Facial Age Estimation by Nonlinear Aging Pattern Subspace

Xin Geng
School of Engineering and Information Technology, Deakin University, Victoria 3125, Australia
xgeng@deakin.edu.au

Kate Smith-Miles
School of Engineering and Information Technology, Deakin University, Victoria 3125, Australia
katesm@deakin.edu.au

Zhi-Hua Zhou
National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093, China
zhouzh@nju.edu.cn

ABSTRACT

Human age estimation by face images is an interesting yet challenging research topic emerging in recent years. This paper extends our previous work on facial age estimation (a linear method named AGES). In order to match the nonlinear nature of the human aging progress, a new algorithm named KAGES is proposed based on a nonlinear subspace trained on the aging patterns, which are defined as sequences of individual face images sorted in time order. Both the training and test (age estimation) processes of KAGES rely on a probabilistic model of KPCA. In the experimental results, the performance of KAGES is not only better than all the compared algorithms, but also better than the human observers in age estimation. The results are sensitive to parameter choice however, and future research challenges are identified.

Categories and Subject Descriptors: I.2.10 [Computing Methodologies]: Artificial Intelligence — *Vision and Scene Understanding*

General Terms: Algorithms.

1. INTRODUCTION

Facial age estimation is a relatively new research topic in the area of facial image analysis. Compared with other facial information such as identity, expression and gender, estimation of age exhibits certain difficulties. Even a human observer can rarely guess the exact age of a face image. The reason is that the facial aging progress could be affected by various factors, including not only the human gene, but also many external factors, such as health, lifestyle, weather conditions, etc. This makes age estimation a unique and challenging problem.

Despite the fact that automatic facial age estimation has vast potential applications, in reality not much work has been done up to the present. The first true age estimation algorithm was proposed by Lanitis et al. [7] [6] based on function regression, where the aging pattern was represented by a quadratic function called aging function. Our latest work on automatic facial age estimation is the AGES algorithm [4] [3]. The basic idea of AGES is to model the aging pattern, which is defined as a sequence of a particular individual's face images sorted in time order, by constructing a representative linear subspace. The proper aging pattern for a previously unseen face image is determined by the projection in the subspace that can reconstruct the face image with minimum reconstruction error, while the position of the face image in that aging pattern will then indicate its age. One weakness of AGES is that it is based on a *linear* subspace, while the aging patterns intuitively distribute on a *nonlinear* manifold. To deal with this issue, we propose in this paper a novel facial age estimation method based on nonlinear aging pattern subspace, which is named as KAGES (Kernel AGing pattErn Subspace).

As the name implies, in KAGES, the nonlinear subspace is constructed implicitly using the kernel techniques. However, KAGES is not just 'kernelized AGES' analogous to the transformation from PCA [5] to KPCA [10] because of the following two reasons: First, there are massive missing values in the aging pattern vectors; Second, the subspace is constructed for image sequences (aging patterns), but the final estimate is based on a single input image. These problems are solved in KAGES through two different optimizations of a probabilistic model of KPCA. Especially, the age estimation phase of KAGES is very different from that of AGES due to the consideration in computational complexity.

The rest of this paper is organized as follows: First, the AGES algorithm is briefly introduced in Section 2. Then, KAGES is proposed in Section 3. Experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. THE AGES ALGORITHM

The basis of AGES is a data format called *aging pattern*. An aging pattern is a sequence of personal face images sorted in time order. Along the time axis, each age is allocated one position. If face images are available for certain ages, they are filled into the corresponding positions. Otherwise, the positions are left blank. The face images are transformed into feature vectors through a feature extractor called Appearance Model [2], and then concatenated into a long vector with missing values corresponding to the missing images. In

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

reality, since it is very difficult to obtain the 'full' aging face sequence from one individual, the available aging pattern vectors for training are always highly incomplete.

In the training phase, in order to construct a representative subspace for the highly incomplete aging pattern vectors, PCA is iteratively applied. First, the missing parts in the aging pattern vectors are initialized by mean values. Then standard PCA is applied to the initialized data set to get the first subspace. After that, the subspace is used to reconstruct the missing parts through solving a linear system. Then, PCA is applied again to the updated aging pattern vectors. This procedure repeats until the reconstruction error is lower than a predefined threshold.

During age estimation, the algorithm needs to give a single estimate to a single input image. In order to achieve this, the test image is placed at every possible position (each position corresponds to an age) in the aging pattern. At each position, the test image is reconstructed by the aging pattern subspace obtained in the training phase. The minimum reconstruction error then indicates the proper position for the test image, and hence the age of it can be determined.

3. KERNEL AGING PATTERN SUBSPACE (KAGES)

KPCA is an elegant nonlinear extension of PCA through the 'kernel trick'. The basic methodology is to map the input vector $\boldsymbol{x}$ to a higher dimensional feature space via a nonlinear mapping $\Psi(\boldsymbol{x}) : \mathcal{R}^d \rightarrow \mathcal{R}^p$, where p is higher than d and could be infinite. The dot product in the feature space can be implicitly calculated by kernel functions satisfying Mercer's Theory: $\Psi(\boldsymbol{x}_i) \cdot \Psi(\boldsymbol{x}_j) = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Since applying PCA in the feature space only relies on dot products, all calculation can be done in the input space via kernel function instead of explicitly in the high dimensional feature space.

Unfortunately, KPCA could not substitute PCA in AGES. The implementation of AGES relies on applying PCA *bidirectionally*: one direction is to project the original vector from the input space to the subspace, the other is to reconstruct the vector in the input space from the subspace. However, standard KPCA could not be directly used to reconstruct the original vector in the input space. This poses challenges to both the training and test phase of KAGES.

3.1 Training Phase of KAGES

Recently, an alternative interpretation of KPCA was proposed [9] based on Probabilistic PCA (PPCA) [11] to deal with missing data. Suppose the mappings in the feature space $\boldsymbol{y} = \Psi(\boldsymbol{x})$ are centered[1]. In PPCA, the relationship between $\boldsymbol{y}$ and a latent variable $\boldsymbol{z}$ is assumed to be

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{z} + \epsilon, \tag{1}$$

where $\mathbf{W}$ is a $p \times q$ matrix (p is the dimension of $\boldsymbol{y}$ and q is that of $\boldsymbol{z}$, $p > q$) and ϵ is an error term. Assume the distribution of ϵ is an isotropic Gaussian model: $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Then $\boldsymbol{y}$ is also Gaussian distributed:

$$\boldsymbol{y} \sim N(\mathbf{W}\boldsymbol{z}, \sigma^2\mathbf{I}). \tag{2}$$

Instead of placing a Gaussian prior over $\boldsymbol{z}$ as in PPCA, in [9], a Gaussian prior is put on each element of $\mathbf{W}$: $w_{ij} \sim N(0, 1)$.

[1]Centering of $\boldsymbol{y}$ in the feature space can be done implicitly through the kernel function, see details in [10].

Table 1: Pseudocode of the KAGES Algorithm

a) ***Training Phase***

Initialize the missing parts in the training aging pattern vectors by mean values;
while not converged
  Calculate the kernel matrix $\mathbf{K}$;
  Calculate $\mathbf{Z}$ by Eq. (4);
  Calculate $\mathbf{C} = \mathbf{Z}\mathbf{Z}^\mathbf{T} + \sigma^2\mathbf{I}$;
  Minimize Eq. (5) with respect to the missing values;
  Update the missing values with the optimum;
end

b) ***Age Estimation Phase***

Initialize $\boldsymbol{\omega}$ by random values;
Generate candidate aging pattern vector by Eq. (6);
Minimize Eq. (11) with respect to $\boldsymbol{\omega}$ subject to the condition $0 \leq \omega_i \leq 1$;
The position of the maximum element in the optimum $\boldsymbol{\omega}^*$ indicates the age of the test face;

Then the likelihood of Eq. (2) can be marginalized with respect to $\mathbf{W}$:

$$\boldsymbol{y}^{(j)} \sim N(\mathbf{0}, \mathbf{Z}\mathbf{Z}^\mathbf{T} + \sigma^2\mathbf{I}), \tag{3}$$

where $\boldsymbol{y}^{(j)}$ is the j-th column of the matrix $\mathbf{Y}$ with each row corresponding to one feature vector $\boldsymbol{y}_i$ in the training set, and $\mathbf{Z}$ is the matrix with each row corresponding to one latent vector $\boldsymbol{z}_i$. It was proved that the likelihood of Eq. (3) is maximized when

$$\mathbf{Z} = \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{4}$$

where the columns of $\mathbf{U}$ are the principal eigenvectors of the kernel matrix $\mathbf{K}$ with element $k_{ij} = \boldsymbol{y}_i \cdot \boldsymbol{y}_j = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\mathbf{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues of $\mathbf{K}$, and $\mathbf{R}$ is an arbitrary rotation matrix in the latent space. Considering $\mathbf{K}$ is also the empirical covariance matrix of $\boldsymbol{y}^{(j)}$, the covariance matrix in Eq. (3) $\mathbf{C} = \mathbf{Z}\mathbf{Z}^\mathbf{T} + \sigma^2\mathbf{I}$ can be viewed as an approximation of $\mathbf{K}$. Thus the *cross entropy* between the two Gaussian distributions can be used as an objective function

$$\mathcal{H}(N(\mathbf{0}, \mathbf{K}), N(\mathbf{0}, \mathbf{C})) = -\frac{1}{2}(\log|\mathbf{C}| + \mathrm{trace}(\mathbf{K}\mathbf{C}^{-1}). \tag{5}$$

The vector $\boldsymbol{x}_i$ in the input space can be reconstructed through minimizing the above objective function without explicit calculation in the high dimensional feature space. The optimization algorithm used in this paper to minimize Eq. (5) is Scaled Conjugate Gradient (SCG) [8]. The pseudocode of the training phase of KAGES is shown in Table 1 a).

Therefore, by placing Gaussian prior distributions to $\mathbf{W}$ and minimizing the cross entropy between the kernel matrix $\mathbf{K}$ and $\mathbf{C}$, KPCA can also be applied bidirectionally, i.e., finding the low dimensional latent vector $\boldsymbol{z}$ of the input vector $\boldsymbol{x}$, as well as reconstructing $\boldsymbol{x}$ from $\boldsymbol{z}$. Consequently, this new probabilistic interpretation of KPCA can replace the linear PCA in the training phase of AGES to learn a nonlinear subspace for the incomplete aging pattern vectors.

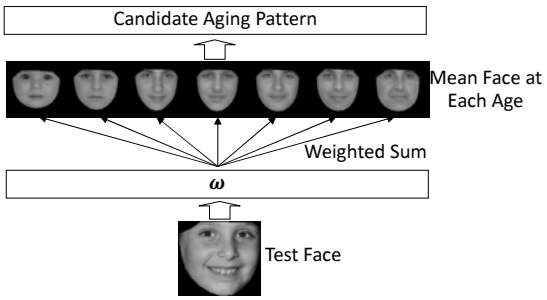

Figure 1: Generation of candidate aging pattern.

3.2 Age Estimation Phase of KAGES

It seems that the KPCA approach used in the training phase of KAGES can also be applied to the age estimation phase: First, place the test image at every possible positions in the aging pattern vector; Then, for each position, reconstruct the test image by minimizing the objective function in Eq. (5); Finally, find the suitable position with minimum reconstruction error. However, this means n times of optimization of Eq. (5) for each test image, where n is the number of ages considered in the aging pattern. Each optimization will cost approximately the same amount of time as the whole training phase described in Section 3.1. Such an approach will be computationally infeasible for most applications. In order to solve this problem, in KAGES we propose an age estimation method completely different from that in AGES.

Suppose the extracted feature vector of the test image is $\boldsymbol{b}_t$ (column vector), the mean feature vector for age $i(i = 1 \ldots n)$ over the training set is $\bar{\boldsymbol{b}}_i$, then the *candidate aging pattern vector* is composed by

$$\boldsymbol{x}_c(\boldsymbol{\omega}) = \left[\boldsymbol{b}_1^{\mathrm{T}}, \boldsymbol{b}_2^{\mathrm{T}}, \ldots, \boldsymbol{b}_n^{\mathrm{T}}\right], \quad (6)$$

$$\boldsymbol{b}_i = \omega_i \boldsymbol{b}_t + (1-\omega_i)\bar{\boldsymbol{b}}_i, i = 1 \ldots n, \quad (7)$$

where $0 \le \omega_i \le 1$. An illustration of the generation of the candidate aging pattern is shown in Fig. 1. The fraction in the candidate aging pattern vector corresponding to age i is a weighted sum of $\boldsymbol{b}_t$ and $\bar{\boldsymbol{b}}_i$ with the weight ω_i. This can be viewed as a combination of personality and commonality of the aging progress. Higher ω_i means more personality brought by the test image $\boldsymbol{b}_t$, while lower ω_i means more commonality is learned from the training set. Consequently, ω_i can be viewed as a indicator of suitability of placing $\boldsymbol{b}_t$ at the position for age i. Higher ω_i means more suitable.

In PPCA, the distribution of the centered feature vector $\boldsymbol{y}$ is $N(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^2\mathbf{I}$. The maximum likelihood can be achieved by minimizing the Mahalanobis distance from the high dimensional mapping of the candidate aging pattern vector $\boldsymbol{y}_c = \Psi(\boldsymbol{x}_c)$ to the mean (for centered $\boldsymbol{y}$, the mean is $\mathbf{0}$):

$$\begin{aligned}\boldsymbol{y}_c^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_c &= \boldsymbol{y}_c^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^2\mathbf{I})^{-1}\boldsymbol{y}_c \\ &= \sum_{i=1}^{q}(\lambda_i^{-1} - \sigma^{-2})(\boldsymbol{y}_c \cdot \boldsymbol{u}_i)^2 + \sigma^{-2}\|\boldsymbol{y}_c\|^2, \quad (8)\end{aligned}$$

where q is the dimension of the subspace, $\boldsymbol{u}_i$ are the principal eigenvectors of the covariance matrix, and λ_i are the corresponding eigenvalues. In KPCA, $\boldsymbol{u}_i$ is calculated as a linear combination of the training vectors

$$\boldsymbol{u}_i = \sum_{j=1}^{N}\alpha_j^i\boldsymbol{y}_j = \sum_{j=1}^{N}\alpha_j^i\Psi(\boldsymbol{x}_j), \quad (9)$$

$$\boldsymbol{\alpha}^i = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{v}_i, \quad (10)$$

where N is the number of training vectors, and $\boldsymbol{v}_i$ is the i-th eigenvector of the kernel matrix $\mathbf{K}$. Substituting Eq. (9) into Eq. (8) leads to the objective function

$$\begin{aligned}\mathcal{L}(\boldsymbol{\omega}) &= \sum_{i=1}^{q}(\lambda_i^{-1} - \sigma^{-2})\left(\sum_{j=1}^{N}\alpha_j^i\Psi(\boldsymbol{x}_c(\boldsymbol{\omega}))\cdot\Psi(\boldsymbol{x}_j)\right)^2 \\ &\quad +\sigma^{-2}\Psi(\boldsymbol{x}_c(\boldsymbol{\omega}))\cdot\Psi(\boldsymbol{x}_c(\boldsymbol{\omega})) \\ &= \sum_{i=1}^{q}(\lambda_i^{-1} - \sigma^{-2})\left(\sum_{j=1}^{N}\alpha_j^i\kappa(\boldsymbol{x}_c(\boldsymbol{\omega}), \boldsymbol{x}_j)\right)^2 \\ &\quad +\sigma^{-2}\kappa(\boldsymbol{x}_c(\boldsymbol{\omega}), \boldsymbol{x}_c(\boldsymbol{\omega})). \quad (11)\end{aligned}$$

An indicator vector $\boldsymbol{\omega}^*$ can be obtained through minimizing Eq. (11) with respect to $\boldsymbol{\omega}$ subject to the condition that $0 \le \omega_i \le 1$. The constrained optimization algorithm used in this paper to minimize Eq. (11) is the interior-reflective Newton method described in [1]. The maximum element in $\boldsymbol{\omega}^*$ will then indicate the suitable age for the test image. The pseudocode of the age estimation phase of KAGES is also shown in Table 1 b).

Figure 2: Typical aging face sequences in the FG-NET Aging Database.

4. EXPERIMENT

The FG-NET Aging Database [7] is used in the experiment. There are 1,002 face images from 82 subjects in this database. Each subject has 6-18 face images at different ages. The ages are distributed in a wide range from 0 to 69. Besides age variation, most of the age-progressive image sequences display other types of facial variations, such as significant changes in 3D pose, illumination, expression, *etc.* Some typical aging face sequences are shown in Figure 2.

In this experiment, KAGES is compared with AGES [4] [3], WAS [7], AAS [6], as well as some conventional classification methods including k-Nearest Neighbors (kNN), Back Propagation neural network (BP), C4.5 decision tree (C4.5), and Support Vector Machines (SVM). The algorithms are tested through the Leave-One-Person-Out (LOPO) mode, *i.e.*, in each fold, the images of one person are used as the test set and those of the others are used as the training set. After 82 folds, each subject has been used as test set once, and the final results are calculated based on all the estimates. In this way, the algorithms are tested in the case similar to real applications, *i.e.*, the subject for whom the

Table 2: Mean Absolute Error (in Years) of Age Estimation on the FG-NET Aging Database

Method	KAGES	AGES	WAS	AAS	kNN	BP	C4.5	SVM	HumanA	HumanB
MAE	**6.18**	**6.77**	**8.06**	14.83	8.24	11.85	9.34	**7.25**	8.13	6.23

algorithms attempt to estimate his/her age is previously unseen in the training set.

The face feature extractor used in the experiments is the Appearance Model [2]. The extracted feature requires 200 model parameters to retain about 95% of the variability in the training data. The dimension of the aging pattern subspace is set to 20. The kernel used in KAGES is the RBF kernel with the inverse width of 0.075. In AAS, the error threshold in the appearance cluster training step is set to 3, and the age ranges for the age specific classification are set as 0-9, 10-19, 20-39 and 40-69. The k in kNN is set to 30. The BP neural network has a single hidden layer of 100 neurons and the same number of output neurons as the number of classes. The parameters of C4.5 are set to the default values of the J4.8 implementation [12]. SVM uses the RBF kernel with the inverse width of 1.

As an important baseline, the human ability in age perception is also tested. 51 face images are randomly selected and presented to 29 human observers. There are two stages in the experiment. In each stage, the 51 face images are randomly shown to the observers, and the observers are asked to choose an age from 0 to 69 for each image. The difference is that in the first stage (HumanA), only the gray-scale face regions are shown, while in the second stage (HumanB), the whole color images are shown. HumanA intends to test age estimation purely based on face, while HumanB intends to test age estimation based on multiple cues including face, hair, skin color, clothes, and background. Note that the information provided in HumanA is the same as that provided to the algorithms, while in HumanB, additional information is available to the observers.

The age estimation performance is evaluated by the Mean Absolute Errors (MAE), i.e., the average absolute difference between the estimated age and the real age. The MAE of the all the compared methods, including the human tests, on the FG-NET Aging Database are tabulated in Table 2. The algorithms performing better than HumanA are highlighted in boldface and those better than HumanB are underlined. As can be seen, both KAGES and AGES are significantly better than all the other algorithms as well as HumanA. By extending AGES from a linear method to a nonlinear one, KAGES achieves around 9% lower MAE than AGES. It is interesting to note that the MAE of KAGES is even lower than that of HumanB, where the human observers are provided with more information than that input into the algorithms. Thus at least under this experimental setting, KAGES performs better than humans in the ability of facial age estimation. It should be noted that we experienced some sensitivity to the initial conditions of the optimization algorithm, and the best results have been reported for KAGES, just like they have been for all the other algorithms reported in Table 2.

5. CONCLUSION

This paper extends our previous work on facial age estimation. The linear AGES algorithm [4] [3] is transformed into a nonlinear algorithm named KAGES. The transformation is not a simple analog of the kernelization of PCA. Both the training and test procedures of KAGES rely on a probabilistic model of KPCA. Especially, the age estimation phase of KAGES is quite different from AGES. Experimental results show that KAGES can achieve better performance than AGES as well as six other compared algorithms. Most interestingly, KAGES even exceeds the human ability in facial age estimation. The sensitivity to the initial conditions of the optimization algorithm however resulted in inconsistent results, and while we have demonstrated the potential of the nonlinear subspace approach, there are some future research challenges that will need to be addressed. The objective function Eq. (11) could be minimized using global search methods to avoid local minima and sensitivity to initial conditions, although the computational expense of doing so may become prohibitive. In addition, different kernels and parameter settings could be explored to arrive at more robust setting.

6. REFERENCES

[1] T. F. Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, 1996.

[2] G. J. Edwards, A. Lanitis, and C. J. Cootes. Statistical face models: Improving specificity. *Image Vision Comput.*, 16(3):203–211, 1998.

[3] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12):2234–2240, 2007.

[4] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *Proc. the 14th ACM Int'l Conf. Multimedia*, pages 307–316, Santa Barbara, CA, 2006.

[5] I. T. Jolliffe. *Principal Component Analysis, 2nd Edition*. Springer-Verlag, New York, 2002.

[6] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. Systems, Man, and Cybernetics - Part B: Cybernetics*, 34(1):621–628, 2004.

[7] A. Lanitis, C. J. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(4):442–455, 2002.

[8] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

[9] G. Sanguinetti and N. D. Lawrence. Missing data in kernel pca. In *Proc. Euro. Conf. Machine Learning*, pages 751–758, Berlin, Germany, 2006.

[10] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[11] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61:611–622, 1999.

[12] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools With Java Implementations*. Morgan Kaufmann, San Francisco, CA, 1999.