Multifeature Anisotropic Orthogonal Gaussian Process for Automatic Age Estimation

ZHIFENG LI and DIHONG GONG, Chinese Academy of Sciences
KAI ZHU, Chinese University of Hong Kong
DACHENG TAO, University of Sydney
XUELONG LI, Chinese Academy of Sciences

Automatic age estimation is an important yet challenging problem. It has many promising applications in social media. Of the existing age estimation algorithms, the personalized approaches are among the most popular ones. However, most person-specific approaches rely heavily on the availability of training images across different ages for a single subject, which is usually difficult to satisfy in practical application of age estimation. To address this limitation, we first propose a new model called *Orthogonal Gaussian Process* (OGP), which is not restricted by the number of training samples per person. In addition, without sacrifice of discriminative power, OGP is much more computationally efficient than the standard Gaussian Process. Based on OGP, we then develop an effective age estimation approach, namely anisotropic OGP (A-OGP), to further reduce the estimation error. A-OGP is based on an anisotropic noise level learning scheme that contributes to better age estimation performance. To finally optimize the performance of age estimation, we propose a multifeature A-OGP fusion framework that uses multiple features combined with a random sampling method in the feature space. Extensive experiments on several public domain face aging datasets (FG-NET, MORPH Album1, and MORPH Album 2) are conducted to demonstrate the state-of-the-art estimation accuracy of our new algorithms.

CCS Concepts: • **Computing methodologies → Biometrics**;

Additional Key Words and Phrases: Age estimation, face image

ACM Reference format:
Zhifeng Li, Dihong Gong, Kai Zhu, Dacheng Tao, and Xuelong Li. 2017. Multifeature Anisotropic Orthogonal Gaussian Process for Automatic Age Estimation. *ACM Trans. Intell. Syst. Technol.* 9, 1, Article 2 (September 2017), 15 pages.
https://doi.org/10.1145/3090311

This work was funded by the External Cooperation Program of BIC, the Chinese Academy of Sciences (172644KYSB20160033), Australian Research Council Projects (FT-130101457, DP-140102164, and LP-150100671), and Natural Science Foundation of Guangdong Province (2014A030313688).

Authors' addresses: Z. Li, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; email: zhifeng.li@siat.ac.cn; D. Gong, Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; email: gongd@ufl.edu; K. Zhu, Department of Information Engineering, Chinese University of Hong Kong, Hong Kong; email: zk013@ie.cuhk.edu.hk; D. Tao, UBTECH Sydney Artificial Intelligence Centre and the School of Information Technologies in the Faculty of Engineering and Information Technologies at The University of Sydney, J12, 6 Cleveland St, Darlington, NSW 2008, Australia; email: dacheng.tao@sydney.edu.au; X. Li, Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China; email: xuelong_li@opt.ac.cn.

https://doi.org/10.1145/3090311

1 INTRODUCTION

Face analysis and recognition is an interesting yet challenging research problem in multimedia and intelligent systems (Ni et al. 2011; Fu and Huang 2008; Hsieh et al. 2009; Tawari and Trivedi 2013; Chen et al. 2015; Kotropoulos et al. 2000; Berretti et al. 2012; Liu et al. 2015a; Ewerth et al. 2012; Li et al. 2016; Liu et al. 2015b; Fu and Zheng 2006; Su et al. 2010). As part of this research topic, automatic age estimation from an input face image often plays a crucial role in multimedia communication and human-computer interaction (HCI). Automatic age estimation refers to the act of estimating age based on an input face image, which has many potential applications ranging from forensics to social media.

However, despite advances in automatic age estimation, it still remains a challenging problem for the following reasons. First, in addition to age-related variations, face images have other kinds of significant variations (caused by lighting, expressions, and poses). Second, living habits and working conditions usually have a significant effect on the aging process of the human face. The appearance of persons of the same age is frequently very different. And third, age-related discriminative information is usually more difficult to obtain since it has a slightly weaker influence on people's appearance than other factors. Figure 1 shows some example face images from the well-known FG-NET Aging Database (2015). These example images demonstrate several types of challenge in age estimation: poor image quality, large variation of poses and expressions, and relatively slow facial aging in older age ranges (which makes the prediction for older faces more difficult), and so on. Recently, several works (Zhang and Yeung 2010; Geng et al. 2007, 2008; Lanitis et al. 2002; Geng and Smith-Miles 2009; Lanitis et al. 2004; Yan et al. 2007, 2008; Guo et al. 2008; Xiao et al. 2009; Thukral et al. 2012; Wu et al. 2012; Chen and Hsu 2013; Chang et al. 2011; Guo and Mu 2010, 2011; Zhu et al. 2014; Fu et al. 2010; Guo et al. 2009a; Han et al. 2015; Ni et al. 2011; Lu and Tan 2010; Fu and Huang 2008; Guo et al. 2009b; Fu et al. 2007; Li et al. 2010; Yi et al. 2014; Yang et al. 2015; Chang and Chen 2015) have addressed the age estimation problem, as summarized in the following.

Classification Versus Regression. The age estimation problem can be treated as a multiclass classification problem, in which each age is taken as a class label, or as a regression problem, in which each age is used as a regression value. For the classification approaches, they first learn the individual models for different classes and then estimate the age by finding the best-matching model that fits the testing samples. For example, Geng et al. (2007) proposed the AGES algorithm to learn 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. In 2009, Geng and Smith-Miles (2009) proposed another classification-based algorithm, called *MSA*, to learn aging patterns through a series of multilinear subspace analysis operating on three-dimensional tensors for age prediction. In both of these methods, the age estimation for a testing subject is based on model comparison—that is, the age of the testing subject is predicted as the corresponding age of the model that can best fit the testing subject. Different from the classification-based approaches, the regression-based methods treat the ages as the regression values, and the age is predicted as the output of the regression models learned from training data. For example, Zhang and Yeung (2010) proposed a regression method based on a warped Gaussian Process (WGP). Earlier algorithms like AGES or MSA are based on classification models, but most recent approaches, which generally achieve significantly better prediction accuracy, are based on regression models, suggesting that regression approaches are more effective for age estimation than the classification approaches. From our point of view, this is mostly because the regression approaches can preserve the intrinsic ordering of ages.

Global Versus Person Specific. Another perspective to view the existing approaches is to examine whether the learned age estimator is specific to subjects. The person-specific approaches learn

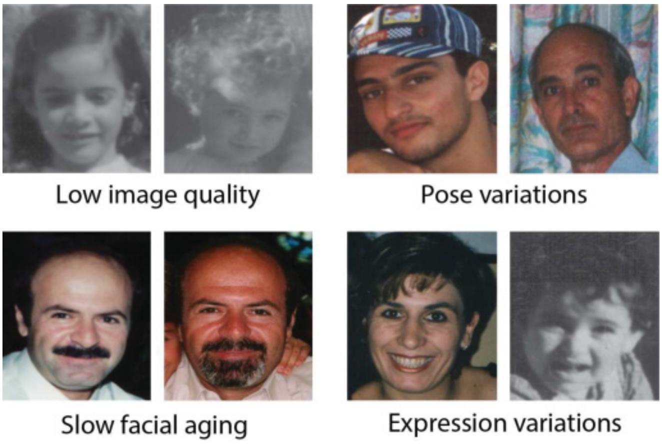

Fig. 1. Example face images from the FG-NET database demonstrating various difficulties in age prediction.

different models for different persons in the training stage, and at the testing stage the identity information of the testing person is used to choose the appropriate model that matches the testing subject. The rationale for this treatment is based on the observation that the aging process of human beings is quite person specific. Therefore, different subjects have different aging patterns associated with the person's genes or living habits. A representative person-specific method is the AGES algorithm (Geng et al. 2007), which learns one aging subspace for each person, then at the testing stage the age of the testing person is predicted by a model whose corresponding training person can match the testing person the best. Another representative work (Zhang and Yeung 2010) is multitask warped Gaussian Process (MTWGP), which was proposed in 2010 by Zhang et al. [2010]. MTWGP assigns the same noise level of Gaussian Process (GP) for samples from the same subject while assigning different noise levels for samples from different subjects. There are two major limitations for person-specific approaches. First, they rely heavily on the availability of training faces for a single person with different ages. Practically, it is usually difficult to satisfy. Second, the model complexity will increase significantly as the number of training samples increases, which may limit the application of these methods.

Unlike the person-specific approaches, the global algorithms treat all training samples equally. The global algorithms differ from the person-specific approaches in that the preceding approach trains one single model for all different subjects, whereas the latter approach trains one model for each subject. Hence, the global algorithms do not need to introduce new models or model parameters to accommodate the newly available training samples. For example, Guo et al. (2008) developed an age manifold learning scheme for extracting face aging features and designed a locally adjusted robust regressor to predict human ages. Xiao et al. (2009) proposed an mkNN algorithm for age prediction. This method learns the distance metric to preserve the local neighborhoods of training samples and at the same time maximizes the distances between the data that are not in the same neighborhood in a semantic space. A recent work (Chen and Hsu 2013) learns a global support vector regression (SVR) model for age estimation, which demonstrates the state-of-the-art prediction accuracy.

In addition to learning age classification or regression models, some research attempts have focused on extracting age informative facial features. The idea of age informative features is that greater prediction may be achieved if the facial features capture crucial information relating to human aging. Representative works include Guo et al. (2009a) and Han et al. (2015). In Guo et al. (2009a), biologically inspired features (BIF) containing rich age information are extracted. Although BIF are mostly handcrafted, Han et al. (2015) exploit a boosting algorithm to select age-informative features, the features of which are selected adaptively. Either handcrafted or adaptive approaches, they are designed to capture as much age-related information as possible so that the models are able to better discriminate between different ages.

In this article, we first present a model called *Orthogonal Gaussian Process* (OGP) (Zhu et al. 2014), which is based on the GP for regression. The optimization of OGP is over an orthogonal space, which is much more efficient than the standard GP. Based on OGP, an anisotropic OGP (A-OGP) algorithm is then proposed for effective age estimation. To further improve the age estimation performance, we propose a multifeature A-OGP framework. This is an improvement on, and an extension of, the A-OGP algorithm using multiple features combined with a random sampling method in the feature space. In this framework, we first represent each face image using a patch-based local feature representation scheme, in which three feature descriptors—scale-invariant feature transform (SIFT), multiscale local binary pattern (MLBP), and BIF—are used. A-OGP is then performed on a collection of random subspaces to construct an ensemble of classifiers for automatic age estimation. Extensive experiments across several public domain face aging databases are conducted to evaluate the effectiveness of our algorithms.

The major contributions of this article are summarized as follows. First, we propose an OGP algorithm that is more efficient than the standard GP. Second, based on OGP, we propose an anisotropic noise level learning algorithm called *A-OGP*, which is very beneficial in improving age estimation accuracy. Third, based on the preceding new methods, we propose a robust age estimation framework, called the *multifeature A-OGP fusion framework*, demonstrating state-of-the-art age estimation performance across several public domain face aging databases.

The rest of this article is organized as follows. In Section 2, we introduce the related work. In Section 3, we present the proposed approaches. In Section 4, we introduce the experimental results. We conclude in Section 5.

2 THE GP REVISITED

2.1 Gaussian Process

In this section, we first briefly revisit the GP for regression. Suppose that we are given a training set of $X_N = \{\vec{x}_n\}_{n=1}^{N}$ and real-valued targets $Y_N = \{y_n\}_{n=1}^{N}$, where X_N is the training features and Y_N is the training target. In the GP regression model, we define a latent variable z_n for each data point. The prior distribution for the latent variable is $Z \sim N(0_n, K)$, where $Z = (z_1 \ldots z_n)^T$ and K is a kernel matrix defined on X using a kernel function k_θ. The marginal likelihood is calculated as

$$p(Y|X) = \int p(Y|Z)p(Z|X)dZ = N(0_n, K + \sigma^2 I).$$

The negative log-likelihood of all data points can be expressed (with some constant terms removed) as

$$l = \frac{1}{2}[Y^T(K + \sigma^2 I)Y + \ln|K + \sigma^2 I|].$$

By minimizing the preceding negative log-likelihood with respect to parameter θ, we can make a prediction for any unseen test data point x^*. The distribution of corresponding regression value y^* is given as

$$\begin{pmatrix} Y \\ y^* \end{pmatrix} = N\left(0_{n+1}, \begin{pmatrix} K + \sigma^2 I & k^* \\ k^{*T} & k_\theta(x^*, x^*) + \sigma^2 \end{pmatrix}\right).$$

We then obtain the predictive distribution $p(y^*|x^*, X, Y)$ as a Gaussian distribution with mean $m^* = k^{*T}(K + \sigma^2 I)^{-1} Y$.

2.2 Warped Gaussian Process

WGP (Snelson et al. 2004) is an extension of the standard GP for regression. In WGP, the regression targets are warped into a latent space by a nonlinear monotonic function so that the transformed data can be well modeled by a GP.

Suppose that we are given a dataset D, consisting of N pairs of input vectors $X_N = \{\vec{x}_n\}_{n=1}^N$ and real-valued targets $Y_N = \{y_n\}_{n=1}^N$ that are transformed into latent space $G_N = \{g_\theta(y_n)\}_{n=1}^N$ with a monotonic function $g_\theta : R \rightarrow R$ parameterized by θ. Our training aim is to learn a set of model parameters (α, β, θ) where $\alpha, \beta > 0$ such that the following negative log-likelihood function can be minimized:

$$l = \frac{1}{2}\left[\vec{g}^T(\alpha K + \beta I_N)^{-1}\vec{g} + \ln|\alpha K + \beta I_N|\right] - \sum_{i=1}^{N} \ln g'_\theta(y_n), \tag{1}$$

where $\vec{g} = (g_\theta(y_1), \ldots, g_\theta(y_n))^T, K_{mn} = exp^{-\frac{||\vec{x}_m - \vec{x}_n||^2}{2\sigma^2}}$, and I_N is the $N \times N$ identity matrix. Based on the successful application of WGP in Zhang and Yeung (2010), we set the transformation function as

$$g_\theta(y) = a\ln(by + c) + d,$$

where $a, b > 0$ to ensure the function is monotonically increasing. The derivatives with regard to model parameters are given by

$$\frac{\partial l}{\partial \alpha} = \frac{1}{2}tr\left(\left[\tilde{K}^{-1} - \tilde{K}^{-1}\vec{g}\vec{g}^T\tilde{K}^{-1}\right]K\right)$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{2}tr\left(\left[\tilde{K}^{-1} - \tilde{K}^{-1}\vec{g}\vec{g}^T\tilde{K}^{-1}\right]\right)$$

$$\frac{\partial l}{\partial \theta} = \vec{g}^T\tilde{K}^{-1}\frac{\partial \vec{g}}{\partial \theta} - \sum_{i=1}^{N}\frac{\partial \ln g'_\theta(y_n)}{\partial \theta},$$

where $\tilde{K} = \alpha K + \beta I_N$. The problem can be solved by gradient-based optimization algorithms such as conjugate methods. At the testing stage, suppose that we are given a testing feature $\vec{t}$, then the prediction value is given by $m = \vec{k}_t^T\tilde{K}^{-1}\vec{g}$,

where

$$\vec{k}_t = \left[\alpha exp^{-\frac{||\vec{x}_1 - \vec{t}||^2}{2\sigma^2}}, \ldots, \alpha exp^{-\frac{||\vec{x}_N - \vec{t}||^2}{2\sigma^2}}\right]^T,$$

In our experiments, we find that it achieves the best performance when $\sigma^2 = \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N ||\vec{x}_i - \vec{x}_j||^2$

3 PROPOSED APPROACH

3.1 Orthogonal Gaussian Process

In this section, we propose OGP, which is significantly more efficient than WGP. For the WGP method, the major computational cost in each iteration step involves an inversion of matrix $\tilde{K}$ that is of size $N \times N$, where N is the number of training samples. Hence, when N is large, the computational cost would be extremely high. To alleviate this problem, we developed OGP, which can optimize the model parameters in a transformed space in which more efficient computation is enabled.

We start by using singular value decomposition (SVD) to decompose K in (1) into three matrices:

$$\mathrm{K} = \mathrm{USU^T} \tag{2}$$

where $S = diag\,(\lambda_1, \ldots, \lambda_N)$ is a diagonal matrix whose diagonal elements are real-valued positive eigenvalues as K is symmetric positive definite. U is an orthogonal matrix such that $UU^T = I$. We can obtain (3) by substituting (2) into (1),

$$l = \frac{1}{2}\sum_{n=1}^{N}\left(\frac{h_n^2}{\alpha\lambda_n + \beta} + ln\,(\alpha\lambda_n + \beta)\right) - \sum_{i=1}^{N} lng'_\theta\,(y_n)\,, \tag{3}$$

where $h_n = \vec{u}_n^T\vec{g}$ and $\vec{u}_n$ is the $n\text{-}th$ column vector of U. Given the transformed representation, we could easily find the derivatives:

$$\frac{\partial l}{\partial \alpha} = \frac{1}{2}\sum_{n=1}^{N}\frac{\left(\alpha\lambda_n + \beta - h_n^2\right)\lambda_n}{(\alpha\lambda_n + \beta)^2},$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{2}\sum_{n=1}^{N}\frac{\left(\alpha\lambda_n + \beta - h_n^2\right)}{(\alpha\lambda_n + \beta)^2},$$

$$\frac{\partial l}{\partial \theta} = \sum_{n=1}^{N}\frac{g_n}{\alpha\lambda_n + \beta}\frac{\partial g_n}{\partial \theta} - \sum_{i=1}^{N}\frac{\partial \mathrm{ln} g'_\theta\,(y_n)}{\partial \theta},$$

where g_n is the $n\text{-}th$ element of $\vec{g}$.

It can easily be seen that our algorithm only needs to compute a number of scalar multiplications at each iteration step, which significantly improves the effectiveness and the efficiency. The most computational-cost step of OGP is the SVD of matrix K of size $N \times N$; however, it only need to be computed once. To compare to the standard WGP that requires one to compute inversion of $N \times N$ matrix in each iteration step, it is proved that our approach has a much better computational efficiency in practice. In addition, since the computation of each iteration in WGP is costly, we may need to stop it before it converges, which usually would slightly affect the accuracy of the given solution.

3.2 Anisotropic OGP

It should be noted that the OGP method is inherently an isotropic model based on the assumption that the samples from the same subject should share the same noise level. This limitation would decrease the performance of age estimation in real applications. To overcome this problem, we propose an A-OGP algorithm in this section to improve the age estimation performance, described as follows. Note that l in (3) can be minimized with respect to α and β when $\alpha\lambda_n + \beta = h_n^2$ for $n = 1, \ldots, N$, which can be written in a vector form:

$$\vec{h}_2 = \alpha\vec{\lambda} + \beta\vec{1}, \tag{4}$$

where $\vec{h}_2 = (h_1^2, \ldots, h_N^2)^T$ and $\vec{\lambda} = (\lambda_1, \ldots, \lambda_N)^T$. Note that this equation does not hold generally, as $\vec{h}_2$ is of dimension N and cannot be represented by $span(\vec{\lambda}, \vec{1})$ when $N > 2$. The positive scalar β could be replaced by a vector whose elements are all positive values in an ARD anisotropic GP model, which gives

$$\vec{h}_2 = \alpha\vec{\lambda} + \vec{\beta}. \tag{5}$$

However, this treatment will introduce additional $(N - 1)$ free parameters to the system, which usually causes an overfitting problem (in our experiments, we found that model (5) performs much worse than model (4)).

The idea of our anisotropic model is to limit the number of introduced free parameters so that l can be dramatically reduced and control model complexity at the same time. Suppose that there are M ($M \ll N$) free parameters introduced; based on the current estimate of α, we can observe that the $\vec{h}_2$ is more accurate with $\alpha\vec{\lambda} + \vec{\beta}$ as the angle between $\vec{\beta}$ and $\vec{h}_2 - \alpha\vec{\lambda}$ becomes smaller. According to this observation, we first cluster the elements of the $\vec{h}_2 - \alpha\vec{\lambda}$ vector into M groups and then assign a same value to each element in the same group of $\vec{\beta}$, which gives a $\vec{\beta}$ of M distinct values. For instance, if $M = 3$ and $\vec{h}_2 - \alpha\vec{\lambda} = [1\ 2\ 4\ 9\ 5\ 8]$, then we can form three groups: 1,2, 4,5, and 8,9. The corresponding $\vec{\beta}$ is $[\beta_1\ \beta_1\ \beta_2\ \beta_3\ \beta_2\ \beta_3]$. Here we restrict $\vec{\beta}$ to contain only $M = 3$ free parameters; assigning the average value of each group, such as $\vec{\beta} = [1.5,\ 1.5,\ 4.5,\ 8.5,\ 4.5,\ 8.5]$, gives the optimal approximation to $\vec{h}_2$ with $\alpha\vec{\lambda} + \vec{\beta}$.

The proposed anisotropic model has absolute advantages over both the ARD anisotropic model and the method in Zhang and Yeung (2010). Compared to our model, the ARD anisotropic model has a high risk of overfitting due to the inevitable high model complexity, which would drop its performance. For the method in Zhang and Yeung (2010), the authors introduced a limited amount of freedom for $\vec{\beta}$, based on the assumption that samples from the same subject should share the same noise level. This kind of treatment needs to introduce extra priors over $\vec{\beta}$ to avoid the risk of overfitting that often occurs in the ARD anisotropic model. In addition, the number of free parameters of the method in Zhang and Yeung (2010) for $\vec{\beta}$ would grow linearly with the number of training subjects, which would lead to high model complexity in large-scale age estimation applications. In addition, unlike the model in Zhang and Yeung (2010), our anisotropic model is based on OGP, without relying on any assumption and with a clearer mathematic explanation.

3.3 Multifeature A-OGP Fusion Framework

Inspired by the successful application of multifeature fusion in the face research community, we propose a multifeature fusion framework in this section, using subspace sampling to further enhance age prediction accuracy. We extract multiple features for each face image using state-of-the-art face descriptors SIFT (Lowe 2004), MLBP (He and Wang 1990), and BIF (Mu et al. 2009). All of the descriptors are able to extract age informative features for age estimation. The SIFT features capture the gradient information, and the MLBP features encode the orientation of local patterns, whereas BIF has been shown to be very effective in age estimation. In this article, we have applied the dense version of SIFT features (DSIFT) where features are extracted densely over the entire face (e.g., grids). The final feature representation of an image is a concatenation of all of these features, which we refer to as super-long feature representation. The concatenation of different features captures much richer information than a single feature, as these features are extracted from descriptors of complementary functionalities.

A straightforward approach is to treat the whole super-long feature representation as a single large feature vector and apply the aforementioned method to it for age estimation. Although this approach would seem to utilize all of the information in the super-long feature representation, it has several problems. First, the data size of the original feature vector is very large. Directly handling such a long feature vector is too costly. Second, the original feature vector contains three kinds of local features of different scales and measurements. The incompatibility problem of different scales and measurements would be encountered in directly handling such a complex feature vector.

To alleviate these problems, we develop a multifeature fusion framework based on a random subspace method (Ho 1998) to construct multiple subclassifiers by randomly sampling the feature

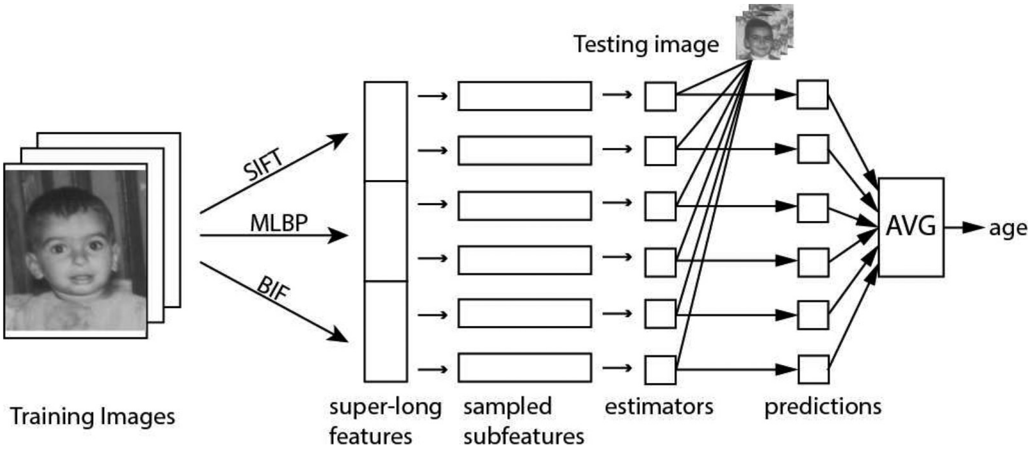

Fig. 2. Illustration of the multifeature A-OGP fusion framework pipeline.

space. The final matching score is the sum of matching scores from individual subclassifiers. The detailed algorithm of the multifeature fusion framework is described as follows:

1. Apply PCA [Turk and Pentland 1991] on each kind of feature (DSIFT, MLBP, or BIF) in a super-long feature and keep all eigenvectors with nonzero eigenvalues as the candidate to construct the random subspaces.
2. For each PCA-compressed feature (DSIFT, MLBP, or BIF) with D components, randomly select K components from the D components with $K \approx 0.8D$. Here, D is the dimension of features to selected from, and K is dimension of sampled features (e.g., random subfeatures).
3. Combine the selected components (from the three kinds of feature descriptors) into a new feature vector.
4. For each super-long feature vector, repeat steps 1 through 3 to obtain the new feature vector.
5. Construct a training set based on the new feature vectors from the preceding steps, denoting this training set as $\hat{X}$.
6. Train an A-OGP–based age estimator based on the training set $\hat{X}$ obtained from the preceding steps.

The preceding procedures are repeated M times, producing M age estimators trained on the different sets of data with overlapping components. The age estimators are trained on different subcomponents (e.g., random subspace) of facial features randomly. Since subcomponents are corresponding to specific location in the face, these age estimators are trained to capture discriminative features of different facial regions.

Each of the estimators can give a single age estimation, and the final age estimation value is an average over M outputs of the classifiers. Figure 2 illustrates the pipeline of the multifeature A-OGP fusion framework.

The proposed random subspace procedure is useful for boosting age estimation performance for the following reasons:

1. It makes better use of the limited amount of training data by repeatedly constructing different training data with some level of overlap.

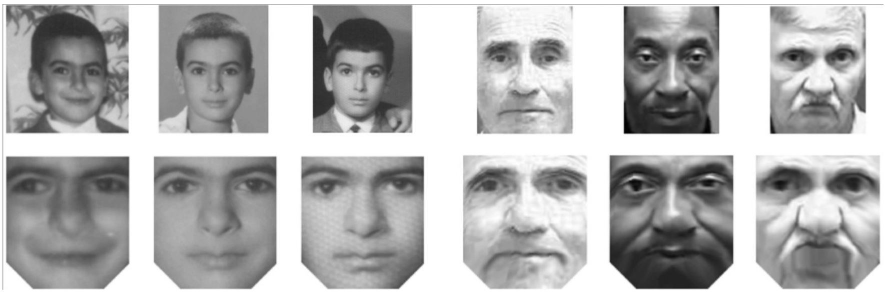

Fig. 3. The first row shows the original example faces for the FG-NET dataset (left) and the MORPH Album 2 dataset (right). The second row shows the corresponding normalized faces.

2. More importantly, by averaging the prediction values, the prediction becomes more robust and resistant to noises and outliers.

3.4 Discussion

The proposed multifeature A-OGP fusion framework has several advantages. First, the fusion framework is an improvement and extension of the A-OGP method, so it is expected to inherit the positive property (an anisotropic noise level learning scheme) of A-OGP and provide more flexibility. Second, the idea of using random subspace sampling techniques to construct an ensemble of subspaces allows us to work on more manageable-size data with respect to the number of training samples. Third, the multifeature A-OGP method has the capability to effectively combine the rich information conveyed by the densely sampled feature descriptors (DSIFT, MLBP, and BIF), which are complementary to some extent. The good performance described in the experiment section demonstrates the effectiveness of the multifeature A-OGP method for age estimation.

4 EXPERIMENTS

In this section, we conduct extensive experiments to investigate the performance of our approaches (OGP, A-OGP, and multifeature A-OGP). In OGP and A-OGP, we use the features of SIFT and MLBP, whereas for the proposed multifeature A-OGP framework, we use the fusion of three kinds of features (SIFT, MLBP, and BIF).

4.1 Databases and Evaluation Metric

For face aging research, there are two well-known public domain databases: FG-NET [FG-NET Aging Database 2015] and MORPH [Ricanek and Tesafaye 2006]. The FG-NET dataset contains 1,002 face images from 82 different individuals. Each person has multiple images taken at different ages. The MORPH database has two separate datasets: Album 1 and Album 2. There are 1,690 face images from 625 different subjects in MORPH Album 1. The MORPH Album 2 that we used in this work is an extended version of the original one. Specifically, it is the same as the one we used in Li et al. (2011). It consists of more than 78,000 face images from about 20,000 different persons and is currently the largest public domain face aging dataset. Figure 3 shows some examples for FG-NET and MORPH, as well as the corresponding normalized face images.

We use the leave-one-person-out (LOPO) testing strategy for the FG-NET and MORPH Album 1 datasets, which is a common practice in this field. For the MORPH Album 2 dataset, we randomly select 1,000 subjects as the testing set and the remaining 9,000 subjects as the training set (note that there is no overlap of subjects between the training and testing sets).

Table 1. Comparison of Training Time (in Seconds) at Different Numbers of Training Samples

	WGP	OGP	A-OGP
100	0.223	0.054	0.063
200	0.462	0.071	0.084
300	0.920	0.109	0121
400	2.171	0.210	0.231
500	2.533	0.287	0.314
600	4.802	0.594	0.613

Table 2. MAE on the FG-NET Dataset

	DSIFT	MLBP	DSIFT+MLBP
WGP	5.24	5.38	5.13
OGP	5.12	5.30	4.98
A-OGP	4.87	5.02	**4.76**

Table 3. MAE on the MORPH Album 1 Dataset

	DSIFT	MLBP	DSIFT+MLBP
WGP	5.08	5.42	4.88
OGP	5.02	5.16	4.76
A-OGP	4.34	4.57	**4.12**

Mean age error (MAE) is used for all results reported in this article, which is the average of the absolute difference between the estimated age and the ground-truth age.

4.2 Comparison for Computational Efficiency

First, we compare the computational efficiency between our algorithms (OGP and A-OGP) and the WGP algorithm. The time is reported as the average training time on the same machine over five random splits of the MORPH Album 1 dataset to make fair comparisons. The comparative results are reported Table 1, from which we can observe that there is a significant improvement in computational efficiency in our approaches over WGP.

4.3 Performance Evaluation of Our Algorithms Under Different Settings

To better explore the effectiveness of our algorithms (OGP and A-OGP), we use different configurations to compare their MAE performance. The results are reported in Tables 2 and 3. From these results, we make the following observations. First, our baseline algorithm OGP has slightly better performance than WGP, which reveals the fact that the optimization over an orthogonal space in OGP can obtain a better local optima solution. Second, it is clear that A-OGP has better performance over both OGP and WGP, which proves the effectiveness of our anisotropic algorithm. Third, the fusion of local feature descriptors can also improve age estimation accuracy.

4.4 Benchmark Comparison on the FG-NET Database

The comparison results of our algorithms with the state-of-the-art algorithms on the FG-NET dataset are shown in Table 4. We use all data in the FG-NET dataset for performance evaluation of the proposed method. All algorithms in Table 4 follow the same LOPO testing strategy. In addition,

Table 4. Benchmark Comparison on the FG-NET Dataset

Methods	MAE
AAS (Lanitis et al. 2004)	14.83
WAS (Lanitis et al. 2002)	8.06
AGES (Geng et al. 2007)	6.77
KAGES (Geng et al. 2008)	6.18
RUN1 (Yan et al. 2007)	5.78
MSA (Geng and Smith-Miles 2009)	5.36
RUN2 (Yan et al. 2008)	5.33
LARR (Guo et al. 2008)	5.07
mkNN (Xiao et al. 2009)	4.93
MTWGP (Zhang and Yeung 2010)	4.83
TSR (Wu et al. 2012)	5.90
The method in Han et al. (2015)	4.80
A-OGP	**4.76**
Multifeature A-OGP fusion framework	**4.69**

Table 5. Benchmark Comparison on the MORPH Album 1 Dataset

Methods	MAE
AAS (Lanitis et al. 2004)	20.93
WAS (Lanitis et al. 2002)	9.32
AGES (Geng et al. 2007)	8.83
RUN1 (Yan et al. 2007)	8.34
LARR (Guo et al. 2008)	7.94
mkNN (Xiao et al. 2009)	10.31
MTWGP (Zhang and Yeung 2010)	6.28
The method in Chen and Hsu (2013)	5.41
OGP (Zhu et al. 2014) (FG-NET training)	5.47
OGP (Zhu et al. 2014) (LOPO)	4.76
A-OGP (FG-NET training)	**5.25**
Multifeature A-OGP fusion framework (FG-NET training)	**5.18**
A-OGP (LOPO)	**4.12**
Multi-feature A-OGP fusion framework (LOPO)	**3.95**

for fair comparison, the result of the method in Han et al. (2015) is the one obtained on all data of the FG-NET dataset without removing the low-quality face images. From these results, we make several observations. First, compared to the MTWGP algorithm, our algorithm is not only more effective but also is more efficient in computation (because MTWGP is theoretically slower than WGP in Table 1). Second, it is clear that our algorithm achieves state-of-the-art accuracy over all existing methods.

4.5 Benchmark Comparison on the MORPH Database

The comparison result of our algorithms with the state-of-the-art algorithms on the MORPH Album 1 dataset are shown in Table 5. Since each person in the MORPH Album 1 dataset has very limited face images (approximately two images), all listed algorithms except for ours are trained with the FG-NET face dataset. For our algorithm, we present MAE results for training with both

Table 6. Benchmark Comparison on the MORPH Album 2 Dataset

Methods	MAE
OHRank (Chang et al. 2011)	6.07
KPLS (Guo and Mu 2011)	4.18
RMIR (Ni et al. 2011)	6.06
The method in Guo and Mu (2010)	4.44
OPMFA (Lu and Tan 2010)	5.24 (for those of white race) 4.18 (for those of black race)
The method in Chen and Hsu (2013)	4.42
OGP (Zhu et al. 2014)	4.15
A-OGP	**3.92**
Multifeature A-OGP fusion framework	**3.81**

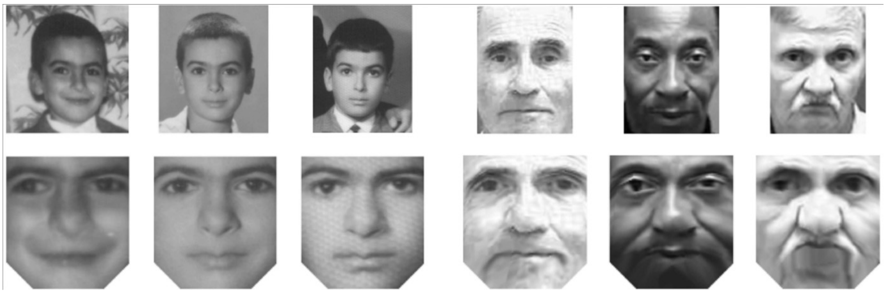

Fig. 4. The top row shows example images for which the age has been correctly identified, whereas the bottom row shows images for which the age prediction error is more than 10 years.

FG-NET data and MORPH Album 1 data (LOPO). Several points need to be highlighted. First, under the same testing configuration (the age estimators are trained from the FG-NET dataset), our algorithms obtain significant performance improvement over existing state-of-the-art methods. Second, our algorithm achieves another major boost in accuracy when trained with face images from Album 1 using the LOPO strategy. Third, compared to the A-OGP method, the multifeature A-OGP fusion framework improves the accuracy of age estimation in both configurations. Last, in Table 6, we compare our approaches to the existing results on the MORPH Album 2 dataset. Our approaches also outperform the other approaches on this dataset. This confirms our observation on the FG-NET dataset and MORPH Album 1 dataset. Compared to the A-OGP algorithm, the proposed multifeature A-OGP fusion framework obtains a much better result on this large-scale face aging dataset. This confirms the superiority of the multifeature A-OGP fusion framework over the A-OGP algorithm.

4.6 Some Discussions

To further explore which factors will achieve good prediction results and which factors will cause bad prediction results, we present good and bad prediction examples from the FG-NET database, shown in Figure 4. The top row illustrates good prediction examples, whereas the bottom row illustrates bad examples. We make the following observations from Figure 4. First, the good prediction examples are mostly of young ages with good image quality (frontal faces, good image focusing, etc.), with fewer significant age-unrelated variations (pose, expression, lighting, etc.). Second, the bad prediction examples are usually in the older age range (e.g., older than 40 years) or of poor image quality. Note that the aging process is very complex and highly person dependent. As a result, different persons may have very different rates of aging, depending on many intrinsic and extrinsic factors. In addition, the progression of the aging process for an adult is usually slower than it is for a child, so the face images at relatively old ages are more difficult to predict than those at young ages. Our research results are consistent with previous works in the literature.

5 CONCLUSIONS AND FUTURE WORK

In this article, we first proposed a new age estimation approach called *OGP*, which is much more efficient than the standard Gaussian approach. Based on OGP, we proposed two enhanced age estimation approaches: A-OGP and multifeature A-OGP. Extensive experiments were conducted on several public domain face aging databases to demonstrate the state-of-the-art estimation accuracy of our new algorithms. In future work, we would like to explore the possibility of applying the proposed models to other machine learning regression tasks.

REFERENCES

Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. 2012. Distinguishing facial features for ethnicity-based 3D face recognition. *ACM Transactions on Intelligent Systems and Technology* 3, 3, Article No. 45.

K. Y. Chang and C. S. Chen. 2015. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing* 24, 3 (2015), 785–798.

Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. 2011. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, Los Alamitos, CA, 585–592.

Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia* 17, 6 (2015), 804–815.

Yu-Lun Chen and Cheng-Ting Hsu. 2013. Subspace learning for facial age estimation via pairwise age ranking. *IEEE Transactions on Information Forensics and Security* 8, 12, 2164–2176.

FG-NET Aging Database. 2015. FG-NET Aging Database. Retrieved July 17, 2017, from http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html.

Ralph Ewerth, Markus Mühling, and Bernd Freisleben. 2012. Robust video content analysis via transductive learning. *ACM Transactions on Intelligent Systems and Technology* 3, 3, Article No. 41.

Yun Fu, Guodong Guo, and Thomas S. Huang. 2010. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 11, 1955–1976.

Yun Fu and Thomas S. Huang. 2008. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia* 10, 4, 578–584.

Yun Fu, Ye Xu, and Thomas S. Huang. 2007. Estimating human age by manifold analysis of face pictures and regression on aging features. In *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo*. IEEE, Los Alamitos, CA, 1383–1386.

Yun Fu and Nanning Zheng. 2006. M-face: An appearance-based photorealistic model for multiple facial attributes rendering. *IEEE Transactions on Circuits and Systems for Video Technology* 16, 7, 830–842.

Xin Geng and Kate Smith-Miles. 2009. Facial age estimation by multilinear subspace analysis. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*. IEEE, Los Alamitos, CA, 865–868.

Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. 2008. Facial age estimation by nonlinear aging pattern subspace. In *Proceedings of the 16th ACM International Conference on Multimedia*. ACM, New York, NY, 721–724.

Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. 2007. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 12, 2234–2240.

Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang. 2008. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing* 17, 7, 1178–1188.

Guodong Guo and Guowang Mu. 2010. Human age estimation: What is the influence across race and gender? In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'10)*. IEEE, Los Alamitos, CA, 71–78.

Guodong Guo and Guowang Mu. 2011. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, Los Alamitos, CA, 657–664.

Guodong Guo, Guowang Mu, Yun Fu, Charles Dyer, and Thomas Huang. 2009b. A study on automatic age estimation using a large database. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision.* IEEE, Los Alamitos, CA, 1986–1991.

Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang. 2009a. Human age estimation using bio-inspired features. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09).* IEEE, Los Alamitos, CA, 112–119.

Hu Han, Christina Otto, Xindong Liu, and Abhishek Jain. 2015. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3, 6, 1148–1161.

Dong-Chen He and Li Wang. 1990. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing* 28, 4, 509–512.

Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8, 832–844.

Chao-Kuei Hsieh, Shang-Hong Lai, and Yung-Chang Chen. 2009. Expression-invariant face recognition with constrained optical flow warping. *IEEE Transactions on Multimedia* 11, 4, 600–610.

Constantine L. Kotropoulos, Anastasios Tefas, and Ioannis Pitas. 2000. Frontal face authentication using discriminating grids with morphological feature vectors. *IEEE Transactions on Multimedia* 2, 1, 14–26.

Andreas Lanitis, Chrisina Draganova, and Chris Christodoulou. 2004. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 34, 1, 621–628.

Andreas Lanitis, Chris J. Taylor, and Timothy F. Cootes. 2002. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 4, 442–455.

Zhen Li, Yun Fu, and Thomas S. Huang. 2010. A robust framework for multiview age estimation. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'10).* IEEE, Los Alamitos, CA, 9–16.

Zhifeng Li, Dihong Gong, Qiang Li, Dacheng Tao, and Xuelong Li. 2016. Mutual component analysis for heterogeneous face recognition. *ACM Transactions on Intelligent Systems and Technology* 7, 3, Article No. 28.

Zhifeng Li, Park Unsang, and Anil K. Jain. 2011. A discriminative model for age invariant face recognition. *IEEE Transactions on Information Forensics and Security* 6, 3, 1028–1037.

Fan Liu, Jinhui Tang, Yan Song, Liyan Zhang, and Zhenmin Tang. 2015a. Local structure-based sparse representation for face recognition. *ACM Transactions on Intelligent Systems and Technology* 7, 1, Article No. 2.

Kuan-Hsien Liu, Shuicheng Yan, and C.-C. Jay Kuo. 2015b. Age estimation via grouping and decision fusion. *IEEE Transactions on Information Forensics and Security* 10, 11, 2408–2423.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2, 91–110.

Jiwen Lu and Yap-Peng Tan. 2010. Ordinary preserving manifold analysis for human age estimation. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'10).* IEEE, Los Alamitos, CA, 90–95.

Guowang Mu, Guodong Guo, Yun Fu, and Thomas S. Huang. 2009. Human age estimation using bio-inspired features. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09).* IEEE, Los Alamitos, CA, 112–119.

Bingbing Ni, Zheng Song, and Shuicheng Yan. 2011. Web image and video mining towards universal and robust age estimator. *IEEE Transactions on Multimedia* 13, 6, 1217–1229.

Karl Ricanek Jr. and Tamirat Tesafaye. 2006. MORPH: A longitudinal image database of normal adult age-progression. In *Proceedings of the 2006 7th International Conference on Automatic Face and Gesture Recognition (FGR'06).* IEEE, Los Alamitos, CA, 341–345.

Edward Snelson, Carl Edward Rasmussen, and Zoubin Ghahramani. 2004. Warped Gaussian Processes. *Advances in Neural Information Processing Systems* 16, 337–344.

Ya Su, Yun Fu, Qi Tian, and Xinbo Gao. 2010. Cross-database age estimation based on transfer learning. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing.* IEEE, Los Alamitos, CA, 1270–1273.

Ashish Tawari and Mohan Manubhai Trivedi. 2013. Face expression recognition by cross modal data association. *IEEE Transactions on Multimedia* 15, 7, 1543–1552.

Pavleen Thukral, Kaushik Mitra, and Rama Chellappa. 2012. A hierarchical approach for human age estimation. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12).* IEEE, Los Alamitos, CA, 1529–1532.

Matthew Turk and Alex P. Pentland. 1991. Face recognition using eigenfaces. In *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'91).* IEEE, Los Alamitos, CA, 586–591.

Tao Wu, Pavan Turaga, and Rama Chellappa. 2012. Age estimation and face verification across aging using landmarks. *IEEE Transactions on Information Forensics and Security* 7, 6, 1780–1788.

Bo Xiao, Xiaokang Yang, Yi Xu, and Hongyuan Zha. 2009. Learning distance metric for regression by semidefinite programming with application to human age estimation. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, New York, NY, 451–460.

Shuicheng Yan, Huan Wang, Thomas S. Huang, Qiong Yang, and Xiaoou Tang. 2007. Ranking with uncertain labels. In *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo*. IEEE, Los Alamitos, CA, 96–99.

Shuicheng Yan, Xi Zhou, Ming Liu, Mark Hasegawa-Johnson, and Thomas S. Huang. 2008. Regression from patch-kernel. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE, Los Alamitos, CA, 1–8.

Huei-Fang Yang, Bo-Yao Lin, Kuang-Yu Chang, and Chu-Song Chen. 2015. Automatic age estimation from face images via deep ranking. In *Proceedings of the 2015 26th British Machine Vision Conference (BMVC'15)*. IEEE, Los Alamitos, CA.

D. Yi, Z. Lei, and S. Z. Li. 2014. Age estimation by multi-scale convolutional network. In *Proceedings of the 2004 12th Asian Conference on Computer Vision (ACCV'14)*. IEEE, Los Alamitos, CA.

Yu Zhang and Dit-Yan Yeung. 2010. Multi-task warped Gaussian Process for personalized age estimation. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, Los Alamitos, CA, 2622–2629.

Kai Zhu, Dihong Gong, Zhifeng Li, and Xiouou Tang. 2014. Orthogonal Gaussian Process for automatic age estimation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, NY, 857–860.

Received December 2015; revised September 2016; accepted April 2017