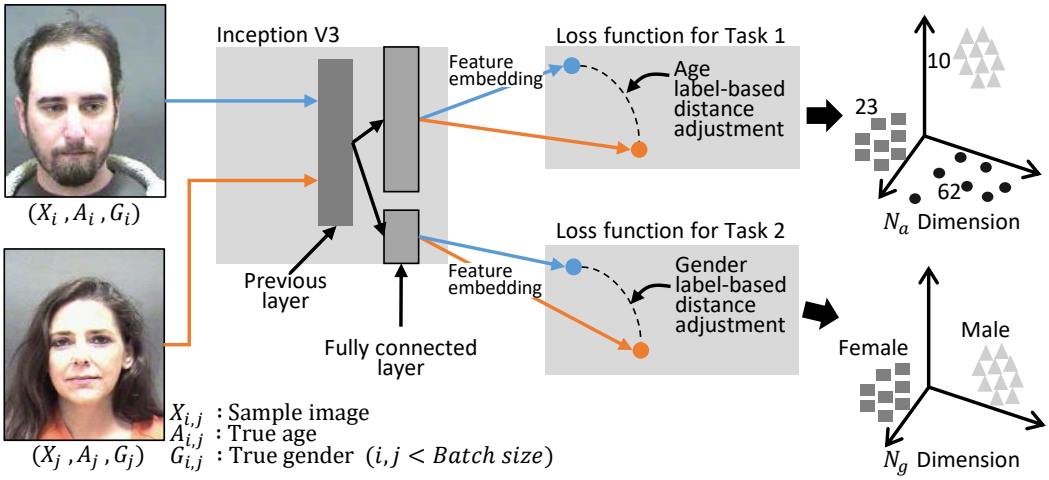

Figure 9. Multi-task learning for the age algorithm considering age and gender simultaneously.

3. Experimental Results and Discussion

The purpose of this experiment was to verify the age estimation accuracy performance based on our architecture in an open image database. We implemented our algorithm using TensorFlow [15], an open source deep learning framework based on Python. We use Inception V3 for CNN model [13] which is pre-trained with ImageNet [14]. The batch size was 128, the image size was 227 and dropout was performed with a probability of 50%. The first fully connected layer's dimension, N_a, which is task with measuring age similarity was 70, and N_g, which is the dimension of second fully connected layer for measuring gender similarity, was 10. Each dimension was experimentally selected. In the gradient descent procedure to optimize network weights, the Adadelta [16] method was used. The margin-allowed error newly defined in our proposed method was set to 4. This means that, if the difference between age is smaller than 4, two ages are considered to be in the same class. In the test step, mean age of the nearest 20 ($M = 20$) was calculated for prediction. The age estimation performance was evaluated by mean absolute error (MAE) generally used in previous research as defined in the following equation. The meaning of MAE is how close a prediction is to the true age.

$$MAE = \frac{\sum_{i=1}^{n} |A_i - \tilde{A}_i|}{n} \tag{5}$$

$\tilde{A}_i$ and A_i are the estimate and true age of the sample image j, and n is the total number of samples. We also calculated the cumulative score (CS) [17–19]. CS indicates the percentage of sample correctly estimated in the range of $[A_i - T, A_i + T]$, a neighbor range of the true age where T is the parameter representing the tolerance. CS was calculated using the following equation.

$$CS(T) = 100 \times \frac{\sum_{i=1}^{n} [|A_i - \tilde{A}_i| <= T]}{n} \tag{6}$$

Here, **[.]** is the truth-test operator. A higher value of $CS(T)$ means a better performance of the architecture. We experimented with two public datasets. The first was the MORPH database [20]. There are 55,132 face images from more than 13,000 subjects in this database. The ages of the face images range from 16 to 77. Front face images are from different races, among which African faces account for about 77%, European faces account for about 19% and the remaining 4% include Hispanic,

Asian, Indian, and other races [11]. The second was MegaAge-Asian [21]. It contains 40,000 face images of Asians with ages from 0 to 70. Table 1 shows the size of each dataset and the corresponding splits for training and testing. At first, we selected test images randomly and remainders were used as training images. Therefore, there is no intersection between training and test sets.

Table 1. The proposed method was evaluated using two datasets.

DB Name	The Number of Training Images	The Number of Test Images
MegaAge Asian	40,000	4000
MORPH	45,132	10,000

The proposed architecture was trained with each dataset in Table 1. We experimented to verify the performance of our method, as described in the following sections.

3.1. Toy Example: Visualization of Feature Embedding Computed by Our Method Using a Subset of the MORPH Dataset

To verify the fact that the clustering process improves the accuracy of the margin value, feature vectors were visualized using a small subset of the MORPH dataset. For visualization on two-dimensional space, to facilitate convergence, we collected face images with ages from 16 to 63 (only 48 classes) and each class had 1–3 images randomly. Hyper parameters for the toy example are as follows. The batch size was 48, the dimension of the feature vector was 2 for visualization on two-dimensional space. In the case of the toy example, the margin value was set as 2. After the training step with these conditions, extracted feature vectors were clustered, as shown in Figure 10. The vertical axis is the true age of each feature vector; the others are axes of feature space. Most of the feature vectors were well-clustered, as shown in the zoomed graph (red box). The clustering process had an accuracy of one year but our CNN model had an accuracy of two years, thus putting those that were two years younger or two years older in the same class.

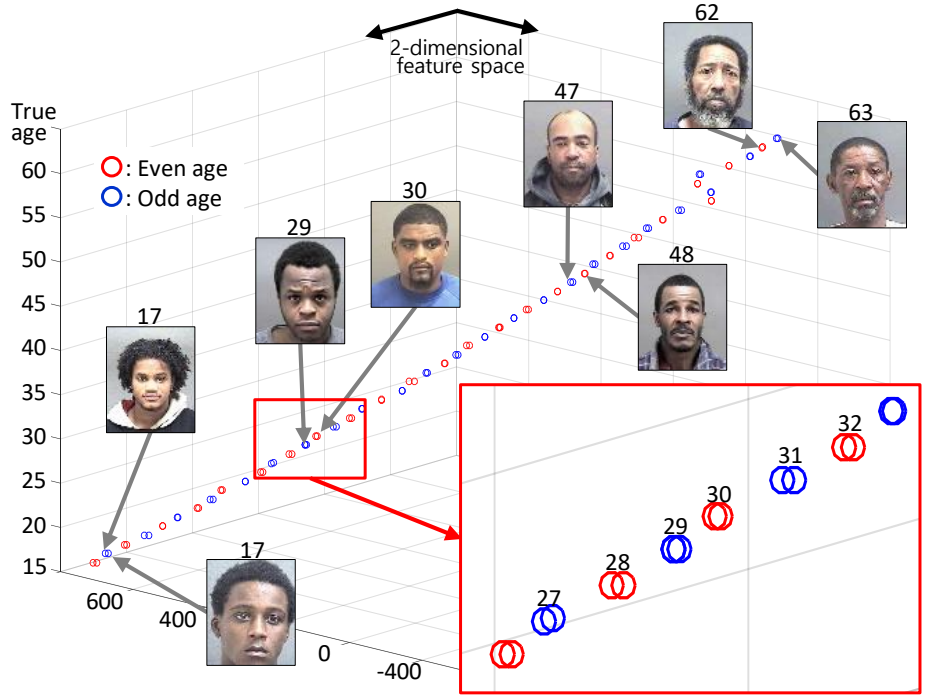

Figure 10. Visualization of feature embedding with the toy example.

3.2. Multi-Task Learning for Age and Gender Estimation

The first row of Table 2 is the result of the gender classification rate on the MORPH dataset using only age data. Even though gender data were not used, the gender classification rate was quite high. The classification rate was much lower than in another CNN model using gender data. Even AlexNet, which is a relatively simple model, had a better classification rate. However, 80% accuracy means that the age estimation is tightly coupled with the gender. Therefore, we tried to use gender data in the CNN model by applying multi-task learning for age and gender estimation simultaneously. The results of the experiment before and after applying the multi-task learning method are shown in Figure 11. The MAE of our method with multi-task learning slightly decreased from 2.24 to 2.28, but $CS(T)$ values were improved. In particular, the $CS(5)$ value increased by about 2%. Therefore, performance was improved by using gender data to estimate age through multi-task learning.

Table 2. Gender classification rates on the MORPH dataset.

Method	Accuracy (%)
Our w/o gender data	81.23
Alexnet [5] w gender data	97.38
Inception V3 [7] w gender data	99.1

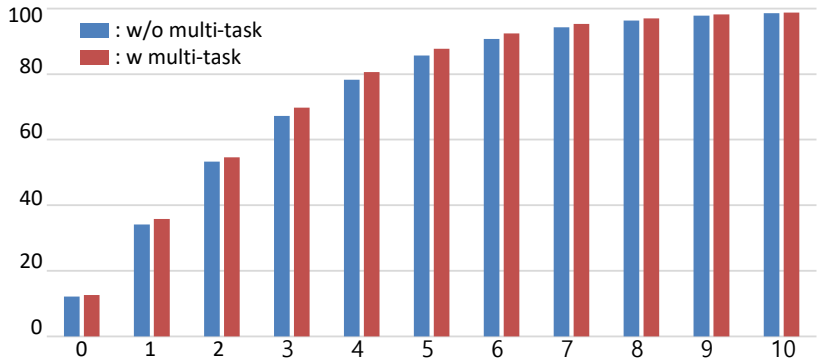

Figure 11. Comparison of our method with and without multi-task learning.

3.3. Comparison with Deep Metric Learning-Based Approaches on the MORPH Dataset

Table 3 shows the age estimation result of the dataset and a comparison with traditional methods based on deep metric learning. The MAE of our method was 2.24, indicating better accuracy than the MAE of the CRCNN [11], which is 3.74. This means that applying our method of deep metric learning based on a Siamese network is suitable for age estimation. Moreover, M-LSDML [22], the latest age estimation method based on deep metric learning, has a slightly lower MAE than our method. Additionally, the MAE of the ResNet with each loss function for deep metric learning is shown.

Table 3. Age estimation results on test images of the dataset and a comparison with traditional deep metric learning methods.

Method	Kinds of Loss Function	MORPH(MAE)
Our	Revised contrastive loss function	2.24
Our w multi-task learning	Revised contrastive loss function	2.28
CRCNN [11]	Contrastive loss function	3.74
M-LSDML [22]	Custom-defined loss function	2.89
ResNet (contrastive loss) [22]	Contrastive loss function	3.72
ResNet (triplet hinge loss) [22]	Triplet hinge loss function	3.59
ResNet (lifted structural loss) [22]	Lifted structural loss function	3.24

3.4. Comparison with State-of-Art Method on Each Datasets

In addition, we compared the state-of-the-art methods. Most techniques using the MegaAge-Asian dataset evaluate age estimation performance by $CS(T)$, as shown in Table 4. Our method achieved a slightly higher score than the other method on the MegaAge-Asian dataset. In the case of techniques using the MORPH dataset, the MAE is widely used to evaluate the age estimation performance. In Table 5, the MAE of each technique is shown. In the experiment on MORPH dataset, our method achieved the best MAE 2.24.

Table 4. Comparison of $CS(T)$ with state-of-the-art methods on the MegaAge Asian dataset (* face alignment method is applied, ** additional labels are used).

Method	$CS(3)$	$CS(5)$
Our	69.70	84.64
MobileNet [23]	44.0	60.6
DenseNet [24]	51.7	69.4
Zhang et al. [25] **	64.08	82.43
SSR-Net [26] *	54.9	74.1

Table 5. Comparison of MAE with state-of-the-art methods on the MORPH dataset (* face alignment method is applied, ** additional labels are used).

Method	MAE
Our	2.24
Our w multi-task **	2.28
Ranking-CNN [9]	2.96
DEX [4] *	3.25
DEX w IMDB [4] *	2.68
Zhang et al. [25] **	2.87
Zhang et al. w IMDB-WIKI [25] **	2.52
SSR-Net [26] *	2.52

In terms of age estimation, the accuracy of our method is improved with respect to CS value and MAE by using more data from relationship between images. However, to deal with bigger datasets, comparing all images may not be an efficient strategy because of the increased computation and clustered data. Because our architecture has disadvantage in terms of the training time, in the case of applying our multi-task method in MORPH datasets, our architecture needs 275 epochs to converge. In future work, to reduce the training time, we will consider a strategy of automatically selecting images which can be references to compare with training dataset and using for gallery. This strategy can be more appropriate to apply for bigger and more varied datasets (e.g., FG-net and IMDB-WIKI). Additionally, for optimizing our method, more analysis on dimension of feature vector, the consideration of simpler networks with statistical significance according to random initialization and more efficient loss function are needed, which will be researched in future work.

4. Conclusions

This paper is motivated by the fact that training a CNN model based on age comparison is easier than directly estimating the absolute age. The proposed approach trained the CNN model for age comparison using a Siamese network-based deep metric learning method. We designed a binary classifier, which was applied to train the Siamese network, to cluster the classes within the margin of tolerance as the same class so that we could successfully train the Siamese network by adopting L^1-norm instead of using L^2-norm. The experimental test indicated that the proposed approach itself performs the gender classification in processing the age estimation, so we tried to train the CNN model by comparing age and gender simultaneously using the multi-task learning technique.

The proposed method was evaluated using the MORPH dataset. Although our architecture needs a large number of epoch, it results in better performance and additional enhancement using multi-task learning for age and gender, which is compared to that of the CRCNN of original Siamese network-based deep metric learning and the latest M-LSDML. Additionally, our method is also the best results compared with the results of the state of art on MegaAge Asian and MORPH datasets. In future work, the more analysis is needed to reduce the training time by selecting reference images to compare rather than comparing all images.

Author Contributions: Y.J. designed the entire core architecture and performed the hardware/software implementation and experiments; S.L. validated the experimental results by the proposed framework; K.H.P. proposed the key concept and algorithm of the proposed architecture; and D.P. was the corresponding author.

Funding: This study was supported by the BK21 Plus project funded by the Ministry of Education, Korea (21A20131600011).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
DEX	Deep EXpectation
CRCNN	Comparative Region Convolution Neural Network
MAE	Mean Absolute Error
CS	Cumulative Score

References

1. Ling, H.; Soatto, S.; Ramanathan, N.; Jacobs, D.W. A Study of Face Recognition as People Age. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
2. Alkass, K.; Buchholz, B.A.; Ohtani, S.; Yamamoto, T.; Druid, H.; Spalding, K.L. Age estimation in forensic sciences: Application of combined aspartic acid racemization and radiocarbon analysis. *Mol. Cell. Proteom.* **2010**, *95*, 1022–1030. [CrossRef] [PubMed]
3. Han, H.; Otto, C.; Jain, A.K. Age estimation from face images: Human vs. machine performance. In Proceedings of the 2013 International Conference on Biometrics (ICB), Madrid, Spain, 4–7 June 2013; pp. 1–8.
4. Rothe, R.; Timofte, R.; Van Gool, L. Deep Expectation of Real and Apparent Age from a Single Image without Facial Landmarks. *Int. J. Comput. Vis.* **2018**, *126*, 144–157. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
6. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. [1409.1556].
7. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2015**, arXiv:1409.4842. [1409.4842].
8. Yin, X.; Liu, X. Multi-Task Convolutional Neural Network for Face Recognition. *arXiv* **2017**, arXiv:1702.04710. [1702.04710].
9. Chen, S.; Zhang, C.; Dong, M.; Le, J.; Rao, M. Using Ranking-CNN for Age Estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 742–751.
10. Song, H.O.; Xiang, Y.; Jegelka, S.; Savarese, S. Deep Metric Learning via Lifted Structured Feature Embedding. *arXiv* **2015**, arXiv:1511.06452. [1511.06452].

11. Abousaleh, F.S.; Lim, T.; Cheng, W.H.; Yu, N.H.; Hossain, M.A.; Alhamid, M.F. A novel comparative deep learning framework for facial age estimation. *EURASIP J. Image Video Process.* **2016**, *2016*, 47. [CrossRef]
12. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; pp. 539–546.
13. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:1512.00567. [1512.00567].
14. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
15. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A System for Large-Scale Machine Learning. In Proceedings of the OSDI'16 Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
16. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method. *arXiv* **2012**, arXiv:1212.5701. [1212.5701].
17. Zhang, Y.; Yeung, D.Y. Multi-task warped Gaussian process for personalized age estimation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2622–2629.
18. Guo, G.; Mu, G. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 657–664.
19. Guo, G.; Fu, Y.; Dyer, C.R.; Huang, T.S. Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression. *IEEE Trans. Image Process.* **2008**, *17*, 1178–1188. [CrossRef] [PubMed]
20. Ricanek, K.; Tesafaye, T. MORPH: A longitudinal image database of normal adult age-progression. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 341–345.
21. Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; Brossard, E. The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. *arXiv* **2015**, arXiv:1512.00596. [1512.00596].
22. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Label-Sensitive Deep Metric Learning for Facial Age Estimation. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 292–305. [CrossRef]
23. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [1704.04861].
24. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993. [1608.06993].
25. Zhang, Y.; Liu, L.; Li, C.; Loy, C.C. Quantifying Facial Age by Posterior of Age Comparisons. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017.
26. 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 Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, Stockholm, Sweden, 13–19 July 2018; pp. 1078–1084.