Automatic Age Estimation from Face Images via Deep Ranking

Huei-Fang Yang [2], Bo-Yao Lin [1], Kuang-Yu Chang [1], and Chu-Song Chen [1,2]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan.
[2] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.

BMVC 2015

Problem

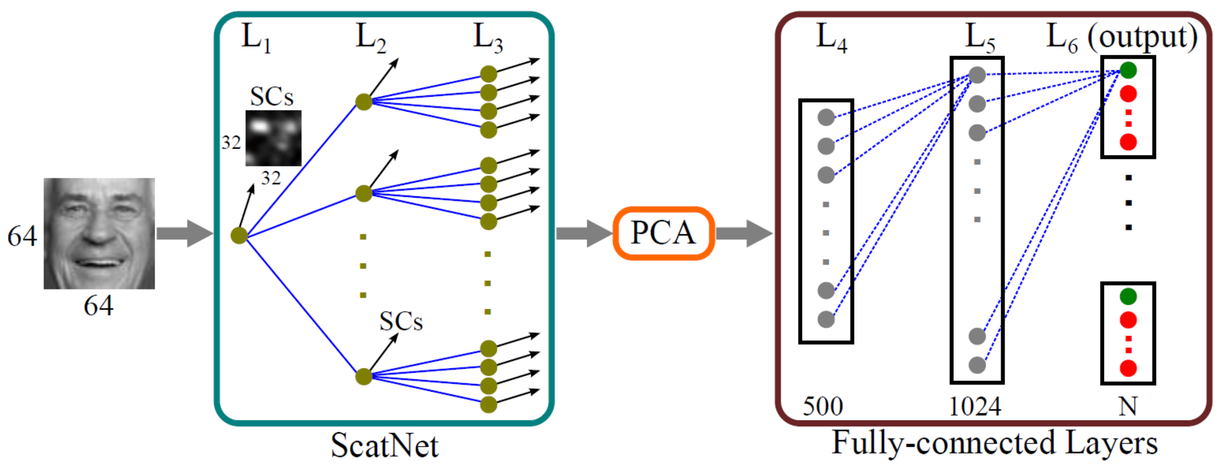
- Automatic age estimation (AAE) from face images is a challenging problem because of large facial appearance variations resulting from a number of factors, e.g., gender and facial expressions.
- We propose a generic, deep ranking model which first extracts features from the face through a scattering network (ScatNet), then reduces the feature dimension by principal component analysis, and finally predicts the age via category-wise rankers.

Contribution

1. Develop a generic, deep ranking model that combines ScatNet and the ordinal ranking for AAE from face images.
2. Our deep ranking model is general and can be applied to age estimation from faces with large facial appearance variations as a result of aging or facial expression changes.
3. We show that the high-level concepts learned from large-scale neutral faces can be transferred to estimating ages from faces under expression changes, leading to improved performance.

Deep Ranker

- Our network model for human age prediction from face images



ScatNet as Facial Descriptors

- ScatNet [1] is a deep convolutional network of specific characteristics.
  - use predefined wavelets
  - compute by a cascade of wavelet transforms and modulus pooling operators from shallow to deep layers
- ScatNet can eliminate variability from translations, rotations or scaling.
- Unlike standard CNNs, ScatNet requires no learning of the parameters.

DeepRank: An Ordinal Regression Ranker

- **Rank Encoding**: We employ the reduction framework [2] to conduct a deep ranker. Given a set of training samples $\boldsymbol{X} = \{(\boldsymbol{x_i}, \boldsymbol{y_i}), \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{N}\}$, let $\boldsymbol{x_i} \in \boldsymbol{R^D}$ be the input image and $\boldsymbol{y_i}$ be a rank label ($\boldsymbol{y_i} \in \{\mathbf{1}, \ldots, \boldsymbol{K}\}$), respectively, where $\boldsymbol{K}$ is the number of age ranks. For rank $\boldsymbol{k}$, we separate $\boldsymbol{X}$ into two subsets, $\boldsymbol{X_k^+}$ and $\boldsymbol{X_k^-}$, as follows:

$$\begin{aligned} \boldsymbol{X_k^+} &= \{(\boldsymbol{x_i}, +\mathbf{1}) | \boldsymbol{y_i} > \boldsymbol{k}\} \\ \boldsymbol{X_k^-} &= \{(\boldsymbol{x_i}, -\mathbf{1}) | \boldsymbol{y_i} \leqslant \boldsymbol{k}\}. \end{aligned} \quad (1)$$

- $\boldsymbol{K} - \mathbf{1}$ such binary classifiers are required for $\boldsymbol{K}$ age ranks.
- The output layer can be encoded in an intuitive-to-understand structure: For a face image with true age $\boldsymbol{k}$, the teaching vector with length $\boldsymbol{K} - \mathbf{1}$ is designed as $\left[\mathbf{1}, \ldots, \mathbf{1}, -\mathbf{1}, \ldots, -\mathbf{1}\right]$, where the first $\boldsymbol{k} - \mathbf{1}$ values are 1 and the remaining $-\mathbf{1}$.
- **Prediction**: Aggregation the network's binary outputs into a rank $\boldsymbol{r}(\boldsymbol{x_i})$:

$$\boldsymbol{r}(\boldsymbol{x_i}) \equiv \mathbf{1} + \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}-\mathbf{1}} [\![\boldsymbol{O_k}(\boldsymbol{x_i}) > \mathbf{0}]\!], \quad (2)$$

where $\boldsymbol{O_k}(\boldsymbol{x_i})$ is the output of the $\boldsymbol{k}$th node, and $[\![\cdot]\!]$ is 1 if the condition is met and 0 otherwise.

DeepRank+: A Multi-task Ranker

- **Category-wise Label Encoding**: Assume there are $\boldsymbol{C}$ category-wise rankers. The encoding for each ranker consists of two constituents: the category label(s) and the age rank.
- For category $\boldsymbol{j}$, its encoding is given as $\boldsymbol{e_j} = [\boldsymbol{g_j}, \boldsymbol{r_j}]$, where $\boldsymbol{g_j}$ denotes the desired output of the category and $\boldsymbol{r_j}$. Concatenating the $\boldsymbol{C}$ encoding sets forms a final encoding: $\boldsymbol{E} = \left[\boldsymbol{g_0}, \boldsymbol{r_0}, \ldots, \boldsymbol{g_j}, \boldsymbol{r_j}, \ldots, \boldsymbol{g_C}, \boldsymbol{r_C}\right]$.
- When a face image with age $\boldsymbol{k}$ belongs to category $\boldsymbol{j}$, the teaching vector is defined as:

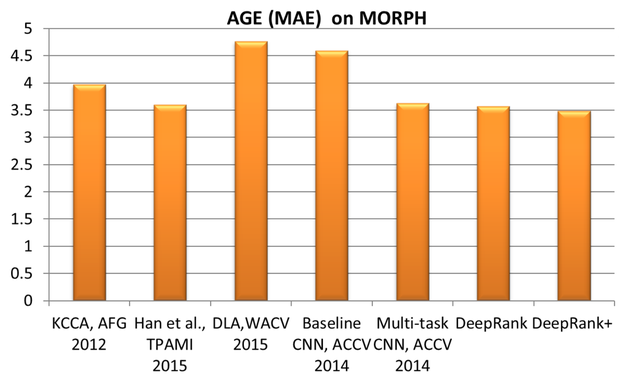
$$\left[\overbrace{\underbrace{g_1}_{-1}}, \overbrace{\underbrace{r_1}_{0}, \ldots, \underbrace{r_1}_{0}}^{K-1}, \ldots, \overbrace{\underbrace{g_j}_{1}}, \overbrace{\underbrace{r_j}_{1}, \ldots, \underbrace{r_j}_{1}, \underbrace{r_j}_{-1}, \ldots, \underbrace{r_j}_{-1}}^{K-1}, \ldots, \overbrace{\underbrace{g_C}_{-1}}, \overbrace{\underbrace{r_C}_{0}, \ldots, \underbrace{r_C}_{0}}^{K-1}\right], \quad (3)$$

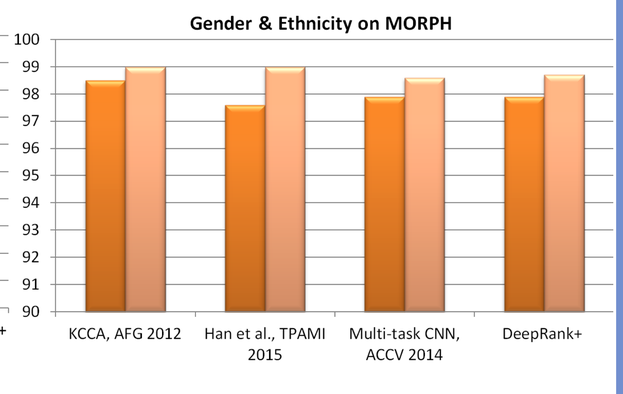
where **0**s denote "don't care".

- **Prediction**: First determine to which category the face image belongs. Then, we use Equation (2) to aggregate the ranking results of the determined category.

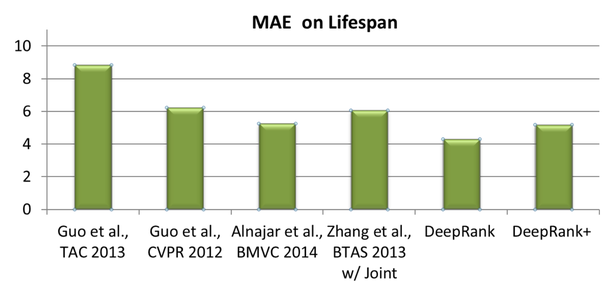
Experimental Results

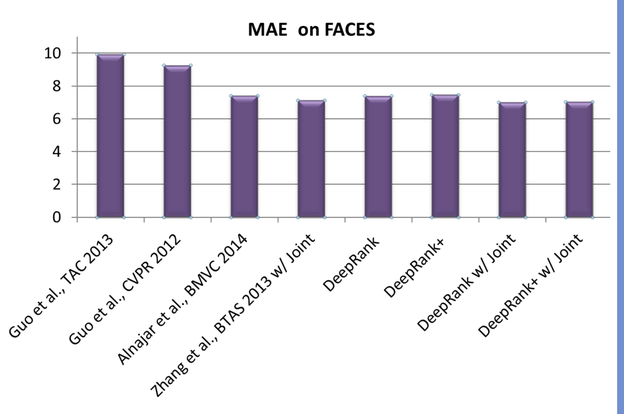
- We conducted two sets of experiments on three face datasets.
  1. Age estimation with different gender and races on a large-scale MORPH dataset.
  2. Age estimation under expression changes on the Lifespan and FACES datasets.
- Age Estimation with Different Gender and Races on MORPH
  - **MORPH** comprises more than 55,000 face images of 13,000 individuals aged from 16 to 77 years (and thus there are 62 age ranks).
  - Joint learning of age, gender, and race not only boosts the performance on age estimation but also allows our network to accurately obtain other facial attributes from a face image.
  - Network Models on MORPH: combinations of ethnicity (Black (B) and White (W)) and gender (Female (F) and Male (M)) the category labels, which results in 4 categories. Layer $\boldsymbol{L_6}$ consists of 248 (i.e., $(\mathbf{1} + \mathbf{61}) \times \mathbf{4}$) nodes.





- Age Estimation under Expression Changes on Lifespan & FACES
  - **Lifespan** contains 1,046 images of 580 individuals in various expressions (e.g., anger, annoyed, disgusted, grumpy, happy, neutral, sad, and surprise), and the ages range from 18 to 93.
  - Network Models on Lifespan: the number of nodes in $\boldsymbol{L_6}$ of DeepRank is 75, and the number of nodes in $\boldsymbol{L_6}$ of DeepRank+ is set to 152 (i.e., $(\mathbf{1} + \mathbf{75}) \times \mathbf{2}$).
  - **FACES** contains 2,052 images of 171 subjects ranging from ages 19 to 80 in two sets of six facial expressions (neutrality, sadness, happiness, disgust, fear, and anger).
  - Network Models on FACES: DeepRank has 61 nodes in $\boldsymbol{L_6}$ (62 age groups), and DeepRank+ has 372 nodes because 62 (1 category and 61 rank labels) nodes are required for each category-wise ranker and there are 6 expressions.
  - Network Models on FACES with Extended Data: DeepRank has 75 nodes in $\boldsymbol{L_6}$ (76 age groups) with additional data from Lifespan, and DeepRank+ has 456 (i.e., (1 (category) + 75 (ranks)) $\times$ 6 (expressions)) nodes.





[1] Joan Bruna and Stéphane Mallat, Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[2] H.-T. Lin and L. Li, Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 2012.