applied sciences

MDPI

Article

# Multiple Network Fusion with Low-Rank Representation for Image-Based Age Estimation

**Chaoqun Hong *,†, Zhiqiang Zeng, Xiaodong Wang and Weiwei Zhuang**

School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; zqzeng@xmut.edu.cn (Z.Z.); xdwangjsj@xmut.ecu.cn (X.W.); zhuangweiwei@xmut.edu.cn (W.Z.)

* Correspondence: cqhong@xmut.edu.cn; Tel.: +86-592-6291390

† Current address: Ligong Road #600, Houxi Town, Jimei District, Xiamen 361024, Fujian Province, China.

Received: 9 August 2018; Accepted: 3 September 2018; Published: 10 September 2018


**Featured Application: The proposed method is used in biometric feature recognition of people.**

**Abstract:** Image-based age estimation is a challenging task since there are ambiguities between the apparent age of face images and the actual ages of people. Therefore, data-driven methods are popular. To improve data utilization and estimation performance, we propose an image-based age estimation method. Theoretically speaking, the key idea of the proposed method is to integrate multi-modal features of face images. In order to achieve it, we propose a multi-modal learning framework, which is called Multiple Network Fusion with Low-Rank Representation (MNF-LRR). In this process, different deep neural network (DNN) structures, such as autoencoders, Convolutional Neural Networks (CNNs), Recursive Neural Networks (RNNs), and so on, can be used to extract semantic information of facial images. The outputs of these neural networks are then represented in a low-rank feature space. In this way, feature fusion is obtained in this space, and robust multi-modal image features can be computed. An experimental evaluation is conducted on two challenging face datasets for image-based age estimation extracted from the Internet Move Database (IMDB) and Wikipedia (WIKI). The results show the effectiveness of the proposed MNF-LRR.

**Keywords:** age estimation; multi-modal features; deep learning; low-rank representation

## 1. Introduction

Image-based age estimation tries to compute the age or age group with facial images. It can be widely used in many applications such as biometric feature recognition, human–computer interaction (HCI), and so on. Although a number of studies have been conducted [1–3], image-based age estimation is still a challenging task due to the following aspects. First, it often lacks sufficient training samples since each person may be captured by several images in a wide range of ages. Second, facial appearance may not indicate the age accurately since some people may look younger than they actually are and some people may look older. Third, facial images are often captured in wild conditions so they are influenced by large variations such as occlusion, lighting, shadow, and complex backgrounds.

Similar to many other applications of computer vision, most existing image-based age estimation approaches focus on two key stages: feature description and feature mapping. Feature description tries to represent facial images without losing details. Traditional methods usually uses texture features or shape features, such as the active appearance model (AAM) [4], holistic subspace features [5,6], local binary patterns (LBPs) [7], Gabor wavelets [4], bio-inspired features (BIFs) [8], and so on. However, most of them make use of hand-crafted features. In this way, strong prior knowledge is required. To solve this problem, learning-based feature descriptors [9,10] have been proposed to compute descriptive features directly from images. Recently, neural networks have

been efficient in exploring descriptive representations in natural images, such as autoencoders [11], Convolutional Neural Networks (CNNs) [12], and so on. Among these methods, Liu et al. proposed group-aware deep feature learning (GA-DFL) to estimate ages with facial images [13]. Different from most previous methods using hand-crafted features for facial image description, GA-DFL uses a deep CNN framework to compute a discriminative feature descriptor per image automatically from raw pixels of facial images. Although a large number of feature descriptors have been proposed, most of them can only describe a part of the information inherent in images. Therefore, researchers look into representing images with multiple features. Traditional methods make use of multiple features by directly concatenating them, which is oversimplified. To solve this problem, researchers also apply manifold learning to combine different types of features [14,15].

On the other hand, feature mapping tries to learn the mapping relationship from face images to age labels. With descriptive representations of facial images, age estimation is usually considered as a regression or classification problem [5,16]. Linear regression and twin Gaussian processes are also used for pose estimation [17,18]. Tian et al. proposed conducting age estimation by taking both ordinality and locality into consideration [19]. Previous approaches made an over-simplified assumption that the mapping from images into poses is linear. To tackle this nonlinear issue, methods based on deep learning have been applied. They can train a series of nonlinear mapping models [20–24]. However, these models cannot explicitly define the ordinal relationship between facial images and chronological ages, because they usually suffer from insufficient and unbalanced training data. In this way, they still cannot be used in practical scenarios.

Although many methods for age estimation with images have been proposed, they usually use only a single type of feature. Even with popular neural networks, they apply only a single structure of neural networks, which still suffers the so-called "semantic gap". Currently, multiple types of features have been used in many applications. Inspired by it, we proposed a Multiple Network Fusion with Low-Rank Representation (MNF-LRR) for the age estimation method. The contributions of this paper can be summarized by the following:

- The first and key contribution is a novel framework that estimates ages with a single image by fusing multiple deep neural networks. This framework is flexible and the hidden representations are computed independently. In this way, different types of neural networks, different network structures, and different features can be used in this framework.
- The second contribution of the proposed method is multiple-network fusion with low-rank learning. Low-rank representation is naturally sparse. Besides, different types of features are extracted by different networks and their distributions can be observed clearly. To improve traditional low-rank learning, we introduce a hypergraph manifold. In this way, samples can be represented in a unified low-rank space and the process of fusion can be achieved in this space.
- The third contribution is that the performance of the proposed method is verified on datasets from the Internet Movie Database (IMDB) and Wikipedia (WIKI). They are challenging datasets since the images are collected in natural scenarios and not all of the faces are frontal. The performance on this dataset indicates that the proposed MNF-LRR is suitable for practical and complicated applications.

## 2. Multiple Network Learning with Low-Rank Representation

### 2.1. Overview of the Proposed Method

The process of the proposed method (MNF-LRR) can be summarized in Figure 1. To get rid of the influences of background, we should extract faces in images first. This process depends on the definitions of different datasets. In some datasets, such as IMDB and WIKI, the positions and sizes are provided and they can be used directly. However, in some other datasets or real scenarios, we need face detection or face tracking to determine the face area. We then utilize different networks to extract deep features of facial images. Finally, we use manifold learning based on low-rank representation

to integrate the outputs of these networks. In this way, a unified multi-modal representation can be obtained.

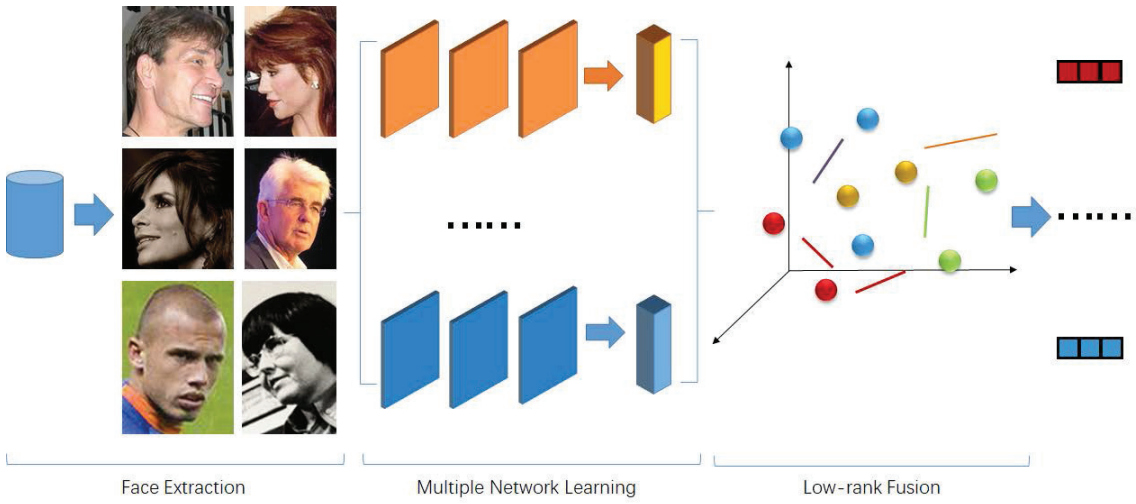



**Figure 1.** The flowchart of the proposed method (Multiple Network Fusion with Low-Rank Representation (MNF-LRR)).

### 2.2. Definitions

In age estimation with regression, given a set of images $X = \{x_1, x_2, ..., x_n\}$ and the corresponding labels $Y = \{y_1, y_2, ..., y_n\}$ with $n$ pairs of samples, we try to learn a model that minimizes the loss:

$$\underset{\delta}{\operatorname{argmin}} |Y - \mathcal{F}(\overline{X})| \tag{1}$$

where $\mathcal{F}$ is the regression function, $\delta$ is the regression parameter, and $\overline{X}$ is the feature representation of $X$. Therefore, to minimize Equation (1), we need a descriptive $\overline{X}$ and a reasonable $\mathcal{F}$. In the proposed method, we focus on $\overline{X}$.

### 2.3. Multiple Network Learning

As mentioned in the introduction, multiple feature fusion has been proved to be effective in image representation. In this way, to compute $\overline{X}$, we propose feature learning by fusing multiple neural networks, which compute features with different neural networks and integrate them to form new features. Neural networks [25,26] have been widely used to explore hidden representations of images and the effectiveness has been proved. Generally speaking, neural networks compute hidden representation by minimizing the loss function:

$$\sum_{i}^{n} \| x_i - \overline{x}_i \|^2 \tag{2}$$

where $\overline{x}_i = Wx_i$ is the hidden representation by mapping $x_i$ with weight $W$. The key to neural networks is optimizing $W$, which is defined differently by different neural networks. However, they depend on a large number of training data. Usually, age estimation with a single image is achieved with insufficient training samples or classification information. Therefore, we adopt different types of neural networks to extract different types of features and fuse them to improve the descriptive power with a small number of training samples. In MNF-LRR, we use the following neural networks to represent face images.

- Autoencoders (AE). Autoencoders are unsupervised to learn the hidden representation. To solve Equation (2), people usually use denoising autoencoders (DAE). In DAE, inputs $x_1, ..., x_n$ are corrupted by randomly removing some features. After corruption, $x_i$ is converted to $\hat{x}_i$ and $W : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is denoted as the transform matrix to reconstruct $x_i$ with $\hat{x}_i$. In this way, the squared reconstruction loss can be defined as

$$\frac{1}{2n} \sum_{i=1}^{n} \parallel x_i - W\hat{x}_i \parallel^2 . \tag{3}$$

The solution to Equation (3) depends on corrupted features of each input. To lower the variance, Marginal Denoising Autoencoders (MDA) [27] utilize multiple epochs with the training set, each epoch with different corruption settings. In this way, the overall squared loss can be transformed to

$$loss(W) = \frac{1}{2mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \parallel x_i - W\hat{x}_{i,j} \parallel^2 \tag{4}$$

where $\hat{x}_{i,j}$ represents the $j$th corrupted features, and $m$ is the number of epochs.

To represent features with the matrix form, $X = [x_1, ..., x_n] \in \mathbb{R}^{d \times n}$ is denoted as the data matrix, while the $m$-epochs repeated version of $X$ is denoted by $\overline{X} = [X, .., , X]$ and the corrupted version of $\overline{X}$ is denoted by $\hat{X}$. Equation (4) can then be reduced to

$$\begin{aligned} loss(W) &= \frac{1}{2mn} tr[(\overline{X} - W\hat{X})^T (\overline{X} - W\hat{X})] \\ &= \frac{1}{2mn} tr[\overline{X}^T \overline{X} - \overline{X}^T W\hat{X} - \hat{X}^T W^T \overline{X} + \hat{X}^T W^T W\hat{X}] \end{aligned} . \tag{5}$$

We can clearly figure out that Equation (5) is a convex problem, and the global optimal solution to it can be computed by setting its partial derivative for $W$ to 0. We then need to compute partial derivative of $loss(W)$, which is defined as

$$\arg\min_{W} loss(W). \tag{6}$$

$\frac{\partial loss(W)}{\partial W} = 0$ is set, and the close form to compute optimal $W$ is

$$W = \overline{X}\hat{X}^T (\hat{X}\hat{X}^T)^{-1}. \tag{7}$$

- Convolutional Neural Networks (CNNs). CNNs are constructed by alternatively stacking convolutional layers and spatial pooling layers. Convolutional layers are key to CNNs since they generate feature maps by linear convolutional filters. The feature maps are then activated by nonlinear functions, which are called activation functions. Different activation functions are defined, such as rectifier, sigmoid, tanh, and so on. Taking the Rectified Linear Units (ReLUs) as an example, the feature maps can be computed by

$$F(X; W) = \mathcal{R}(x_i) = max(0, wx_i) = \begin{cases} 0 \, , \, x_i < 0 \\ w_i x_i \, , \, x_i \geq 0 \end{cases} . \tag{8}$$

In computational networks, given an input or set of inputs, the activation function of a neuron defines the output of that neuron with these inputs. In the scenario of the deep neural network,

activation functions project $x_i^v$ to a higher level hidden representation step by step with a sequence of non-linear mappings, which can be defined as

$$(x_i)^0 \xrightarrow[W]{\mathcal{R}} (x_i)^1 \xrightarrow[W]{\mathcal{R}} \dots \xrightarrow[W]{\mathcal{R}} (x_i)^l \tag{9}$$

where $l$ is the number of layers, and $\mathcal{R}$ is the mapping function from input to estimated output.

To optimize the weighted matrix $W$, which contains the mapping parameters, we use a back-propagation strategy. For each echo of this process, the weighted matrix is updated by $\Delta W$, which is defined by

$$\Delta W = -\eta \frac{\partial E}{\partial W}. \tag{10}$$

$\eta$ is the learning rate, and we can define

$$\frac{\partial E}{\partial W} = (y_i - \mathcal{R}(x_i))(x_i)^T. \tag{11}$$

In this way, we try to train a model that minimizes the differences between the groundtruth $y_i$ and the estimated output $\mathcal{R}(x_i)$. The back-propagation strategy can be modeled by

$$(x_i)^0 \xleftarrow[W]{\mathcal{R}} (x_i)^1 \xleftarrow[W]{\mathcal{R}} \dots \xleftarrow[W]{\mathcal{R}} (x_i)^l. \tag{12}$$

- Recursive Neural Networks (RNN). RNNs process a structured input with the same set of weights recursively. In this way, we can traverse the given structure into topological order and obtain a structured output or a scalar prediction on it. Different from CNNs, nodes in RNNs are integrated into parents with a weight matrix. This matrix is shared across the whole network. Besides, a non-linearity such as activation functions mentioned above is used. Taking tanh as an example, if $x_i$ and $x_j$ are $n$-dimensional features of nodes, their parent must be an $n$-dimensional feature, too. It can be computed by

$$p_{i,j} = tanh(W[x_i; x_j]) \tag{13}$$

where $W$ is a learned $n \times 2n$ weight matrix, which is usually computed with Stochastic Gradient Descent (SGD). The gradients are calculated using back-propagation through structure (BPTS). BPTS is a variant of the aforementioned back-propagation through time for RNNs.

In the proposed MNF-LRR, by combining Equations (7), (11), and (13), Equation (2) can be rewritten as

$$\alpha \sum_i^n \parallel x_i - W_{ae} x_i \parallel^2 + \beta \sum_i^n \parallel x_i - W_{cnn} x_i \parallel^2 + \gamma \sum_i^n \parallel x_i - W_{rnn} x_i \parallel^2 \tag{14}$$

where $W_{ae}$, $W_{cnn}$, and $W_{rnn}$ are weighting parameters learned by autoencoders, CNNs, and RNNs, respectively. $\alpha$, $\beta$, and $\gamma$ are switches to turn on or off the corresponding neural networks. In this way, we can compute multi-modal feature representation.

### 2.4. Fusion with Low-Rank Representation

As mentioned before, multi-modal feature fusion by computing semantic relationship is more reasonable than simple concatenation. The key to learning the semantic relationship is how to define and compute affinities among data. Many existing methods can be used, such as subspace learning and manifold learning. Recently, low-rank learning attracts plenty of attention. In low-rank representation, assume the data is clean and is drawn from independent subspaces, then there exists $Q^*$, which is block-diagonal, and the rank of each block equals the dimension of the corresponding subspace. Given

the $i$-th modal $X^{(i)}$, computed in the previous sub-section, we can compute the affinities among feature vectors by solving the minimization problem:

$$\begin{aligned} &\min_{Q_0,E_0} \parallel Q_0 \parallel_* + \lambda \parallel E_0 \parallel_{2,1} \\ &\text{s. t. } X^{(i)} = X^{(i)}Q_0 + E_0 \end{aligned} \tag{15}$$

where $\parallel \bullet \parallel_*$ denotes the trace norm, and $\parallel \bullet \parallel_{2,1}$ is the $\ell_{2,1}$-norm to characterize noise. $\lambda > 0$ is the parameter to balance the influences of the two parts. The optimal solution to Equation (15), which is denoted as $Q_0^*$, naturally defines an affinity relationship that implies the pairwise similarities between features. In this way, the similarity $\mathbb{S}_{kl}^{(i)}$ between two features $x_k^{(i)}$ and $x_l^{(i)}$ can be computed by

$$\mathbb{S}_{kl}^{(i)} = |(Q_0^*)_{lk}| + |(Q_0^*)_{kl}| \tag{16}$$

where $(\bullet)_{lk}$ is the $(l,k)$-th element of a matrix.

In previous methods, the above low-rank learning process can be used with only a single type of feature vectors. Therefore, we extend it to the multi-modal scenario and apply it to feature fusion. We define the multi-modal low-rank learning by

$$\begin{aligned} &\min_{\substack{Q^{(1)},\ldots,Q^{(m)} \\ E^{(1)},\ldots,E^{(m)}}} \sum_{i=1}^{m} (\parallel Q^{(i)} \parallel_* + \lambda \parallel E^{(i)} \parallel_{2,1}) + \alpha \parallel Q \parallel_{2,1}^{(i)} \\ &\text{s. t. } X^{(i)} = X^{(i)}Q^{(i)} + E^{(i)}, j = 1, \ldots, m \end{aligned} \tag{17}$$

where $\alpha > 0$ is a balanced parameter and $m$ is the number of modals. In this way, we can infer a set of matrices $Q^{(1)}$,$Q^{(2)}$, ..., $Q^{(m)}$). In this set, each $n \times n$ matrix $Q^{(i)}$ corresponds to the $i$-th modal $X^{(i)}$. The global solution defined by $m \times n^2$ matrix $Q$ is constructed by arranging $Q^1$,$Q^2$, ..., $Q^m$ as follows:

$$Q = \begin{bmatrix} Q_{11}^1 & Q_{12}^1 & \cdots & Q_{nn}^1 \\ Q_{11}^1 & Q_{12}^2 & \cdots & Q_{nn}^2 \\ \vdots & \vdots & \ddots & \vdots \\ Q_{11}^m & Q_{12}^m & \cdots & Q_{nn}^m \end{bmatrix}. \tag{18}$$

$Q_*^1, Q_*^2, \ldots, Q_*^m$ is defined as the optimal solution to Equation (18). A universal affinity matrix can then obtained by quantifying the columns of $Q$:

$$\mathbb{S}_{kl}^{(i)} = \frac{1}{2}\left(\sqrt{\sum_{i=1}^{m}(Q_{lk}^{(i)})^2} + \sqrt{\sum_{i=1}^{m}(Q_{kl}^{(i)})^2}\right). \tag{19}$$

In manifold learning, the key to solving the manifold is computing the affinity matrix. Therefore, with the affinity matrix computed by Equation (19), we can construct the manifold in low-rank space to obtain fused feature descriptors. Specifically, we use affinity matrix $Q$ to construct Laplacian matrix $L$. There are a number of solutions to this problem. In the proposed method, we follow the spectral hypergraph clustering method [28] and use it in the low-rank space. In this way, we consider each feature vector as a vertex $v$ in the low-rank feature space. If some vertices share the same property, they are connected by a hyperedge $e$. In this method, $L$ is defined as

$$L = I - C \tag{20}$$

where $I$ denotes an $n \times n$ identity matrix. Combined with LRR, $C$ in our proposed framework is defined by

$$C = D_v^{-\frac{1}{2}} U Q D_e^{-1} U^T D_v^{-\frac{1}{2}}. \tag{21}$$

In this equation, $U$ is the matrix that indicates when a vertex belongs to a hyperedge if $U_{i,j} = 1$. $D_e$ and $D_v$ are diagonal matrices, which contain degrees of hyperedge $e$ and degrees of vertex $v$, respectively. Degrees of hyperedge are defined as the numbers of vertices that hyperedges connect, while degrees of vertex are defined as the sum of hyperedge weights connected to this vertex.

To compute $U$, we define that the vertices within a certain distance $\sigma$ from a vertex form a hyperedge with this vertex. Therefore, $U$ can be computed by

$$U_{lk}^{(i)} \begin{cases} 1, & \text{if } \| X_k^{(i)} - X_l^{(i)} \| \leq \sigma \\ 0, & \text{if } \| X_k^{(i)} - X_l^{(i)} \| > \sigma \end{cases}. \tag{22}$$

With $U^{(i)}$ for the $i$-th modal, we use logistic $OR$ to compute a unified $U$ for all modals:

$$U = U^1 | U^2 | ... | U^m. \tag{23}$$

Then, we can compute $D_e$ directly with $U$ by summing each row:

$$D_{e_{ll}} = \sum_{k=1}^{n} U_{lk}. \tag{24}$$

$D_v$ can be computed with $Q$ by summing the items within $d$:

$$D_{v_{ll}} = \sum_{k=1}^{n} Q_{lk} \text{ if } U_{lk} \neq 0. \tag{25}$$

With $L$, we apply the standard eigen-decomposition and obtain the eigenvectors corresponding to the $d$ smallest eigenvalues. Finally, we can obtain the multi-modal features $\overline{X}$ with $d \times n$ dimensions.

## 3. Implementation of Age Estimation

In our implementation of age estimation, we use autoencoders, a CNN, and an RNN to extract the hidden representations. Among the activation functions, we use Rectified Linear Units (ReLUs) since ReLUs are inherently sparse and pretraining can be avoided. Then, their low-rank representations are computed and fusion is embedded in the low-rank space. With the unified affinity matrix, we compute $L$ and use eigen-decomposition to obtain the fused features. Finally, the results are computed by softmax regression, which is mentioned as $\mathcal{F}$ in Equation (2). The developed system is implemented based on DeepLearnToolbox, which contains autoencoders and the CNN [29]. We then add the RNN and low-rank learning to it. The settings of three neural networks are shown in Table 1. To make it possible to solve practical issues, TensorFlow version is under construction.

**Table 1.** The structures of different networks implemented by the proposed method.

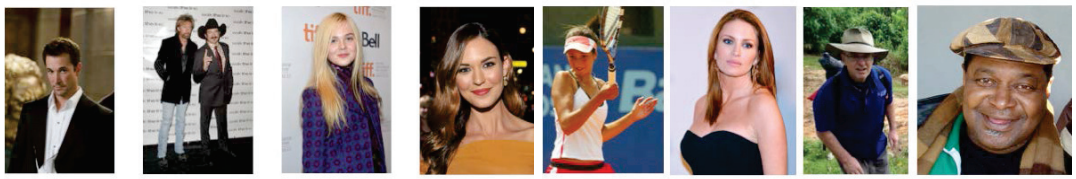
| Network | Structure |
|---|---|
| Autoencoders | Stacked denoising autoencoders with 0.3 corruption level and 5 layers |
| CNN | CNN with 3 convolutional layers and 2 fully-connected layers |
| RNN | RNN with 3 layers |

## 4. Experimental Evaluation

### *4.1. Settings and Datasets*

Images in traditional face datasets are low-resolution and ages are not labeled. Therefore, Rothe et al. collected a large dataset of face images with age information [30], which has been made available for academic research purposes (Available at https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ (accessed on 8 May 2012)). To achieve this, they crawled the data of popular actors on the IMDb

website and fetch their date of birth, name, and gender in their profiles. Besides, they crawled the same meta information and profile images of these people on the Wikipedia website. In this way, 460,723 face images were collected from 20,284 celebrities on IMDb, and 62,328 on Wikipedia were obtained. Sample images of these two datasets are shown in Figure 2. In our experiments, we used the two datasets individually. When we used IMDB, we randomly chose 100,000 images as the training samples and the rest as testing samples. When we used WIKI, we randomly chose 10,000 images as the training samples and the rest as testing samples. This process was repeated 20 times. Then, average performance and standard deviation were recorded. The evaluation was conducted on a desktop with NVIDIA 1080Ti.



(**a**) (**b**)

**Figure 2.** Sample images from datasets from the Internet Movie Database (IMDB) and Wikipedia (WIKI). (**a**) IMDB; (**b**) WIKI.

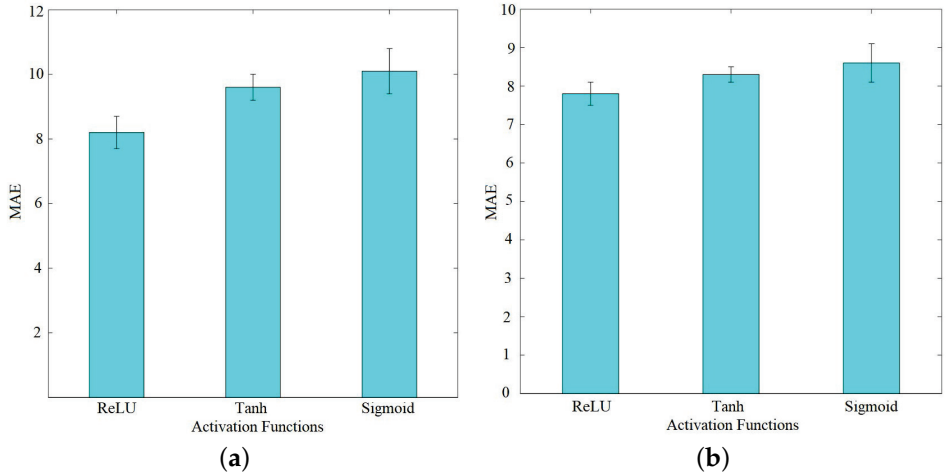
For evaluation, we used mean absolute errors (MAEs), which was computed by

$$MAE = means(|\hat{Y} - Y|) \tag{26}$$

where $\hat{Y}$ is the estimation results. For regression methods, results can be directly computed. For classification methods, we simply treated the age estimation problem as a classification task of 100 classes. $Y$ is the ground truth.

### *4.2. Optimization of Settings*

As mentioned before, activation functions may influence the performance of neural networks. They define the mapped output of a node in different ways. In this way, different activation functions may influence the performance. Therefore, we have tried ReLUs, Sigmoid, and Tanh. The results of these three datasets are shown in Figure 3. Among the three activation functions, ReLUs achieved the best performance among all datasets, which matches recent publications.




(**a**) (**b**)

**Figure 3.** Different activation functions. (**a**) Results of IMDB; (**b**) Results of WIKI.

In the proposed framework, autoencoders, CNNs, and RNNs can be used. Different combinations of them are tested and results are shown in Figure 4. When we integrate the outputs of all neural networks, the performance is the best, which indicates the effectiveness of combining different neural networks.

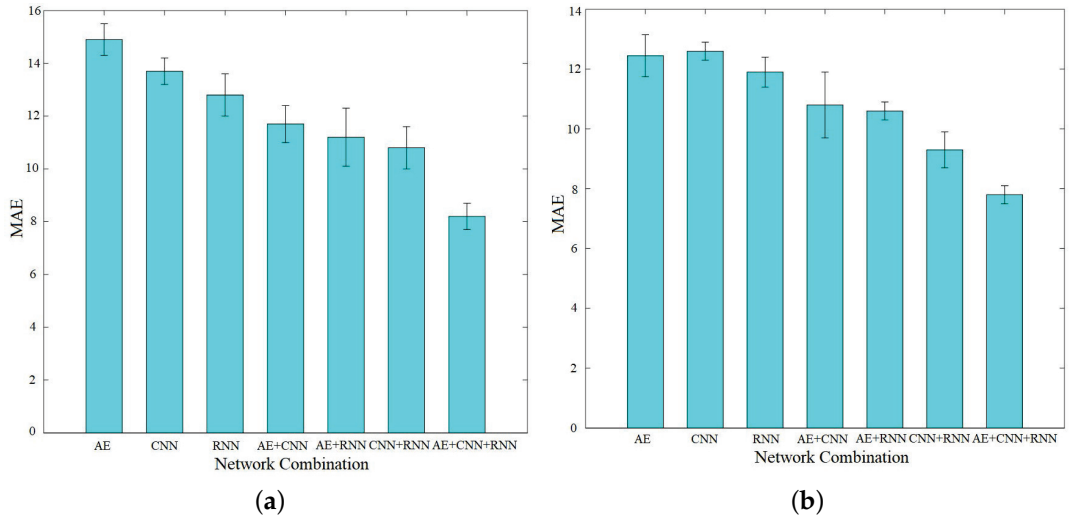



**Figure 4.** Different combinations of networks. (**a**) Results of IMDB; (**b**) Results of WIKI.

### *4.3. Comparison of Multi-Modal Fusion Methods*

We used different methods to integrate outputs of multiple neural networks. The following methods were used:

- Low-Rank Representation (LRR): The proposed method using multiple neural network fusion and low-rank representation.
- Concatenating Different Features (CON): For CON, features from different modals are simply concatenated to construct long features. Principle Component Analysis [31] is then used for dimensionality reduction.
- Multiview Spectral Embedding (MSE) [32]: This method calculates a low-dimensional embedding. In this embedding, the distribution of each modal is sufficiently smooth. The complementary properties of different modals are then explored to obtain a fused representation.
- Multi-View Hypergraph Learning (MHL) [33]: In this method, hypergraph learning is combined with the patch alignment framwork [34]. A multi-view hypergraph Laplacian matrix is constructed, and fused features are computed by solving the standardeigen-decomposition of the multi-view hypergraph Laplacian matrix.

We computed the performance under different dimensions, and the results of different multi-modal fusion methods are shown in Figure 5. According to the figures, all these methods achieved optimal performance among $[400, 600]$. However, optimal performance was not achieved on the same dimensionality. The proposed method uses LRR, and the best performance was achieved on 400 on IMDB and 500 on WIKI. Therefore, these settings were used in the other experiments.

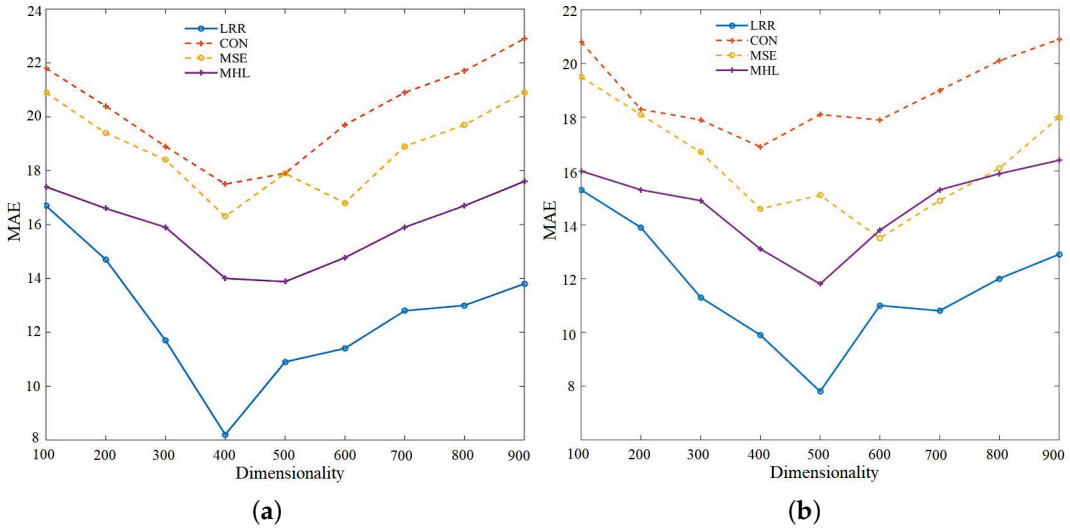



**Figure 5.** Different feature fusion methods. (**a**) Results of IMDB; (**b**) Results of WIKI.

### *4.4. Comparison of Different Methods for Age Estimation*

For age estimation, we compared the following methods, including the proposed Multiple Network Fusion with Low-Rank Representation (MNF-LRR):

- Multiple Network Fusion with Low-Rank Representation (MNF-LRR): The proposed method using multiple neural network fusion and low-rank representation.
- Linear Regression (LR) [17]: This method estimates ages directly by linear regression against feature vectors of facial images. In this paper, HOG [35] was used as image features. Ridge regression (RR-LR) and relevance vector machine (RVM-LR) regression were both implemented by the authors. Their results were similar. We used RVM-LR and set $\nu = 1000$ in the experimental comparison.
- Twin Gaussian Processes (TGPs) [18]: This method applies Gaussian process priors on both covariates and responses. Two Gaussian processes are then modeled as normal distributions over finite index sets of training and testing examples. Finally, outputs can be estimated by minimizing the Kullback–Leibler (K-L) divergence between them. The authors have provided several different implementations of TGP, such as Twin Gaussian Processes with K Nearest Neighbors (TGPKNN), Weighted K-Nearest Neighbor Regression (WKNNRegressor), Gaussian Process Regression (GPR), Hilbert-Schmidt Independent Criterion with K Nearest Neighbors (HSICKNN) and Kernel Target Alignment with K Nearest Neighbors (KTAKNN). We found that TGPKNN outperformed all other methods. Therefore, we used HOG for image features and TGPKNN as the regressor.
- Convolutional Neural Networks (CNNs) [36]. This method uses a simple convolutional net architecture. The network is composed of three convolutional layers and two fully connected layers. The authors have provided the Caffe model for age classification and deployed prototext.
- Deep Expectation (DEX) [30]. The authors here treated age estimation as a classification problem based on deep learning, which was followed by an expected value refinement with softmax. The key to DEX for age regression contains deep learning models with a large amount of data, a robust face alignment process, and softmax-based expected value formulation.

The results of the experimental comparison are shown in Figure 6. Based on the results, we can make the following summarizations:

1. The performance of general mapping learning methods such as LR and TGP is not satisfactory. They are fast and use traditional features such as HOG, but the definition of mapping relationship is oversimplified.

2. The methods based on neural networks such as CNNs and DEX can achieve a stable performance. Neural networks provide descriptive features but require a large amount of training data. Besides, previous neural-network-based methods have not considered multiple features.
3. The performance of the proposed MNF-LRR outperformed the state of the art. We made use of multiple features from different network types and found a reasonable way to fuse them.

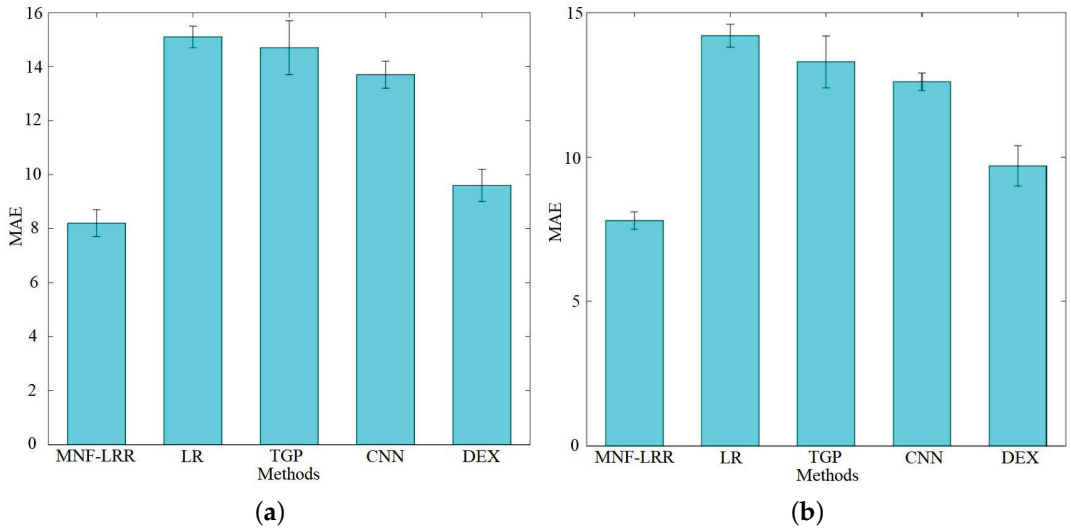



**Figure 6.** Different methods of age estimation. (**a**) Results of IMDB; (**b**) Results of WIKI.

## 5. Discussion

According to the methodology of the developed system and the improvement of experimental performance, the novelty of the proposed method can be shown.

First, the developed system with Multiple Network Fusion with Low-Rank Representation tackles the problem of insufficient descriptive power with insufficient training data. To solve this problem, two types of solutions have been considered in previous methods. One is to improve the descriptive power of a single feature and the other is to fuse multiple features. To improve the descriptive power of a single feature, deep learning has been proved to be effective to represent images in the past few years. To fuse multiple features, manifold learning, low-rank learning, and so on are proposed. However, there have been few attempts to combine the above two solutions. We successfully combine deep learning and low-rank learning. In addition, we implement age estimation. Therefore, the proposed method is theoretically novel.

Second, experimental performance indicates the effectiveness of the proposed method, which can be summarized as follows:

1. We compared different activation functions and different combinations of neural networks to determine the optimal neural network.
2. We compared different feature fusion methods to emphasize the effectiveness of choosing low-rank learning.
3. We compared the proposed method with the state of the art in terms of age estimation to emphasize the overall improvement of the proposed method.

It can be concluded that the proposed method improves age estimation performance.

## 6. Conclusions

In this paper, we propose a data-driven method for image-based age estimation. Multiple Network Fusion with Low-Rank Representation (MNF-LRR) is designed to learn and integrate

multi-modal features. First, multi-modal features are extracted with different neural networks. Second, these features are represented on low-rank space and fused. In this way, a robust representation of facial images for age estimation is computed. In addition, the fused features are connected to softmax, and the estimation results can be obtained. Compared with state of the art, the proposed method is based on multiple features and utilizes multiple neural networks to compute features, which improves the descriptive power of representations. We have conducted experimental evaluation on datasets from the Internet Movie Database (IMDB) and Wikipedia (WIKI). Performance comparison indicates the superiority of the proposed MNF-LRR over previous methods.

**Author Contributions:** Methodology, C.H.; Software, X.W.; Project Administration, Z.Z.; Funding Acquisition, C.H. and Z.Z.; Writing-Review & Editing, C.H. and W.Z.; Visualization, W.Z.

**Funding:** This research was funded by the National Natural Science Foundation of China (61622205), the Fujian Provincial Natural Science Foundation of China (2018J01573, 2016J01327, 2016J01324), the Fujian Provincial High School Natural Science Foundation of China (JZ160472), Fujian Province Universities and Colleges (JK2015033), and the Foundation of Fujian Educational Committee (JAT160357, JAT160358).

**Conflicts of Interest:** No conflict of interest.

## References

1. Chen, B.C.; Chen, C.S.; Hsu, W.H. Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval. *LNCS* **2014**, *8694*, 768–783.
2. Eidinger, E.; Enbar, R.; Hassner, T. Age and Gender Estimation of Unfiltered Faces. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 2170–2179. [CrossRef]
3. Hu, H.; Otto, C.; Jain, A.K. Age estimation from face images: Human vs. machine performance. In Proceedings of the International Conference on Biometrics, Madrid, Spain, 4–7 June 2013; pp. 1–8.
4. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. In Proceedings of the European Conference on Computer Vision, Freiburg, Germany, 2–6 June 1998; pp. 484–498.
5. Fu, Y.; Huang, T.S. Human Age Estimation With Regression on Discriminative Aging Manifold. *IEEE Trans. Multimed.* **2008**, *10*, 578–584. [CrossRef]
6. 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. [PubMed]
7. Ahonen, T.; Hadid, A.; Pietikainen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
8. Guo, G.; Mu, G.; Fu, Y.; Huang, T.S. Human age estimation using bio-inspired features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 112–119.
9. Fu, Y.; Guo, G.; Huang, T.S. Age Synthesis and Estimation via Faces: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976. [PubMed]
10. He, R.; Zheng, W.S.; Tan, T.; Sun, Z. Half-Quadratic-Based Iterative Minimization for Robust Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 261–275. [PubMed]
11. Gupta, K.; Majumdar, A. Imposing Class-Wise Feature Similarity in Stacked Autoencoders by Nuclear Norm Regularization. *Neural Process. Lett.* **2018**, *48*, 615–629. [CrossRef]
12. Kim, J.; Bukhari, W.; Lee, M. Feature Analysis of Unsupervised Learning for Multi-task Classification Using Convolutional Neural Network. *Neural Process. Lett.* **2018**, *47*, 783–797. [CrossRef]
13. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Group-Aware Deep Feature Learning For Facial Age Estimation. *Pattern Recognit.* **2016**, *66*, 82–94. [CrossRef]
14. Yu, J.; Yang, X.; Gao, F.; Tao, D. Deep Multimodal Distance Metric Learning Using Click Constraints for Image Ranking. *IEEE Trans. Cybern.* **2017**, *47*, 4014–4024. [CrossRef] [PubMed]
15. Yu, J.; Kuang, Z.; Zhang, B.; Zhang, W.; Lin, D.; Fan, J. Leveraging Content Sensitiveness and User Trustworthiness to Recommend Fine-Grained Privacy Settings for Social Image Sharing. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 1317–1332. [CrossRef]
16. Kwon, Y.H.; da Vitoria Lobo, N. Age classification from facial images. *Comput. Vis. Image Underst.* **1999**, *74*, 1–21. [CrossRef]

17. Agarwal, A.; Triggs, B. Recovering 3D human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 44–58. [CrossRef] [PubMed]
18. Bo, L.; Sminchisescu, C. *Twin Gaussian Processes for Structured Prediction*; Kluwer Academic Publishers: Alphen aan den Rijn, The Netherlands, 2010; pp. 28–52.
19. Tian, Q.; Xue, H.; Qiao, L. Human Age Estimation by Considering both the Ordinality and Similarity of Ages. *Neural Process. Lett.* **2015**, *43*, 1–17. [CrossRef]
20. Liu, X.; Li, S.; Kan, M.; Zhang, J.; Wu, S.; Liu, W.; Han, H.; Shan, S.; Chen, X. AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 258–266.
21. Kuang, Z.; Huang, C.; Zhang, W. Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 338–343.
22. Yang, X.; Gao, B.B.; Xing, C.; Huo, Z.W. Deep Label Distribution Learning for Apparent Age Estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 344–350.
23. Ranjan, R.; Zhou, S.; Chen, J.C.; Kumar, A.; Alavi, A.; Patel, V.M.; Chellappa, R. Unconstrained Age Estimation with Deep Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 351–359.
24. Wang, X.; Guo, R.; Kambhamettu, C. Deeply-Learned Feature for Age Estimation. In Proceedings of the Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; pp. 534–541.
25. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
26. Yoshua, B. *Learning Deep Architectures for AI*; Foundations and Trends in Machine Learning Series; Now Publishers Inc.: Breda, The Netherlands, 2009; Volume 2, pp. 1–127.
27. Chen, M.; Weinberger, K.Q.; Sha, F.; Bengio, Y. Marginalized Denoising Auto-encoders for Nonlinear Representations. In Proceedings of the IEEE International Conference on Machine Learning, Beijing, China, 3–6 December 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1476–1484.
28. Zhou, D.; Huang, J.; Scholkopf, B. Learning with Hypergraphs: Clustering, Classification, and Embedding. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2007; Volume 19, pp. 1601–1608.
29. Palm, R.B. Prediction as a Candidate for Learning Deep Hierarchical Models of Data. Master's Thesis, Technical University of Denmark, Lyngby, Denmark, 2012.
30. 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*, 144–157. [CrossRef]
31. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Br. J. Educ. Psychol.* **1933**, *24*, 417–520. [CrossRef]
32. Xia, T.; Tao, D.; Mei, T.; Zhang, Y. Multiview Spectral Embedding. *IEEE Trans. Syst. Man Cybern. Part B* **2010**, *40*, 1438–1446.
33. Hong, C.; Yu, J.; Li, J.; Chen, X. Multi-view hypergraph learning by patch alignment framework. *Neurocomputing* **2013**, *118*, 79–86. [CrossRef]
34. Zhang, T.; Tao, D.; Li, X.; Yang, J. Patch Alignment for Dimensionality Reduction. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1299–1313. [CrossRef]
35. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; IEEE Press: Piscataway, NJ, USA, 2005; pp. 886–893.
36. Levi, G.; Hassncer, T. Age and gender classification using convolutional neural networks. In Proceedings of the Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 34–42.