SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation

Tsun-Yi Yang[1,2] **Yi-Hsuan Huang**[1,2] **Yen-Yu Lin**[1] **Pi-Cheng Hsiu**[1] **Yung-Yu Chuang**[1,2]

[1]Academia Sinica, Taiwan [2]National Taiwan University, Taiwan

shamangary@citi.sinica.edu.tw yihsuan32@iis.sinica.edu.tw {yylin,pchsiu}@citi.sinica.edu.tw cyy@csie.ntu.edu.tw

Abstract

This paper presents a novel CNN model called Soft Stagewise Regression Network (SSR-Net) for age estimation from a single image with a compact model size. Inspired by DEX, we address age estimation by performing multi-class classification and then turning classification results into regression by calculating the expected values. SSR-Net takes a coarse-to-fine strategy and performs multi-class classification with multiple stages. Each stage is only responsible for refining the decision of its previous stage for more accurate age estimation. Thus, each stage performs a task with few classes and requires few neurons, greatly reducing the model size. For addressing the quantization issue introduced by grouping ages into classes, SSR-Net assigns a dynamic range to each age class by allowing it to be shifted and scaled according to the input face image. Both the multi-stage strategy and the dynamic range are incorporated into the formulation of soft stagewise regression. A novel network architecture is proposed for carrying out soft stagewise regression. The resultant SSR-Net model is very compact and takes only 0.32 MB. Despite its compact size, SSR-Net's performance approaches those of the state-of-the-art methods whose model sizes are often more than $1500\times$ larger.

1 Introduction

Predicting the real (biological) age of a person from a single face image is a classic problem in computer vision and artificial intelligence [Ramanathan *et al.*, 2009]. It is useful for many applications such as surveillance, product recommendation, human-computer interface, and market analysis. The problem is challenging because there are significant variations on appearance among people of the same age. Older people may look younger than some younger people and vice versa. Thus, the task of estimating the real age from appearance is challenging even for humans.

It is intuitive to formulate the age estimation problem as a regression problem since age is a continuous value rather than a set of discrete classes [Agustsson *et al.*, 2017; Rothe *et al.*, 2016b]. However, as pointed out by previous studies [Chang *et al.*, 2011; Rothe *et al.*, 2016a; Tan *et al.*, 2017], the regression-based age estimation approaches could suffer from overfitting because of randomness in the aging process and ambiguous mapping between face appearance and its real age. On the other hand, people can be easily categorized into several age groups such as teenagers, middle-aged or elderly people. Thus, many studies have addressed the age estimation problem using multi-class classification approaches by quantizing ages into groups [Rothe *et al.*, 2015; Rothe *et al.*, 2016a; Tan *et al.*, 2017]. However, casting age estimation as a multi-class classification problem has to face the issue that age groups are ordinal and highly correlated rather than independent classes. In addition, quantizing ages into age groups could suffer from problems with quantization error and ambiguity among age groups. Some methods adopt ordinal information [Chang *et al.*, 2011; Zhang *et al.*, 2017] for acquiring relative ordering between ages and resolving ambiguity among age groups. Distribution learning approaches model ages as a learnable distribution [Geng *et al.*, 2014; Hou *et al.*, 2017] for addressing the group ambiguity problem. However, approaches using ordinal information and distribution learning usually need additional information such as similarities between distributions or ranks. In addition, they often require more complex loss functions and algorithms.

Most state-of-the-art CNN-based age estimation methods are built upon complex networks or ensembles of networks [Simonyan and Zisserman, 2014; Chen *et al.*, 2017]. Their models are often bulky with model sizes larger than 500 MB. Hence, they are not suitable to be adopted on platforms with limited memory and computation resource such as mobile and embedded devices. Some efforts have been made for compact general-purpose CNN models with small memory footprints [Howard *et al.*, 2017; Huang *et al.*, 2017] so that intelligent applications can be run on such devices. For age estimation, Niu et al. proposed ORCNN whose model consumes around 1.7 MB of memory [Niu *et al.*, 2016].

This paper proposes a novel CNN model called Soft Stagewise Regression Network (SSR-Net) for age estimation. The model is compact with only 0.32 MB memory overhead and achieves the state-of-the-art performance[1]. SSR-Net is inspi-

[1]Source code available at https://github.com/shamangary/SSR-Net

red by DEX [Rothe *et al.*, 2015; Rothe *et al.*, 2016a]. DEX casts age estimation as a multi-class classification problem and turns the classification results into regression by calculating the expected value as the age. One problem with DEX is that it requires a large number of neurons, one for each of the age classes. The number of links at the final fully-connected layer is the product of the number of features and the number of neurons (classes). Thus, many neurons lead to many parameters and a large model. SSR-Net addresses the problem with a coarse-to-fine strategy. Each stage only performs intermediate classification with a small number of classes, say "relatively younger", "about right" and "relatively older" within the current age group. The next stage refines the decision within the age group assigned by the previous stage. This way, each stage only requires a small number of neurons and the model size can be much reduced. Another problem with classification-based approaches is quantization of ages. SSR-Net addresses this issue by introducing a dynamic range to each age group. The age interval of each group can be shifted and scaled depending on the input face image. These ideas are incorporated into a formulation for soft stagewise regression. A novel network structure is proposed for realizing soft stagewise regression. Compared to the approaches based on ordinal information or distribution learning, SSR-Net can be trained with a simple regression loss and an end-to-end fashion without requiring extra information such as distribution/rank similarities. Experiments show that SSR-Net outperforms existing compact networks including MobileNet [Howard *et al.*, 2017], DenseNet [Huang *et al.*, 2017], and ORCNN [Niu *et al.*, 2016]. The performance of SSR-Net approaches those of the state-of-the-art methods with much larger model sizes, usually more than $1500\times$ larger than ours.

2 Related work

In this section, we review the state-of-the-art age estimation methods by organizing them into four categories.

Regression. It is intuitive to cast the age estimation problem as a regression problem. Rothe et al. fed CNN features to support vector regression [Chang and Lin, 2011] for age estimation [Rothe *et al.*, 2016b]. Agustsson et al. proposed Anchored Regression Network (ARN) which combines multiple linear regressors over soft assignments to anchor points [Agustsson *et al.*, 2017]. As pointed out by previous studies [Chang *et al.*, 2011; Rothe *et al.*, 2016a], regression-based approaches often suffer from overfitting because of randomness in the aging process and ambiguous mapping.

Multi-class classification and age grouping. DEX performs age estimation by carrying out multi-class classification and then calculating the expected value as the age estimation [Rothe *et al.*, 2015; Rothe *et al.*, 2016a]. Liu et al. used regression and classification simultaneously for age estimation [Liu *et al.*, 2015]. Malli et al. used age groups and their age-shifted groupings for training an ensemble of deep learning models [Malli *et al.*, 2016]. Tan et al. proposed an age group-n-encoding method in which adjacent ages are grouped together and each age corresponds to several groups [Tan *et al.*, 2017]. Based on the grouping, they transformed age estimation into a series of binary classification tasks. This type of approaches often suffers from problems with group ambiguity and quantization errors.

Distribution learning. Geng et al. proposed two different adaptive distribution learning methods, IIS-ALDL and BFGS-ALDL, for age estimation [Geng *et al.*, 2014]. The standard deviation of a distribution is updated according to KL-divergence. For addressing the problem with shortage of training data with exact ages, Hou et al. proposed Label Distribution Learning to utilize neighboring ages while learning a particular age [Hou *et al.*, 2017].

Ordinal information. It is easier to rank two people by their ages than estimating their ages directly. Some methods focus on learning relative ordering from a dataset for age estimation. OHRank successfully harvests ordering relation for age estimation by developing a cost-sensitive framework with multiple binary classifications [Chang *et al.*, 2011]. Ranking-CNN pre-trains several basic CNNs on a large dataset such as ImageNet and fine-tunes them with ordinal age labels [Chen *et al.*, 2017]. Zhang et al. proposed a paradigm for mapping multiple age comparisons into an age distribution posterior for age estimation [Zhang *et al.*, 2017].

3 Soft stagewise regression network

This section first states the problem. Then, we describe the two key ideas, the stagewise regression and the dynamic range. Finally, the network architecture is given and the formulation for soft stagewise regression is presented.

3.1 Problem formulation

In the problem of real age estimation from a single face image, we are given a set of training face images $X = \{x_n \mid n = 1..N\}$ and the real age $y_n \in Y$ for each image x_n, where N is the number of images and Y is the interval of ages. The goal is to find a function F that predicts $\tilde{y} = F(x)$ as the age for a given image x. For training, we search for the function F by minimizing the mean absolute error (MAE) between the predicted and the real ages,

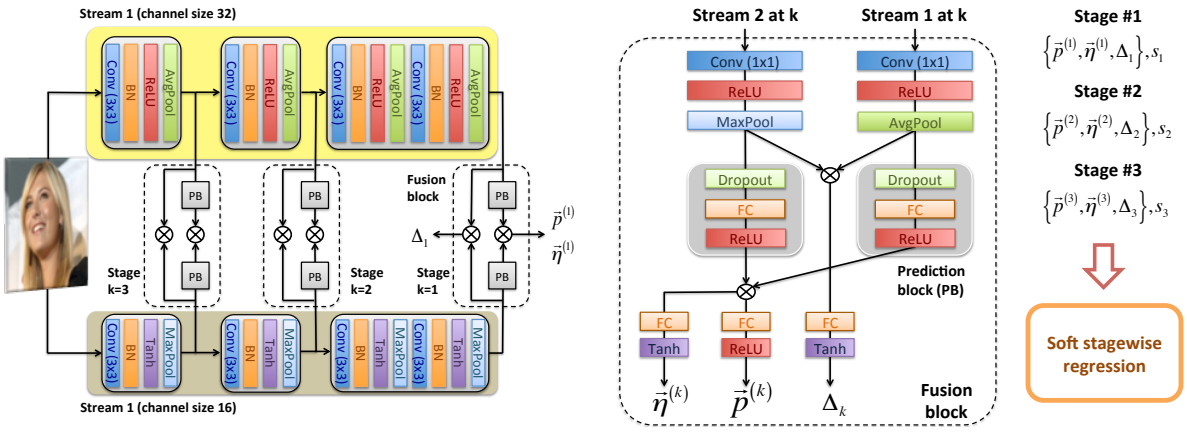
$$J(X) = \frac{1}{N}\sum_{n=1}^{N} |\tilde{y}_n - y_n| \,, \tag{1}$$

where $\tilde{y}_n{=}F(x_n)$ is the predicted age for training image x_n.

3.2 Stagewise regression

Previous work has turned the regression problem of age estimation into solving a multi-class classification problem and then calculating the expected value as the predicted age. For example, DEX [Rothe *et al.*, 2015; Rothe *et al.*, 2016a] divides the age interval $Y = [0, V]$ into s non-overlapping bins uniformly. Thus, the width w of each bin is $\frac{V}{s}$. Let's denote the representative age of the i-th bin as μ_i and DEX chooses $\mu_i = i\left(\frac{V}{s}\right)$. DEX trains a network for the s-class age classification problem. For a given image x, the network outputs a distribution vector $\vec{p} = (p_0, p_1, \ldots, p_{s-1})$ indicating the probability that x belongs to each of age groups. The age is then predicted by calculating the following expected value,

$$\tilde{y} = \vec{p} \cdot \vec{\mu} = \sum_{i=0}^{s-1} p_i \cdot \mu_i = \sum_{i=0}^{s-1} p_i \cdot i\left(\frac{V}{s}\right). \tag{2}$$



(a) SSR-Net (b) Fusion block and regression

Figure 1: (a) The network structure of the proposed Soft Stagewise Regression Network (SSR-Net) with three stages ($K = 3$). The size of pooling is fixed at 2×2 for all stages. (b) The detailed structure of the fusion block in SSR-Net and the structure of the prediction block (PB) within the fusion block.

To have a more accurate estimation, DEX divides the age interval finely and set the bin width as one year old, i.e., there are 101 bins if $Y = [0..100]$. It leads to a large number of parameters for the fully-connected (FC) layer at the final stage and consumes a lot of memory.

To reduce the model size without sacrificing much accuracy, we propose to use a coarse-to-fine strategy with multi-stage prediction. Assume that there are K stages and there are s_k bins for the k-th stage. For each stage, we train a network F_k that generates the distribution $\vec{p}^{(k)} = (p_0^{(k)}, p_1^{(k)}, \ldots, p_{s_k-1}^{(k)})$ for that stage. The age is predicted by the following formula for stagewise regression,

$$\tilde{y} = \sum_{k=1}^{K} \vec{p}^{(k)} \cdot \vec{\mu}^{(k)} = \sum_{k=1}^{K} \sum_{i=0}^{s_k-1} p_i^{(k)} \cdot i \left(\frac{V}{\prod_{j=1}^{k} s_j} \right). \quad (3)$$

The last term in the above equation is the bin width $w_k = \frac{V}{\prod_{j=1}^{k} s_j}$ for the k-th stage and i is the bin index. It is easier to understand the intuition behind stagewise regression by walking through a concrete example. Assume that we want to estimate an age within the range of $0 \sim 90$ years old ($V = 90$). Assume we have two stages ($K = 2$) and there are three bins for either stage ($s_1 = s_2 = 3$). From classification point of view, stage #1 classifies the image as youth ($0 \sim 30$), middle age ($30 \sim 60$) or old age ($60 \sim 90$). For stage #2, each bin in stage #1 is further divided into $s_2 = 3$ bins. Thus, width of the bins in stage #2 is $\frac{90}{3 \cdot 3} = 10$. The classifier of stage #2 classifies an image as relatively younger ($+0 \sim 10$), in the middle ($+10 \sim 20$) or relatively older ($+20 \sim 30$) within the age group assigned by stage #1. Note that there is only one classifier at stage #2, shared by all age groups of stage #1. Turning the stagewise classification into regression gives us the formula in Equation 3. Stage #1 predicts the age with a coarse granularity while stage #2 refines it with a finer granularity. The advantage of stagewise regression is that the number of classes is small at each stage, leading to much fewer parameters and a more compact model.

3.3 Dynamic range

Dividing the age interval uniformly into non-overlapping bins is less flexible on handling age group ambiguity and age continuity. The problem is even more serious at a coarse granularity. We address this issue by introducing a dynamic range for each bin; that is, we allow each bin to be shifted and scaled according to the input image. There are several possible options to apply the shift and the scale to a bin. To use the same stagewise regression formula in Equation 3, we opt to modify the bin index i and bin width w_k in Equation 3 for the adjustment of the bin shift and the bin scale respectively. For adjusting the bin width w_k at the k-th stage, we introduce a term Δ_k to modify s_k into $\bar{s}_k$ as follows,

$$\bar{s}_k = s_k(1 + \Delta_k), \quad (4)$$

where Δ_k is the output of a regression network given the input image. The detail of the regression network will be given in Section 3.4. After modifying s_k, the bin width now becomes

$$\bar{w}_k = \frac{V}{\prod_{j=1}^{k} \bar{s}_j}. \quad (5)$$

Thus, the adjustment of s_k effectively changes the bin width. For shifting bins, we add an offset term η to each bin index i. There are s_k bins at the k-th stage. Thus, we need an offset vector for the k-th stage, $\vec{\eta}^{(k)} = (\eta_0^{(k)}, \eta_1^{(k)}, \ldots, \eta_{s_k-1}^{(k)})$. Again, the offset vector is the output of a regression network on the input image. The bin index i is modified as follows,

$$\bar{i} = i + \eta_i^{(k)}. \quad (6)$$

The modified bin index $\bar{i}$ effectively shifts the i-th bin. Both the scale and shift of bins are regression results of the input image. The input-dependent dynamic range provides more accurate refinement according to the input image.

3.4 Network structure

Figure 1(a) illustrates the overall network structure of the proposed SSR-Net. Motivated by the complementary 2-stream

structure proposed by Yang et al. [Yang *et al.*, 2017], we adopt a 2-stream model where there are two heterogeneous streams. For both streams, the basic building block is composed of 3×3 convolution, batch normalization, non-linear activation and 2×2 pooling. However, different types of activation functions (ReLU versus Tanh) and pooling (average versus maximum) are adopted for each stream to make them heterogeneous. This way, they could explore different features and their fusion could improve the performance.

Features from different levels are adopted for different stages. For each stage, features from both streams at some level are fed into a fusion block which is illustrated in Figure 1(b). The fusion block is responsible for generating stagewise outputs, the distribution $\vec{p}^{(k)}$, the offset vector $\vec{\eta}^{(k)}$, and the scale factor Δ_k, for the k-th stage. In the fusion block, features from both streams first go through 1×1 convolution, activation and pooling for having more compact features. For obtaining Δ_k, the two obtained feature maps are fused by element-wise multiplication $\bigotimes$. The product then goes through a fully-connected layer and then a Tanh function for obtaining a value in $[-1, 1]$ as Δ_k. Both $\vec{p}^{(k)}$ and $\vec{\eta}^{(k)}$ are vectors and more complex. Thus, the features go through an additional prediction block before taking element-wise multiplication, FC layer and activation. Since $\vec{p}^{(k)}$ represents a distribution, ReLU is used as its activation for obtaining positive values. On the other hand, Tanh is used for $\vec{\eta}^{(k)}$ to allow shift on both positive and negative sides.

3.5 Soft stagewise regression

Given the network's stagewise outputs $\left\{\vec{p}^{(k)}, \vec{\eta}^{(k)}, \Delta_k\right\}$ for an input image x and the numbers of bins s_k, the predicted age $\tilde{y}$ for x is calculated as

$$\tilde{y} = \sum_{k=1}^{K} \sum_{i=0}^{s_k - 1} p_i^{(k)} \cdot \bar{i} \left(\frac{V}{\prod_{j=1}^{k} \bar{s}_j} \right), \tag{7}$$

where $\bar{i}$ is the shifted bin index defined in Equation 6 and $\bar{s}_j$ is adjusted bin number defined in Equation 4. We name the formula in Equation 7 soft stagewise regression because the bins are adjusted by fractional numbers. Softness is brought into the bin indexes and the bin widths this way. With the predicted age $\tilde{y}$, by minimizing the L_1 loss defined by MAE in Equation 1, we obtain the SSR-Net model for age estimation.

4 Experiments

This section first describes preprocessing, experimental settings, and competing methods. Next, we report experiments on IMDB-WIKI, MORPH2 and MegaAge-Asian datasets.

4.1 Preprocessing and experimental setting

We performed experiments on several benchmark datasets for age estimation, including the IMDB-WIKI [Rothe *et al.*, 2015; Rothe *et al.*, 2016a], MORPH2 [Ricanek and Tesafaye, 2006], and MegaAge-Asian [Zhang *et al.*, 2017] datasets. Following the procedure suggested by previous work [Zhang *et al.*, 2017; Niu *et al.*, 2016], for preprocessing, all face images were aligned using facial landmarks such as eyes and the nose. After alignment, the face region of each image was cropped and resized to the resolution of 64×64. Note that the size is much smaller than the resolution 224×224 used in previous state-of-the-art methods such as DEX [Rothe *et al.*, 2016a] and ARN [Agustsson *et al.*, 2017]. The lower resolution is necessary for mobile and embedded devices with limited resources.

The experiments were performed on a machine with an Intel i7 CPU and an NVIDIA GTX1080Ti. The program was implemented with Keras. The custom layer for soft stagewise regression is powered by Keras' automatic differentiation. For training, common data augmentation tricks including zooming, shifting, shearing, and flipping were randomly activated. Unless specified otherwise, SSR-Net uses three stages with $s_1 = s_2 = s_3 = 3$, i.e., SSR-Net(3,3,3). The Adam method [Kingma and Ba, 2014] was used for optimizing the network parameters with 90 epochs. The learning rate is 0.002 initially and reduced by a factor 0.1 every 30 epochs. The batch size is 128 for the IMDB dataset and 50 for other datasets. The training time for SSR-Net is around three hours including pre-training.

4.2 Competing methods

We compare the proposed SSR-Net model with a set of state-of-the-art deep-learning-based age estimation methods. The competing methods can be roughly categorized into two groups, bulky models and compact models, according to their model sizes.

The bulky models put more emphasis on prediction accuracy. They are usually more accurate, but at the expense of bulky network models. Many such models are built upon VGG16 [Simonyan and Zisserman, 2014]. **DEX** [Rothe *et al.*, 2015; Rothe *et al.*, 2016a] casts the regression problem of age estimation into a multi-class classification problem and uses the expected value as age estimation. **ARN** [Agustsson *et al.*, 2017] solves the age regression problem by combining multiple linear regressors over soft assignments to anchor points. **AP** [Zhang *et al.*, 2017] trains a network that jointly performs ordinal hyperplane classification and posterior distribution learning. **Hot** [Rothe *et al.*, 2016b] feeds VGG16 features into support vector regression (SVR) for age prediction. **RankingCNN** [Chen *et al.*, 2017] ensembles a series of binary classification networks, each of which is trained with ordinal age labels.

Compact models emphasize reduced memory footprint and could sacrifice accuracy for memory and speed. There are less age estimation models in this category. **ORCNN** [Niu *et al.*, 2016] transforms the ordinal regression problem into a series of binary classification problems and uses a multiple output CNN to collectively solve these sub-problems. **MR-CNN** [Niu *et al.*, 2016] uses a similar network but for metric regression. **MobileNet** [Howard *et al.*, 2017] replaces standard convolution with depthwise separable convolution for reducing parameters and computation overhead. **DenseNet** [Huang *et al.*, 2017] connects each layer to every other layer in a feed-forward fashion and can achieve good performance with substantially fewer parameters. Both MobileNet and DenseNet are general-purpose network models with tunable parameters. We chose the parameters so that their model sizes are roughly 1 MB for fair comparison with SSR-Net.

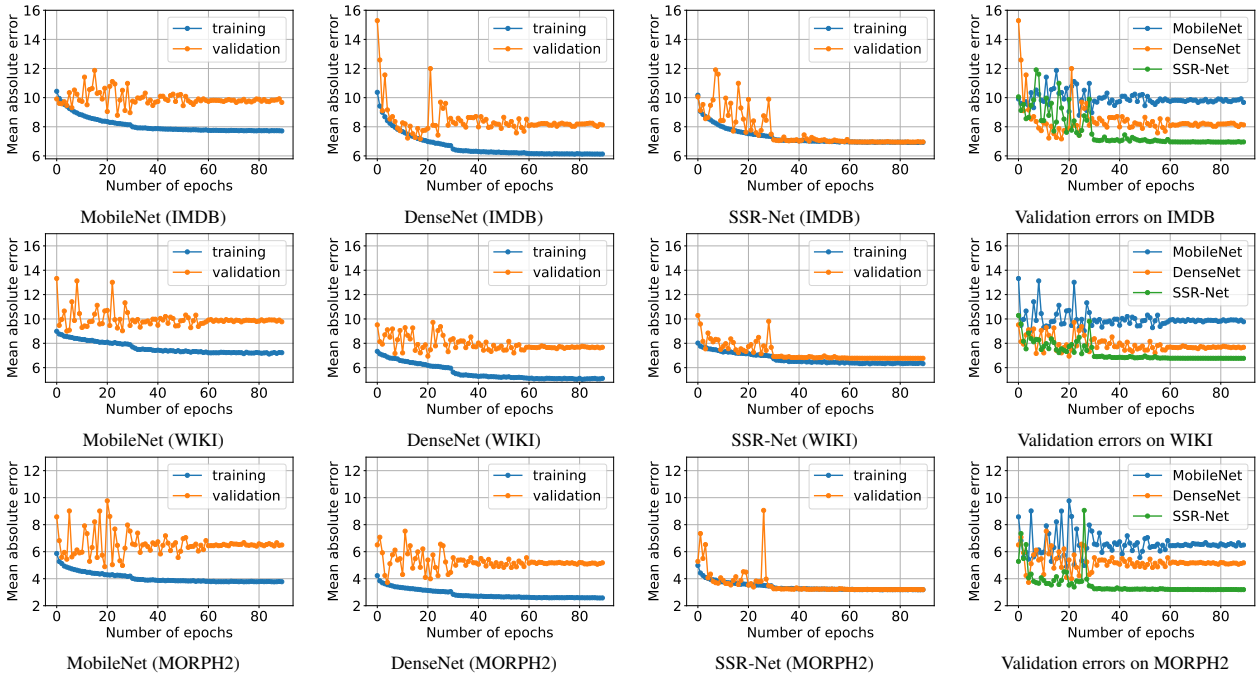


Figure 2: Comparisons of the training progression for MobileNet, DenseNet, SSR-Net (from left to right), and their validation comparisons on IMDB, WIKI and MORPH2 (from top to bottom). For each dataset, 80% of images were used as the training set while the remaining 20% acted as the validation set. For the first three columns, blue curves represent the progression of the training errors in MAE while orange curves are for the validation errors. If the two curves are close, it means that the model obtained from the training data can be better applied to the validation data. The models with this property suffer less from overfitting. From this point of view, SSR-Net outperforms the other two methods on all three datasets. The last column shows that SSR-Net outperforms the others in the MORPH2 validation set.

4.3 Experiments on IMDB-WIKI

The IMDB-WIKI dataset is the largest face image dataset with age labels. It contains 523,051 face images of 20,284 celebrities. Among them, 460,723 images were collected from IMDB and the remaining 62,328 were from Wikipedia.

Although the IMDB-WIKI dataset is the largest for age estimation, as pointed out by previous work [Tan *et al.*, 2017], it contains more noise such as inaccurate ages and images with no face or multiple faces. Therefore, it is not suitable to be used for performance evaluation with MAE. Like most previous work, we used it for pre-training. In addition, we used the IMDB-WIKI dataset for observing the overfitting properties of different methods. The IMDB and WIKI dataset were separately trained with 80% of the images. The other 20% served as the validation set. Since the IMDB dataset is much larger, we trained our model on IMDB first and used the trained model as the starting point for training on the WIKI dataset. Figure 2 compares the progression of the training processes for MobileNet, DenseNet and the proposed SSR-Net on the IMDB, WIKI and MORPH2 datasets. The blue curves show the progression of the training error while the orange ones for the validation sets. It is clear that the blue and orange curves of SSR-Net are closer than other two methods. It means that our model trained on the training set can be applied to the validation set more successfully. The discrepancy between the training errors and the validation errors shows that MobileNet and DenseNet suffer more from overfitting.

4.4 Experiments on MORPH2

MORPH2 is the most popular benchmark dataset for age estimation. It has around 55,000 face images of 13,000 people. Their ages range from 16 to 77 years old. Similiar to previous work [Niu *et al.*, 2016; Zhang *et al.*, 2017], we randomly divided the dataset into independent training (80%) and testing (20%) sets. MAE is used as the metric for performance evaluation.

Table 1 reports MAE values on MORPH2 for a set of state-of-the-art network models for age estimation, including both bulky and compact ones. For better accuracy, the bulky models often use higher-resolution inputs $(224 \times 224 \times 3)$ for retaining more information. They often consume more than 500 MB memory because of a large number of parameters. For training models with massive parameters, more images are required. Thus, in addition to IMDB-WIKI, some of them also use ImageNet [Russakovsky *et al.*, 2015] for pre-training, leading to a much longer training time. Ranking-CNN takes an ensemble of binary classification networks and consumes even more memory, up to 2.2 GB. Although achieving better performance, it is difficult to adopt bulky models on mobile and embedded devices with limited resource because of their bulky model sizes.

On the other hand, for reducing memory footprint, compact models usually take lower-resolution images as inputs $(64 \times 64 \times 3)$ and contains much fewer parameters. The proposed SSR-Net is very compact and only consumes 0.32 MB of memory while MobileNet and DenseNet take roughly 1

		Bulky models						Compact models				
Methods		AP	ARN	DEX		Hot	RankingCNN	SSR-Net	MobileNet	DenseNet	MRCNN	ORCNN
Pre.	ImageNet	√	√	√	√	√	unflitered faces	–	–	–	unknown	unknown
	IMDB-WIKI	√	√	√	–	–		√	√	√		
Input size		$224 \times 224 \times 3$						$64 \times 64 \times 3$				
Model size		VGG16 ≈ 500 MB					2.2 GB	**0.32 MB**	1.0 MB	1.1 MB	1.7 MB	
inference time on GPU/CPU (10^{-3}sec)		–						0.17/2.69	**0.10/1.07**	0.75/28.8	–	
MAE		**2.52**	3.00	2.68	3.25	3.45	2.96	**3.16**	6.50	5.05	3.42	3.27

Table 1: Comparisons of state-of-the-art methods on the MORPH2 dataset. There are two categories, bulky models and compact models. The former takes inputs with a larger resolution and consumes more memory while the latter uses a smaller resolution and has a smaller memory footprint. Both MAE values and inference time are reported.

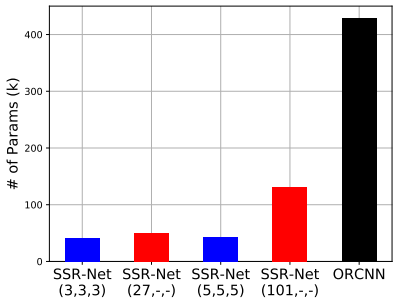


Figure 3: Comparisons on the number of parameters for SSR-Net with different configurations and ORCNN.

MB and MRCNN/ORCNN consumes 1.7 MB. Before training with MORPH2, we used the IMDB-WIKI dataset for pre-training. SSR-Net achieves 3.16 MAE, the best among compact models. It even surpasses several bulky models despite that it consumes less than $1/1500$ of their model sizes. With the extremely compact model (0.32 MB) and reasonable performance, SSR-Net is suitable to be adopted on mobile and embedded platforms. The last row of Figure 2 shows the training/validation curves for MobileNet, DenseNet and SSR-Net on MORPH2. Again, SSR-Net suffers less from overfitting compared to the other two compact models.

Table 1 also reports inference times for compact models if running on a GPU or a CPU. MobileNet is slightly faster than SSR-Net, but its age estimation performance is much worse. DenseNet is much slower than SSR-Net along with its larger model size and worse performance. We are not able to report the inference time and training progression for MRCNN and ORCNN as their source code is not made publicly available.

The use of soft dynamic range is important for SSR-Net. Without it, the MAE is 9.29 for SSR-Net. It shows that the flexibility provided by the dynamic range is essential to multi-stage regression. The coarse-to-fine stagewise regression approach reduces the number of neurons, one for each of classes in classification. For SSR-Net(s_1, s_2, s_3), $s_1 + s_2 + s_3$ neurons are required. With roughly the same number of classes, the single-stage approach would require $s_1 \times s_2 \times s_3$ neurons. Figure 3 compares the number of parameters for SSR-Net with different configurations and ORCNN. The multi-stage design generally leads to fewer parameters than its single-stage counterpart, especially when V is large. In addition, the proposed network structure also helps in reducing parameters. Overall, compared with ORCNN, SSR-Net consumes fewer parameters while achieving better MAE.

	MobileNet	DenseNet	SSR-Net (3,3,3)
CA(3)	0.440	0.517	**0.549**
CA(5)	0.606	0.694	**0.741**

Table 2: Results on the MegaAge-Asian dataset.

4.5 Experiments on MegaAge-Asian

Human races could play an important role on the relationship between ages and appearance. Most face image datasets contain images of Westerners. To validate how the proposed SSR-Net performs for other races such as Asians, we have also performed experiments on the MegaAge-Asian dataset [Zhang *et al.*, 2017]. It contains 40,000 face images of Asians with ages from 0 to 70. Following the protocol of Zhang et al., 3,945 images were reserved for testing and the cumulative accuracy (CA) was used as the evaluation metric. CA is defined as $\text{CA}(n) = K_n/K \times 100$ in which K is the total number of testing images and K_n represents the number of testing images whose absolute errors are smaller than n. The SSR-Net model trained on IMDB-WIKI was used as the starting point for training and the training images of MegaAge-Asian were used to train SSR-Net. The same training procedure was used to train both MobileNet and DenseNet. Table 2 reports CA(3) and CA(5) for the compared methods. It is clear that SSR-Net performs better than the other two compact models.

5 Conclusion

In this paper, we propose a novel method for age estimation, Soft Stagewsie Regression Network (SSR-Net). It is both compact and efficient. It also achieves good performances on multiple age estimation datasets. The stagewise prediction structure avoids a large number of neurons and leads to a more compact model. By leveraging the dynamic range, quantization error can be better addressed so that the performance of SSR-Net can be comparable to those of bulky models. With its compact size and efficiency, SSR-Net is suitable to be employed on mobile or embedded devices for age estimation. In the future, we would like to explore the proposed ideas for other regression problems.

Acknowledgements. This work was supported by Ministry of Science and Technology (MOST) under grants 105-2221-E-001-030-MY2 and 104-2628-E-001-003-MY3, and MOST Joint Research Center for AI Technology and All Vista Healthcare under 107-2634-F-002-007.

References

[Agustsson *et al.*, 2017] Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Anchored regression networks applied to age estimation and super resolution. In *Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV)*, pages 1652–1661, 2017.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[Chang *et al.*, 2011] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 585–592, 2011.

[Chen *et al.*, 2017] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5183–5192, 2017.

[Geng *et al.*, 2014] Xin Geng, Qin Wang, and Yu Xia. Facial age estimation by adaptive label distribution learning. In *International Conference on Pattern Recognition (ICPR)*, pages 4465–4470, 2014.

[Hou *et al.*, 2017] Peng Hou, Xin Geng, Zeng-Wei Huo, and Jia-Qi Lv. Semi-supervised adaptive label distribution learning for facial age estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2015–2021, 2017.

[Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[Liu *et al.*, 2015] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. AgeNet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015.

[Malli *et al.*, 2016] Refik Can Malli, Mehmet Aygün, and Hazim Kemal Ekenel. Apparent age estimation using ensemble of deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[Niu *et al.*, 2016] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016.

[Ramanathan *et al.*, 2009] Narayanan Ramanathan, Rama Chellappa, and Soma Biswas. Age progression in human faces: A survey. *Journal of Visual Languages and Computing*, 15:3349–3361, 2009.

[Ricanek and Tesafaye, 2006] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[Rothe *et al.*, 2015] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015.

[Rothe *et al.*, 2016a] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, 126(2-4):144–157, 2016.

[Rothe *et al.*, 2016b] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Some like it hot-visual guidance for preference prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6663–5561, 2016.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[Tan *et al.*, 2018] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan ZQ Li. Efficient group-n encoding and decoding for facial age estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[Yang *et al.*, 2017] Tsun-Yi Yang, Jo-Han Hsu, Yen-Yu Lin, and Yung-Yu Chuang. DeepCD: Learning deep complementary descriptors for patch representations. In *Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV)*, pages 3314–3322, 2017.

[Zhang *et al.*, 2017] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. *Proceedings of British Machine Vision Conference (BMVC)*, 2017.