Automatic Age Estimation Based on Facial Aging Patterns

Xin Geng, Zhi-Hua Zhou, *Senior Member, IEEE*,
Kate Smith-Miles, *Senior Member, IEEE*

Abstract

While recognition of most facial variations, such as identity, expression and gender, has been extensively studied, automatic age estimation has rarely been explored. In contrast to other facial variations, aging variation presents several unique characteristics which make age estimation a challenging task. This paper proposes an automatic age estimation method named AGES (AGing pattErn Subspace). The basic idea is to model the aging pattern, which is defined as the sequence of a particular individual's face images sorted in time order, by constructing a representative 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. In the experiments, AGES and its variants are compared with the limited existing age estimation methods (WAS and AAS) and some well-established classification methods (kNN, BP, C4.5, and SVM). Moreover, a comparison with human perception ability on age is conducted. It is interesting to note that the performance of AGES is not only significantly better than that of all the other algorithms, but also comparable to that of the human observers.

Index Terms

Computer vision, pattern recognition, machine learning, face and gesture recognition, age estimation.

Manuscript received January 27, 2007 ; revised June 18, 2007. Part of this work was supported by NSFC (60635030,60325207) and FANEDD (200343).

Xin Geng and Kate Smith-Miles are with the School of Engineering and Information Technology, Deakin University, VIC 3125, Australia (e-mail: {xge, katesm}@deakin.edu.au).

Zhi-Hua Zhou is with the National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: zhouzh@nju.edu.cn).

Automatic Age Estimation Based on Facial Aging Patterns

I. Introduction

People can effortlessly extract many kinds of useful information from a face image, such as identity, gender, expression, approximate age, *etc*. The automatic extraction of most of the information has been extensively studied. However, so far there is relatively little work concerning automatic age estimation, despite the fact that it is alone an interesting problem, as well as a challenging subproblem in tasks like face recognition [29]. People's behavior and preference are different at different ages [1], indicating vast potential applications of automatic age estimation.

Compared with other facial variations, aging effects display three unique characteristics:

1) *The aging progress is uncontrollable.* No one can advance or delay aging at will. The procedure of aging is slow and irreversible.
2) *Personalized aging patterns.* Different people age in different ways. The aging pattern of each person is determined by his/her genes as well as many external factors, such as health, lifestyle, weather conditions, *etc*.
3) *The aging patterns are temporal data.* The aging progress must obey the order of time. The face status at a particular age will affect all older faces, but will not affect those younger ones.

Each of these characteristics contributes to the difficulties of automatic age estimation. First, because people cannot freely control aging variation, the collection of sufficient training data for age estimation is extremely laborious. This difficulty is now partly alleviated due to the public dissemination of the FG-NET Aging Database [24]. Nevertheless, each subject in this database only has face images at a few ages, *i.e.*, the data set is highly incomplete in the view of aging patterns. Fortunately, a 'complete' aging face database is *unnecessary* since human beings also learn to perceive facial ages from incomplete aging patterns. Thus the learning algorithm applied to the aging patterns must be able to handle highly incomplete data. Second, the mapping from the instances (face images) to the class labels (ages) is not unique, but complicatedly depends on personalized factors. Thus the selection of a suitable aging pattern for a particular face becomes

a crucial step in age estimation. Third, the set of class labels (ages) is a totally ordered set. Each age has a unique rank in the time sequence. Once the suitable aging pattern for a particular face image is selected, the 'position' of the face in that aging pattern uniquely determines its age. Consistent to the last two characteristics, automatic age estimation should involve at least two main steps: Step 1 is to determine the suitable aging pattern for a particular face; Step 2 is to find the position of the face in that aging pattern.

This paper proposes a subspace approach named AGES (AGing pattErn Subspace) for automatic age estimation. Instead of using isolated face images as data samples, AGES regards each *aging pattern* as a sample. The basic idea is to model the aging patterns by a representative subspace. Each point in the subspace corresponds to one aging pattern. The proper aging pattern for a previously unseen face image is determined by the projection in the subspace that can best reconstruct the face image. Once the proper aging pattern is determined, the position of the face in the aging pattern will then indicate its age.

The rest of this paper is organized as follows. First, the related work is briefly reviewed in Section II. Then the concept of an aging pattern is introduced in Section III. After that, the AGES algorithm is proposed in Section IV. In Section V, the experimental results are reported. Finally in Section VI, conclusions are drawn.

II. Related Work

There are some earlier works aiming to simulate the aging effects on human faces, which is the inverse procedure of age estimation. For example, Burt and Perrett [2] simulated aging variations by superimposing typical aging changes in shape and color on face images. Later, Tiddeman *et al*. [25] extended this work by adopting a wavelet-based approach to add high frequency information to the age progressed images. O'Toole *et al*. [15] described how aging variations can be made by applying a standard facial caricaturing algorithm to the 3D models of faces. Hutton *et al*. [9] proposed a dense surface point distribution model for expressing the shape changes associated with growth and aging. Hill *et al*. [7] presented a statistical approach to age face images along the 'aging direction' in a face model space. Scandrett *et al*. [22] constructed a statistical model in which historical, familial and average growth tendencies of a peer group can be incorporated. Ramanathan and Chellappa [19] proposed a craniofacial growth model that characterizes growth related shape variations observed in human faces during young ages. Although these works

did not attempt age estimation, they did reveal some of the important facts in the relationship between age and face. Some other work tried to partly reveal the mapping from face to age. For example, Ramanathan and Chellappa [18] proposed a method for face verification across age based on a Bayesian classifier. Zana *et al*. [28] proposed a face verification algorithm in polar frequency domain which is robust against aging variation. Shi *et al*. [23] studied how effective are landmarks and their geometry-based approach for face recognition across ages. Kwon and Lobo [11] proposed an age classification method based on well controlled high-quality face images, which can classify faces into one of the three groups (babies, young adults, and senior adults). Zhou *et al*. [31] presented a boosting based algorithm for image based regression (IBR). Although the algorithm was designed for the general purpose of IBR, it can be well applied to the problem of age estimation.

The first *true* age estimation algorithm was proposed by Lanitis *et al*. [13]. In their work, the aging pattern is represented by the *aging function*: $age = f(\boldsymbol{b})$, where $\boldsymbol{b}$ is the vector of the face model parameters, and f is defined as a quadratic function. During the training process, a quadratic function is fitted for each individual in the training set as his/her aging function. To determine the suitable aging function for a previously unseen face image during age estimation, they proposed four different ways. Among the methods that do not rely on the external 'lifestyle profiles', the Weighted Appearance Specific (WAS) method achieved the best performance. Later, Lanitis *et al.* [12] compared their quadratic aging function method with several conventional classification methods in age estimation. The algorithms were tested in the single layer mode as well as three hierarchical modes. As expected, all classifiers performed better in the hierarchical modes because the hierarchical structures handle the face image clusters separately according to the age groups or the appearance or both. Among them, the Appearance and Age Specific (AAS) method achieved the best performance. However, according to the experimental results, the quadratic aging function did not show remarkable superiority over the conventional classifiers in the overall performance. The aging function based approaches regard age estimation as a conventional function regression problem without special design for the unique characteristics of aging variation. This limitation prevents them from obtaining more satisfying results. In detail, there might be four weaknesses in such approaches. First, the formula of the aging function is empirically determined. There is no evidence suggesting that the relationship between face and age is as simple as a quadratic function. Second, the *temporal*

characteristic cannot be well utilized by the aging function. The dependent relationship among the aging faces is monodirectional, *i.e.*, the status of a certain face only affects those older faces. However, the relationship revealed by the aging function is bidirectional: any changes on a particular face will change the aging function, hence affect all other faces. Third, the learning of one person's aging pattern is solely based on the face images of that person. Although people age in different ways, there must be some commonality among all aging patterns, *i.e.*, the general trend of aging. Such commonality is also crucial in age estimation, especially when the personal training data is insufficient. Fourth, the aging function for the previously unseen face image is simply a linear combination of the known aging functions, rather than being generated from a certain model of aging patterns. All of these problems can be solved, from a new point of view, by the AGES algorithm. Changes start from the very beginning: data representation.

III. AGING PATTERN

The aging function based methods regard age estimation as a conventional classification problem: the data are the face images, the target is their age labels. According to the *personalized* characteristic, each image $\mathbf{I}$ should have one more label other than its age label $age(\mathbf{I})$, *i.e.* its personal identity $id(\mathbf{I})$. If the problem is to be solved by supervised techniques like LDA (Linear Discriminant Analysis), then the algorithm must deal with the multi-label data, which is alone a problem in machine learning. On the other hand, if all of these labels can be integrated into the data representation, then the multi-label problem can be transformed into an unsupervised learning problem. Thus we propose a data representation called *Aging Pattern*, which is the basis of AGES. A formal definition is given as follows.

Definition 1. An aging pattern is a sequence of personal face images sorted in time order.

The keywords are 'personal' and 'time'. All face images in an aging pattern must come from the same person, and they must be ordered by time. Take the aging pattern shown in Fig. 1 as an example. Along the t axis, each age (0-8) is allocated one position. If face images are available for certain ages (2, 5 and 8), they are filled into the corresponding positions. If not, the positions are left blank. If all positions are filled, the aging pattern is called a *complete aging pattern*, otherwise it is called an *incomplete aging pattern*. Before the aging pattern can be further processed, the face images in it are first transformed into feature vectors. Obviously

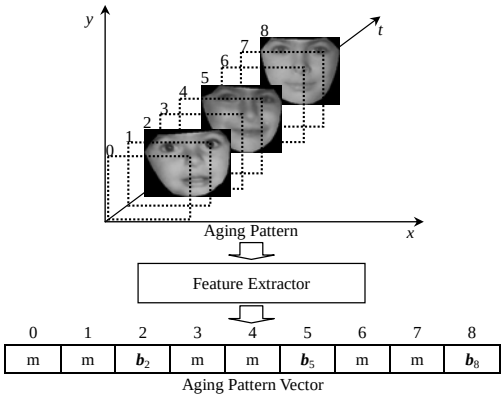


Fig. 1. Vectorization of the aging pattern. The ages (0-8) are marked at the top-left to the corresponding positions and above the corresponding feature vectors. The missing parts in the aging pattern vector are marked by 'm'.

aging is a process related to both the shape and the texture of face. Thus the Appearance Model [4] is used as the feature extractor, whose main advantage is that the extracted feature combines both the shape and the intensity of the face images. Fig. 1 gives an example of the vectorization of the aging pattern, where $\boldsymbol{b}_2$, $\boldsymbol{b}_5$ and $\boldsymbol{b}_8$ represent the feature vectors of the face images at the ages 2, 5 and 8, respectively.

By representing aging patterns in this way, the two labels $age(\mathbf{I})$ and $id(\mathbf{I})$ are naturally integrated into the data without any pre-assumptions. Each aging pattern implies one ID, each age is fixed into a position in the aging pattern, and the position is ordered according to time. Consequently, the *personalized* and *temporal* characteristics can be well utilized. As long as the aging patterns are well sampled, a proper model of aging patterns can be learned and the learning process is unsupervised. However, this brings two other challenges: (1) During training, the learning algorithm applied to the aging patterns must be able to handle highly incomplete training samples; (2) During age estimation on test data, the most suitable aging pattern as well as the most suitable position in that aging pattern must be selected for an unknown face image. The next section mainly tackles these two problems.

IV. THE AGES ALGORITHM

A. Aging Pattern Subspace

A representative model for the aging patterns can be built up by the information theory approach of coding and decoding. One widely adopted technology is using PCA [10] to construct a subspace that captures the main variation in the data set. The projection in the subspace is

computed by

$$\boldsymbol{y} = \mathbf{W}^{\mathbf{T}}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{1}$$

where $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{x}$, and $\mathbf{W}^{\mathbf{T}} = \mathbf{W}^{-\mathbf{1}}$ is the transpose of $\mathbf{W}$, which is composed by the orthogonal eigenvectors of the covariance matrix of $\boldsymbol{x}$. The difficulty is that the aging pattern vector $\boldsymbol{x}$ is highly incomplete. Based on the characteristics of aging patterns, an EM-like algorithm is proposed here to learn a representative subspace.

Suppose the training set has N aging pattern vectors $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Any sample in this set can be written as $\boldsymbol{x}_k = \{\boldsymbol{x}_k^a, \boldsymbol{x}_k^m\}$, where $\boldsymbol{x}_k^a$ are the available features and $\boldsymbol{x}_k^m$ are the missing features of $\boldsymbol{x}_k$. Suppose the transformation matrix is $\mathbf{W}$, the projection $\boldsymbol{y}_k$ of $\boldsymbol{x}_k$ in the subspace can be calculated by Eq. (1), and the reconstruction of $\boldsymbol{x}_k$ is calculated by

$$\hat{\boldsymbol{x}}_k = \boldsymbol{\mu} + \mathbf{W}\boldsymbol{y}_k. \tag{2}$$

$\hat{\boldsymbol{x}}_k$ can also be written as $\hat{\boldsymbol{x}}_k = \{\hat{\boldsymbol{x}}_k^a, \hat{\boldsymbol{x}}_k^m\}$, where $\hat{\boldsymbol{x}}_k^a$ is the reconstruction of $\boldsymbol{x}_k^a$, and $\hat{\boldsymbol{x}}_k^m$ is the reconstruction of $\boldsymbol{x}_k^m$. It is well known that standard PCA can be derived by minimizing the mean reconstruction error (residuals) of the data set D in the subspace [10]. With the presence of the missing features $\boldsymbol{x}_k^m$, the goal is changed into finding a $\mathbf{W}$ that minimizes the mean reconstruction error of the available features

$$\overline{\varepsilon}^a = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k^a - \hat{\boldsymbol{x}}_k^a)^{\mathbf{T}}(\boldsymbol{x}_k^a - \hat{\boldsymbol{x}}_k^a). \tag{3}$$

In case that the number of missing features in different instances is highly uneven, Eq. (3) should be normalized by the dimensionality of the missing part. This is equivalent to a preprocess of dividing each instance by its missing dimensionality. The FG-NET Aging database used in this paper has a similar number of missing features in each aging pattern, thus there is no significant difference observed in the experiments with/without the normalization.

When initializing, $\boldsymbol{x}_k^m$ is replaced by the mean vector $[\boldsymbol{\mu}(_k^m)]$, calculated from other samples whose corresponding features are available. Then standard PCA is applied to the full-filled data set to get the initial transformation matrix $\mathbf{W}_0$ and mean vector $\boldsymbol{\mu}_0$. In the iteration i, the projection of $\boldsymbol{x}_k$ in the subspace spanned by $\mathbf{W}_i$ is estimated first. Since there are many missing features in $\boldsymbol{x}_k$, the projection cannot be computed directly by Eq. (1). Note that the aging patterns are highly redundant, it is possible to estimate $\boldsymbol{y}_k$ only based on part of $\boldsymbol{x}_k$ [14], say $\boldsymbol{x}_k^a$. Instead

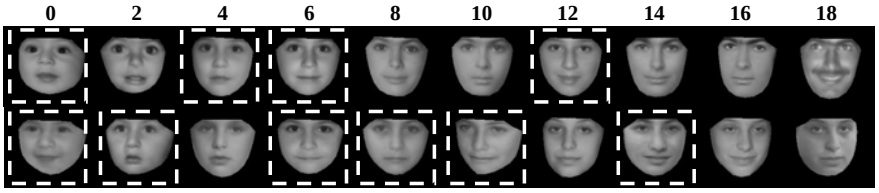


Fig. 2. The 'full-filled' aging patterns. Each line shows the aging pattern of one person. The ages are marked above the corresponding faces. The faces learned by the algorithm are surrounded by the dashed squares.

of using inner product, $\boldsymbol{y}_k$ is solved as the least squares solution of

$$[\mathbf{W}_i(^a_k)]\boldsymbol{y}_k = \boldsymbol{x}^a_k - [\boldsymbol{\mu}_i(^a_k)], \tag{4}$$

where $[\mathbf{W}_i(^a_k)]$ is the part in $\mathbf{W}_i$ and $[\boldsymbol{\mu}_i(^a_k)]$ is the part in $\boldsymbol{\mu}_i$ that correspond to the positions of $\boldsymbol{x}^a_k$. After getting the estimation of $\boldsymbol{y}_k$, $\hat{\boldsymbol{x}}_k$ is calculated by Eq. (2) and $\boldsymbol{x}^m_k$ is updated by $\hat{\boldsymbol{x}}^m_k$. Then, standard PCA is applied to the updated data set to get the new transformation matrix $\mathbf{W}_{i+1}$ and mean vector $\boldsymbol{\mu}_{i+1}$. The whole process repeats until the maximum iteration τ is exceeded or $\overline{\varepsilon}^a$ is smaller than a predefined threshold θ. The convergence of this algorithm is proved in Appendix I.

During the training process of AGES, the missing faces in the training aging patterns can be simultaneously learned by reconstructing the whole aging pattern vectors through Eq. (2). Fig. 2 shows some typical examples of the 'full-filled' aging patterns when AGES is applied to the FG-NET Aging Database [24]. For clarity, only the faces in the most changeable age range from 0 to 18 with 2 year increments are shown. Since remarkable variations other than the aging effects exist in the FG-NET Aging Database and the feature extractor does not treat them separately, some generated faces present noticeable variations in expression, pose or illumination. These variations can be dealt with, as will be discussed in Section V, by applying LDA to the Appearance Model parameters. It can be seen that the learned faces inosculate with those real faces very well in the aging patterns. Thus this learning algorithm can also be used to simulate aging effects on human faces.

The process of the learning algorithm is actually a process of interaction between the global aging pattern model and the personalized aging patterns. As mentioned in Section II, although different persons age in different ways, the commonality (modeled by the subspace) of all aging patterns is also crucial for age estimation, especially when the aging patterns are highly

incomplete. In each iteration, the missing part of the personal aging pattern is first estimated by the current global aging pattern model. Then, the global model is further refined by the updated personal aging patterns. In this way, the commonality and the personality of the aging patterns are alternately utilized to learn the final subspace.

B. Age Estimation

The aging pattern subspace is a global model for aging patterns, each of which corresponds to a sequence of age labels. But the task of age estimation is based on a single face input, and expects a single age output. This section will describe how this can be done with the aging pattern subspace.

Given a previously unseen face image $\mathbf{I}$, its feature vector $\boldsymbol{b}$ is first extracted by the feature extractor. Recall the two steps of age estimation mentioned in Section I. The first step is to find a proper aging pattern for $\mathbf{I}$. Note that each point in the subspace corresponds to one aging pattern. Thus the proper aging pattern for $\mathbf{I}$ can be selected through finding a point in the subspace that can best reconstruct $\boldsymbol{b}$, *i.e.*, minimizing the reconstruction error. However, without knowing the position of $\mathbf{I}$ in the aging pattern, which should be determined in the second step, the reconstruction error cannot be actually calculated. Thus $\mathbf{I}$ is placed at every possible position in the aging pattern, getting p aging pattern vectors $\boldsymbol{z}_j (j = 1 \dots p)$ by placing $\boldsymbol{b}$ at the position j in $\boldsymbol{z}_j$. Note that $\boldsymbol{b}$ is the only available feature in $\boldsymbol{z}_j$, the projection $\boldsymbol{y}_j$ can be estimated by Eq. (4), and the reconstruction error can be calculated by

$$\varepsilon^a(j) = (\boldsymbol{b} - \boldsymbol{\mu}_{(j)} - \mathbf{W}_{(j)}\boldsymbol{y}_j)^{\mathbf{T}}(\boldsymbol{b} - \boldsymbol{\mu}_{(j)} - \mathbf{W}_{(j)}\boldsymbol{y}_j), \tag{5}$$

where $\boldsymbol{\mu}_{(j)}$ is the part in $\boldsymbol{\mu}$ and $\mathbf{W}_{(j)}$ is the part in $\mathbf{W}$ that corresponds to the position j. Then the projection $\boldsymbol{y}_r$ that can reconstruct $\boldsymbol{b}$ with minimum reconstruction error over all the p possible positions is determined by

$$r = \arg\min_j(\varepsilon^a(j)). \tag{6}$$

Thus the suitable aging pattern for $\mathbf{I}$ is $\boldsymbol{z}_r$. Step 2 afterward becomes trivial because r also indicates the position of $\mathbf{I}$ in $\boldsymbol{z}_r$. Finally the age associated to the position r is returned as the estimated age of $\mathbf{I}$. As a byproduct of age estimation, the whole aging pattern vector can be reconstructed as $\mathbf{W}\boldsymbol{y}_r$, which can be used to simulate faces at different ages of the subject in $\mathbf{I}$.

(a) (b)

Fig. 3. Typical aging face sequences in (a) the FG-NET Aging Database and (b) the MORPH database.

During the age estimation process of AGES, the proper aging pattern for the test image is generated based on both the aging pattern subspace and the face image feature. The subspace defines the general trend of aging, and the face image feature represents the personalized factors. By placing the feature vector at different positions, candidate aging patterns specified to the test face are generated. Among these candidates, only one is consistent with the general aging trend, which can be detected via minimum reconstruction error by the aging pattern subspace. At the same time, the position of the test image in that aging pattern can be determined.

V. EXPERIMENTS

A. Methodology

The FG-NET Aging Database [24] and the MORPH database [20] are used in the experiments. The FG-NET Aging Database contains $1,002$ face images from 82 subjects. In the MORPH database, there are $1,724$ face images from 515 subjects. Each subject has around 3 aging images, which are too few for the training of AGES. Thus in the experiment, MORPH is only used to test the algorithms trained on the FG-NET database. Because all subjects in the FG-NET database are of Caucasian descent, only the 433 images from Caucasian descent in the MORPH database are used as the test set. The ages in both databases are distributed highly unevenly in wide ranges: 0-69 for FG-NET and 15-68 for MORPH. The age range distribution is tabulated in Table I. Typical aging face sequences from both databases are shown in Fig. 3. As can be seen, besides the aging variation, most aging sequences display other variations in pose, illumination, expression, occlusion, *etc*. Although these variations might be harmful, all images are used in the experiment because data insufficiency is more serious a problem.

The face feature extractor used in the experiments is the Appearance Model [4]. Sixty-eight landmark points of each face image are used to train the shape model, and $5,000$ pixels are used in the shape-normalized faces. The extracted feature requires 200 ($n = 200$) model parameters to retain about 95% of the variability in the training data. In order to deal with variations other

TABLE I

AGE RANGE DISTRIBUTION OF THE IMAGES IN THE DATABASES AND THE HUMAN OBSERVERS PARTICIPATING THE EXPERIMENT

Age Range	FG-NET (%)	MORPH (%)	Observers (%)
0-9	37.03	0	0
10-19	33.83	24.71	0
20-29	14.37	47.34	93.10
30-39	7.88	18.94	3.45
40-49	4.59	6.47	3.45
50-59	1.50	1.85	0
60-69	0.80	0.69	0

than aging in the training set, LDA can be applied to the 200-dimensional feature vectors with age labels. LDA tries to find a subspace where images of different ages scatter while those of the same age converge. In this sense, the resulted *discriminant parameters* are expected to be more related to the aging variation and hence the effect of other variations can be partially suppressed. The AGES based on such discriminant parameters is denoted by AGES$_{lda}$. In both AGES and AGES$_{lda}$, the dimension of the aging pattern subspace is set to 20 ($d = 20$), the maximum iteration $\tau = 50$, and the error threshold $\theta = 10^{-3}$.

In this experiment, AGES is compared with WAS [13], AAS [12], as well as some conventional classification methods including k-Nearest Neighbors (kNN) [16], Back Propagation neural network (BP) [21], C4.5 decision tree (C4.5) [17], and Support Vector Machines (SVM) [26]. The algorithms are first tested on the FG-NET Aging Database 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 estimations. In this way, the algorithms are tested in the case similar to real applications, *i.e.*, the subject for whom the algorithms attempt to estimate his/her age is previously unseen in the training set. In order to further test the generalization ability, the algorithms trained on the FG-NET Aging Database are then tested on the MORPH database. Note that as a test set, the possible age range of the MORPH data is assumed to be the same as that of the FG-NET Aging Database (0-69), although the actual range is much smaller (15-68).

For all the comparative algorithms, several parameter configurations are tested and the best

TABLE II

MAE OF STANDARD AGE ESTIMATION ON FG-NET AND MORPH

Method	AGES	$AGES_{lda}$	WAS	AAS	kNN	BP	C4.5	SVM	HumanA	HumanB
FG-NET (LOPO)	**6.77**	**<u>6.22</u>**	**8.06** (1, 1)	14.83 (1, 1)	8.24 (1, 1)	11.85 (1, 1)	9.34 (1, 1)	**7.25** (1, 1)	8.13 (1, 1)	6.23 (-1, 0)
MORPH (Test Set)	8.83	**8.07**	9.32 (1, 1)	20.93 (1, 1)	11.30 (1, 1)	13.84 (1, 1)	12.69 (1, 1)	9.23 (1, 1)	— —	— —

result is reported. For 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 architecture of 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 [27]. SVM follows the 1-against-1 method [8] for multi-class classification and uses the RBF kernel with the bias 1.

As an important baseline, the human ability in age perception is also tested. From each age range listed in Table I, 5% of the face images are randomly selected from the FG-NET Aging Database. In total, 51 face images are selected and presented to 29 human observers (24 males and 5 females). The age range distribution of the observers themselves is listed in the 4th column of Table I. None of them received training on the task before the experiment. 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.

B. Standard Age Estimation

First, the algorithms and the human ability are evaluated by the criterion used in [12] and [13], *i.e.*, the Mean Absolute Error (MAE), which is tabulated in Table II. The algorithms performing better than HumanA are highlighted in boldface and those better than HumanB are underlined. The results of the pairwise one-tailed t-test at the significance level 0.025 are listed in the parentheses under the corresponding MAEs. The first number is the t-test result of AGES paired

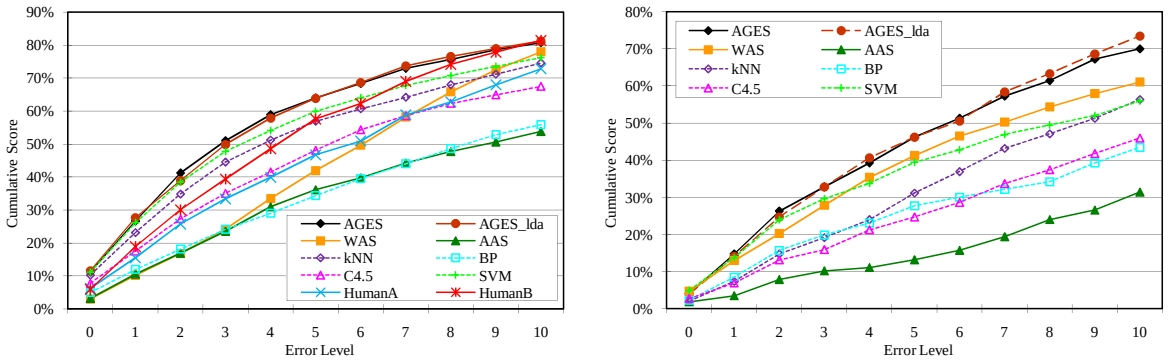
with other algorithms, the second number is that of AGES$_{lda}$ paired with other algorithms. 1, 0 and -1 represent significantly better, not significantly different, and significantly worse, respectively. Note that the t-tests related to the human test are performed only on those images used in the human test. When tested on the FG-NET Aging Database (by the LOPO mode), Both AGES$_{lda}$ and AGES are significantly better than all the other algorithms and HumanA. Although AGES is significantly worse than HumanB, AGES$_{lda}$ is not significantly different with HumanB, which indicates the effectiveness of using LDA to extract the aging-related features. It is worth mentioning that even lower MAE (5.81) on the FG-NET Aging Database was reported in [31]. However, the test mode in [31] was different from the LOPO mode, where the whole data set was randomly divided into 800 training images and 202 test images. This brought the advantage that the aging pattern of the test image would be included in the training set. In order to verify this, we also test AGES in such mode and get a lower MAE of 5.27. When the algorithms are trained on the FG-NET Aging Database but tested on the MORPH database, all of them get higher MAE as expected. This time, only AGES$_{lda}$ can get lower MAE than that of HumanA. However, the relative performance of the algorithms is similar as that on the FG-NET Aging Database. Recall that in HumanA, the observers are provided with the same information as that fed into the algorithms. So the comparison between the algorithms and HumanA is more meaningful. But can the conclusion be drawn from the results on the FG-NET Aging Database that all of the best 4 algorithms (AGES$_{lda}$, AGES, WAS, and SVM) perform better than the human observers? Perhaps not.

MAE is only an indicator of the average performance of the age estimators. It does not provide enough information on how accurate the estimators might be. Suppose there are M test images, $M_{e\leq l}$ is the number of test images on which the age estimation makes an absolute error no higher than l (years), then the *cumulative score* at error level l is calculated by

$$CumScore(l) = M_{e\leq l}/M \times 100\%. \tag{7}$$

If the 'correct estimation' is defined as the estimation with an absolute error no higher than l, then $CumScore(l)$ is actually the accuracy rate. Thus the cumulative score can be viewed as an indicator of the accuracy of the age estimators. Since the acceptable error level is unlikely to be very high, the cumulative scores at lower error levels are more important.

The cumulative scores of the algorithms and human observers at the error levels from 0 to 10



(a) (b)

Fig. 4. Cumulative scores of standard age estimation on (a) FG-NET (LOPO) and (b) MORPH (Test Set).

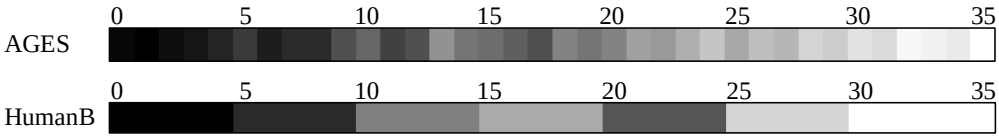


Fig. 5. σ values of AGES and HumanB on the FG-NET Aging Database when the real age is from 0 to 35.

(years) are compared in Fig. 4. The situation at higher error levels is not shown because in general, age estimation with an absolute error higher than 10 (a decade) is not acceptable. Fig. 4(a) reveals that the result of AGES$_{lda}$ on FG-NET is very similar to that of AGES. AGES/AGES$_{lda}$ is the most accurate age estimator at almost all error levels. This is impressive since more information is provided in HumanB than that fed into AGES. All of the best 4 algorithms (AGES$_{lda}$, AGES, SVM, and kNN) perform better than HumanA at all error levels. Although the MAE of WAS is lower than that of HumanA, its cumulative scores are worse than those of HumanA in most cases, especially at the relatively important low error levels. The situation on MORPH (Fig. 4(b)) is similar to Fig. 4(a) with the exception that WAS performs remarkably better relative to other algorithms, which indicates good generalization ability of WAS. But its cumulative scores at all error levels are still worse than both AGES and AGES$_{lda}$.

C. *Imbalanced Age Estimation*

It is a common sense that the changing rates of facial appearance at different aging stages are different. Usually young faces change faster than older ones. Consequently, age estimation is more error-vulnerable at older ages. This phenomenon is called *Imbalanced Age Estimation*. Since AGES is based on the aging patterns rather than isolated face images, the imbalanced

TABLE III

MAE OF AGES AND AGES$_r$ IN DIFFERENT AGE RANGES

Method	0-5	6-30	31-69	All Ages
AGES	1.87	4.88	24.97	6.77
AGES$_r$	1.17	4.48	7.93	4.15 (6.52)

appearance changing rate in the aging patterns can be naturally learned. Suppose when the real age is α, the estimated age $\hat{\alpha}$ follows a normal distribution centered at α, *i.e.*, $\hat{\alpha} \sim N(\alpha, \sigma^2)$. The standard deviation σ can be estimated from $\hat{\alpha}$. After MIN-MAX normalization and histogram equalization, the σ values of AGES and the human observers (HumanB) on the FG-NET Aging Database are shown as gray-scale images in Fig. 5. The darker the intensity, the more accurate the estimation. Since the images shown to the observers do not include all ages, an average value of σ is calculated for each age range rather than exact age. It can be seen that the imbalanced age estimation phenomenon of AGES is similar to that of the human observers. One exception is that in the age range 20-24, the σ value of the human observers is relatively lower. The reason might be that most observers themselves are within this age range (refer to Table I), thus they are most familiar with faces in the range. For AGES, there are two remarkable jumps of σ with the increase of age. The first is at 5 (σ increases by 139%), and the second is at 30 (σ increases by 74%). These two points divide all ages into three groups with different aging rates: infant to child (0-5, fastest), child to middle-age (5-30, fast), and middle-age to senior (≥ 30, slow), which are roughly the three main stages in human facial aging.

D. Age Range Based Estimation

The above approaches might encounter a problem that the resulted model would be skewed toward the age range that has more instances in the training set. In order to verify this, the ages (0-69) in the FG-NET Aging Databse are divided into 3 age ranges as 0-5, 6-30, 31-69, which are consistent with the age groups found by analyzing the changes of σ in Section V-C. The MAEs of AGES in different age ranges are shown in the first line of Table III. As can be seen, the MAE in 31-69 is much higher than others due to the fact that the training samples in that range are insufficient (refer to Table I).

One way to solve this problem is to build separate subspaces for different age groups. The result of applying AGES separately to each age group (denoted by AGES$_r$) is given in the second

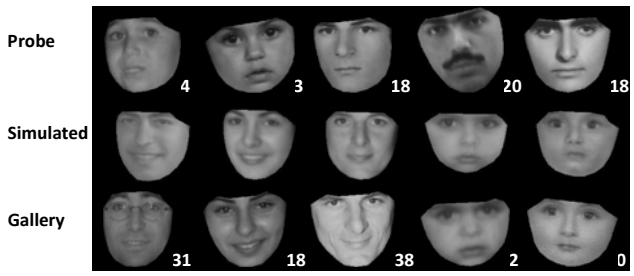


Fig. 6. Typical examples of aging effects simulation by AGES. The ages are marked at the right-bottom corner of the images

line of Table III. As can be seen that the MAEs of AGES$_r$ in the first two age ranges with relatively abundant images are similar to those of AGES. But its MAE in 31-69 is remarkably lower than that of AGES because the independent training in this range prevents the model from being biased to other age ranges with more training samples. Note that the MAEs of AGES$_r$ are obtained under the assumption that the age range of the test image is known. When this does not hold, an age range estimator is needed. AGES can be directly used for age range estimation after re-labeling the training data with the age ranges where the corresponding ages fall into. The resulted prediction will be an age range label indicating which subspace to use for further estimation of the exact age. The MAE of such two-layer age estimation is 6.52 (shown in parentheses in Table III), which improves the overall MAE of AGES (6.77).

E. Aging Effects Simulation and Face Recognition

As mentioned in Section IV-B, given a face image, AGES can be used to simulate face images at different ages. Besides the direct applications of aging effects simulation, such as aging missing children, it can be used for face recognition systems across ages.

For each subject in the FG-NET Aging Database, 10 pairs of face images are randomly selected, the first one as ‘gallery’ face and the second one as ‘probe’ face. Usually there is remarkable age difference between them. Given a probe face, the objective of aging effects simulation is to generate a face image at the age of the gallery face. Some typical results of the simulation by AGES are shown in Fig. 6. As can be seen that the simulated faces look quite similar to the real faces (the gallery faces), only with slight difference in pose, illumination, or expression. It is noteworthy that for the first probe face, the simulated face looks relatively more different from the gallery face. This might be because that the gallery face wears glasses, which is impossible to be predicted based on the 4-year-old probe face. To evaluate the simulation

quantitatively, the difference between images is calculated as the Mahalanobis Distance (MD) between the Appearance Model parameters. The average MD from the original probe faces to the gallery faces is 18.83, while that between the simulated faces to the gallery faces is 11.92, which reveals that the simulation makes the probe faces more similar to the gallery faces.

If one gallery face from each subject (82 subjects in the FG-NET Aging Database) is selected and composes a database, then each probe face can be recognized by this database. The most common implementation is to calculate the similarity between a probe face and each gallery face in the database, then recognize the probe as the person in the most similar gallery image. Here the Mahalanobis Distance is used again as the similarity measure. Also, we use the same 10 gallery-probe pairs selected from each subject in the aging simulation experiment. Note that the gallery set and the probe set are both selected randomly and they do not have intersection. Each time, the gallery face in one pair from each subject is used to build a database and the probe face in that pair is used to constitute a test set corresponding to the database. In total, 10 gallery databases and 10 corresponding probe sets are composed. One face recognition test is performed on each pair of them. The average recognition rate of the 10 tests without aging simulation is 14.39%. If the probe face is first simulated by AGES to the age of the gallery face, then the average recognition rate can be improved to 38.05%. Of course, the assumption that the ages of both the probe and the gallery faces are known before the recognition is sometimes unsatisfactory. One possible way to solve this problem is to simulate the whole aging pattern from the gallery/probe face and recognize the 'probe aging pattern' based on the database of the 'gallery aging patterns'.

VI. Conclusion

This paper proposes an automatic age estimation method named AGES, which improves our earlier work [5]. It is interesting to note that, at least under the experimental configuration in this paper, the performance of AGES is not only significantly better than that of the state-of-the-art algorithms, but also comparable to that of the human observers.

The current preprocessing method in AGES relies on many landmark points in the face images, eventually these landmarks should be determined by applying automatic landmarking algorithms like [3]. Moreover, the current preprocess does not retain the information about the outer contour size of the face. However, face size varies across ages, especially during formative years. Hence as

a future work, taking the size and shape of the face contour into consideration might significantly improve the accuracy of AGES, especially for age estimation on children's faces.

Besides age estimation, AGES can be utilized in other computer vision tasks. For example, with the ability to simulate facial aging effects, AGES can be used for face recognition across ages, which has been tested in the experiment. More generally, pose and illumination variations are always troublesome in computer vision systems. Similar to AGES dealing with images at different ages, images under different pose and illumination conditions can be treated as a whole (analogous to an aging pattern). This idea has been explored in face recognition, known as the 'Eigen Light-field' [6], [30]. In order to model the light-field, a 'generic training data set' is required in such works, which contains face images under all possible pose and illumination conditions. But this is not always available in reality. By the algorithm dealing with missing data proposed in this paper, the light-field based approaches can be generalized to the case when not all pose and illumination variations are included in the training set. This will be further investigated in the future work.

ACKNOWLEDGMENT

The authors would like to thank Y. Zhang, G. Li and H. Dai for their help, and Dr. A. Lanitis for providing the FG-NET Aging Database. Part of the work was done when Xin Geng was at the LAMDA Group, Nanjing University.

REFERENCES

[1] B. Bruyer and J.-C. Scailquin, "Person recognition and ageing: The cognitive status of addresses - an empirical question," *Int'l Journal of Psychology*, vol. 29, no. 3, pp. 351–366, 1994.

[2] M. Burt and D. Perrett, "Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and color information," *Proc. the Royal Society of London B: Biological Sciences*, vol. 259, no. 1355, pp. 137–143, 1995.

[3] N. Duta, A. K. Jain, and M.-P. Dubuisson-Jolly, "Automatic construction of 2D shape models." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 5, pp. 433–446, 2001.

[4] G. J. Edwards, A. Lanitis, and C. J. Cootes, "Statistical face models: Improving specificity," *Image Vision Comput.*, vol. 16, no. 3, pp. 203–211, 1998.

[5] 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*, Santa Barbara, CA, 2006, pp. 307–316.

[6] R. Gross, I. Matthews, and S. Baker, "Eigen light-fields and face recognition across pose," in *Proc. the 5th IEEE Int'l Conf. Automatic Face and Gesture Recognition*, Washington, DC, 2002, pp. 3–9.

[7] C. Hill, C. Solomon, and S. Gibson, "Aging the human face - a statistically rigorous approach," in *Proc. the IEE Int'l Symp. on Imaging for Crime Detection and Prevention*, London, UK, 2005, pp. 89–94.

[8] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[9] T. J. Hutton, B. F. Buxton, P. Hammond, and H. W. W. Potts, "Estimating average growth trajectories in shape-space using kernel smoothing," *IEEE Trans. Med. Imag.*, vol. 22, no. 6, pp. 747–753, 2003.

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

[11] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," *Comput. Vis. Image Und.*, vol. 74, no. 1, pp. 1–21, 1999.

[12] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, no. 1, pp. 621–628, 2004.

[13] A. Lanitis, C. J. Taylor, and T. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 4, pp. 442–455, 2002.

[14] A. Leonardis and H. Bishof, "Robust recognition using eigenimages," *Comput. Vis. Image Und.*, vol. 78, no. 1, pp. 99–118, 2000.

[15] A. J. O'Toole, T. Vetter, H. Volz, and E. Salter, "Three-dimensional caricatures of human heads: Distinctiveness and the perception of age," *Perception*, vol. 26, no. 6, pp. 719–732, 1997.

[16] E. A. Patrick and F. P. Fischer, "A generalized k-nearest neighbor rule," *Information and Control*, vol. 16, no. 2, pp. 128–152, 1970.

[17] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann, 1993.

[18] N. Ramanathan and R. Chellappa, "Face verification across age progression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 462–469.

[19] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, New York, NY, 2006, pp. 387–394.

[20] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. the 7th Int'l Conf. Automatic Face and Gesture Recognition*, Southampton, UK, 2006, pp. 341–345.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by backpropagating errors," *Nature*, vol. 323, no. 9, pp. 318–362, 1986.

[22] C. M. Scandrett, C. J. Solomon, and S. J. Gibson, "A person-specific, rigorous aging model of the human face," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1776–1787, 2006.

[23] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?" *Comput. Vis. Image Und.*, vol. 102, no. 2, pp. 117–133, 2006.

[24] The FG-NET Aging Database: http://sting.cycollege.ac.cy/∼alanitis/ fgnetaging/index.htm.

[25] B. Tiddeman, M. Burt, and D. Perret, "Prototyping and transforming facial texture for perception research," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 42–50, 2001.

[26] V. Vapnik, *Statistical Learning Theory*. New York: John Wiley and Sons, 1998.

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

[28] Y. Zana, R. M. C. Junior, R. S. Feris, and M. Turk, "Local approach for face verification in polar frequency domain," *Image Vision Comput.*, vol. 24, no. 8, pp. 904–913, 2006.

[29] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–459, 2003.

[30] S. K. Zhou and R. Chellappa, "Illuminating light field: Image-based face recognition across illuminations and poses," in *Proc. the 6th IEEE Int'l Conf. Automatic Face and Gesture Recognition*, Seoul, Korea, 2004, pp. 229–234.

[31] S. K. Zhou, B. Georgescu, X. S. Zhou, and D. Comaniciu, "Image based regression using boosting method," in *Proc. the 10th IEEE Int'l Conf. Computer Vision*, Beijing, China, 2005, pp. 541–548.

APPENDIX I

Proof: Suppose in iteration i, the training data is $\boldsymbol{x}_k^{(i)}$, the reconstruction of $\boldsymbol{x}_k^{(i)}$ by $\mathbf{W}_i$ is $[\hat{\boldsymbol{x}}_k^{(i)}(\mathbf{W}_i)]$, the reconstruction error of $\boldsymbol{x}_k^{(i)}$ by $\mathbf{W}_i$ is $\varepsilon(\boldsymbol{x}_k^{(i)}, \mathbf{W}_i)$, and the reconstruction error of the available features is $\varepsilon^a(\boldsymbol{x}_k^{(i)}, \mathbf{W}_i)$. Note that $[\hat{\boldsymbol{x}}_k^{(i)}(\mathbf{W}_i)]$ and the data of the next iteration, $\boldsymbol{x}_k^{(i+1)}$, share the same values at the positions of missing features, so

$$\varepsilon^a(\boldsymbol{x}_k^{(i)}, \mathbf{W}_i) = U([\hat{\boldsymbol{x}}_k^{(i)}(\mathbf{W}_i)], \boldsymbol{x}_k^{(i+1)}), \tag{8}$$

where $U(\boldsymbol{v}_1, \boldsymbol{v}_2)$ denotes the squared Euclidean distance between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Consequently, $\overline{\varepsilon}_i^a = \overline{U}$, where $\overline{\varepsilon}_i^a$ is the $\overline{\varepsilon}^a$ of iteration i and $\overline{U}$ is the mean value of $U([\hat{\boldsymbol{x}}_k^{(i)}(\mathbf{W}_i)], \boldsymbol{x}_k^{(i+1)})$. If $\boldsymbol{x}_k^{(i+1)}$ is also reconstructed by $\mathbf{W}_i$, then

$$\begin{aligned} \varepsilon(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_i) &= U([\hat{\boldsymbol{x}}_k^{(i+1)}(\mathbf{W}_i)], \boldsymbol{x}_k^{(i+1)}) \\ &\leq U([\hat{\boldsymbol{x}}_k^{(i)}(\mathbf{W}_i)], \boldsymbol{x}_k^{(i+1)}) \end{aligned} \tag{9}$$

because the line between $[\hat{\boldsymbol{x}}_k^{(i+1)}(\mathbf{W}_i)]$ and $\boldsymbol{x}_k^{(i+1)}$ is orthogonal to the subspace spanned by $\mathbf{W}_i$ so that they have the minimum Euclidean distance. Consequently, $\overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_i) \leq \overline{U}$, where $\overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_i)$ is the mean reconstruction error of $\boldsymbol{x}_k^{(i+1)}$ by $\mathbf{W}_i$. After applying PCA on $\boldsymbol{x}_k^{(i+1)}$, the new transformation matrix $\mathbf{W}_{i+1}$ minimizes the mean reconstruction error, thus

$$\overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_{i+1}) \leq \overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_i). \tag{10}$$

Obviously, $\overline{\varepsilon}_{i+1}^a \leq \overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_{i+1})$. So

$$\overline{\varepsilon}_{i+1}^a \leq \overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_{i+1}) \leq \overline{\varepsilon}(\boldsymbol{x}_k^{(i+1)}, \mathbf{W}_i) \leq \overline{U} = \overline{\varepsilon}_i^a. \tag{11}$$

Thus the algorithm will converge to minimize $\overline{\varepsilon}^a$. ■