Age Estimation with Expression Changes Using Multiple Aging Subspaces

Chao Zhang and Guodong Guo
West Virginia University
Lane Dept. of CSEE, Morgantown, WV
cazhang@mix.wvu.edu, guodong.guo@mail.wvu.edu

Abstract

Image-based human age estimation has become one of the interesting but challenging problems in computer vision and biometrics. It is even harder when the faces have different expressions. In this paper, we propose a weighted random subspace method to solve the relatively new problem: cross-expression age estimation. The proposed method does not depend on the learning of correlation between different expressions, and thus could work in the situation when the expression-correlation does not exist in the training data. We also explore the use of data from multiple datasets to further improve the estimation performance. Experiments on two aging datasets with explicit expression changes demonstrate that the proposed approach gives superior performance over the state-of-the-art method.

1. Introduction

Image-based human age estimation has been an active topic in the fields of computer vision and biometrics [4, 5, 15, 3]. Accurate age estimation has several practical applications, such as business intelligence, security, and age-specific HCI. Most work on age estimation only considers the neutral expression for both training and testing [5, 7, 6]. However, using neutral expression facial data for age estimation is not sufficient in real applications, as people naturally present various expressions, i.e., happy and sadness. In [9], the study shows that the changes of expression adversely affect the accuracy of age estimation, and the traditional direct estimation methodology cannot work well. Therefore, developing a method to work well under the changes of expressions is helpful and necessary in practice. In [9], an age estimation method consisting of correlation learning and discriminant mapping was proposed to estimate age under expression changes. The limitation of their method is the dependency of correlation between pairs of expressions, e.g., neutral and happy, from the same subject. In many cases, this assumption may be violated, for example, the correlation does not exist when two expressions are not present by the same group of people. Comparing to the data collection in other applications, such as object recognition and face detection, aging dataset is especially hard to build and collect, due to the nature of slow aging process. As a classical method, the random subspace [12] is an effective way to solve image-based classification problems [18], especially with a small sample size. In this study, we explore the feasibility of applying the random subspace idea to solve the age estimation problem. Besides, a cross-expression discriminative learning method is developed to handle the expression changes, and make aging features more discriminative. At last, we propose a graph-based classifier weighting method to fuse the multiple subspaces and make age estimation robust to expression changes. Our framework can be easily extended to use extra data for training and further improve the accuracies.

1.1. Major Contributions

1. We propose a random subspace based framework for cross-expression age estimation. It is for the first time that the random subspace technique is applied to age estimation.
2. We present a graph-based classifier fusion strategy to improve the performance of age estimation under expression changes, and show the superiority of the method over the state-of-the-art method.
3. A new experiment is conducted to utilize data from another dataset to improve the age estimation accuracy, demonstrating the flexibility of the proposed framework; While the correlation-based, state-of-the-art approach [9] cannot conduct this experiment.

2. Proposed Method

In order to develop a cross-expression human age estimator, we propose a new framework, shown in Figure 1. First, the training data $D_{N\times d}$ is fed into a feature selection module (FSM). The FSM could generate several feature sets $FS = [D^1_{N\times t}, \ldots, D^{NC}_{N\times t}]$, where d is the dimension of original feature vector, t is the number of selected features, N is the number of training samples, NC is the number of feature sets, the same as the number of random subspaces in this paper. Second, we build subspaces $RS = [RS_1, \dots, RS_{NC}]$ upon the feature sets. Then, within each single random subspace, we learn a classifier to predict human age. We use the SVMs (Support Vector Machines) [17] as our age estimator. During testing, multiple subspaces should give several predictions for the single test pattern, while variations of expressions will result in poor estimation for some base estimators. Therefore, inspired by the successful application of graph structure, we propose a graph-based weighting method, based on the adaptability of the base age estimators to a specific new expression.

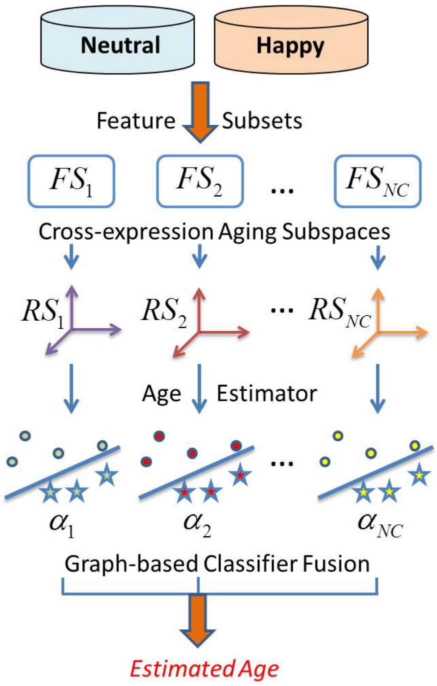

Figure 1. Framework of weighted random subspace method. "FS*" indicates feature sets; "RS*" indicates random subspaces; "α*" represents the optimal weights for multiple classifiers. The optimal weights are learned based on a graph smoothness assumption).

2.1. Feature Subsets Generation

The dimensionality of the extracted features is often large, while the sample size at each age class is small. The random sampling idea fits our problem very well, so we propose to use the random sampling scheme [12] to deal with the high dimensionality of features and prevent the problem of over-fitting. To the best of our knowledge, this is the first time that the random sampling scheme is applied to age estimation, although random sampling is a classical method. Let d be the dimensionality of the original feature space, for each subset we randomly choose t features with replacement. After NC feature subsets are sampled, each subset of them is used to build a discriminative subspace. In work [10], the authors argued the randomly selected features may not be optimal for some application, and they apply the reinforcement learning algorithm to select features by constructing a feature discrimination map. Other feature selection methods like Fisher score [1] and Laplacian score [11] could also be used to select better feature sets for random subspace method. Our experimental results show that when the number of selected features is set large enough (we set t = 500 in our experiments), the results of random sampling method are comparable with [10]. Also note that the random sampling process is much faster than the method proposed in [10] when the dimensionality of original feature space is as large as 5k or more (the dimensionality of feature we used is 4796).

2.2. Cross-Expression Discriminative Analysis

After the feature subsets are built, the possibility of overfitting is reduced because the discrepancy between the training data set size (i.e., the number of training examples) and feature dimensionality is reduced. Next, we deal with two new problems: the first one is how to make each single subspace discriminative from the perspective of age estimation, and at the same time make it invariant to expression changes; the other one is what fusion scheme to use given multiple predictions from multiple subspaces. Given these subsets, different methods could be used to learn the discriminative subspaces, such as LDA [13] and MFA [19]. Wang et al. [18] has shown that the LDA is useful for face recognition. We adopt the Fisher linear discriminant and adapt it to our problem where two factors should be considered. We aim to project the samples of different expressions to a common space in which the aging information is kept while the influence of expression changes is reduced. Notice here we do not learn the discriminative subspaces for each expression independently. The aim of this process is to make the aging patterns of different expressions "move" nearby if they belong to the same age, whereas move far away if they belong to different ages. We extend the classic Linear Discriminant Analysis (LDA), and adapt it to handle our cross-expression age analysis.

For the c-th $(1 < c < NC)$ feature subset containing "Neutral" and "Happy" expression, the mean feature vector for those samples with age equal to i is

$$\mu_i^c = \frac{1}{n_i^N + n_i^H} \left(\sum_{j=1}^{n_i^N} X_{i,j}^N(c) + \sum_{k=1}^{n_i^H} X_{i,k}^H(c) \right) \quad (1)$$

where n_i^N is the number of samples with "Neutral" expression, with the age equal to i. Similarly, n_i^H is for "Happy" expression. The mean feature vector for all ages is $\mu^c = \frac{1}{n} \sum_{i=1}^{n} \mu_i^c$. In the c-th feature subset, $X_{i,j}^N(c)$ is the j-th sample of the i-th age with "Neutral" expression.

$X_{i,j}^{H}(c)$ is the j-th sample of the i-th age with "Happy" expression.

The between-class scatter matrix S_B^c and the within-class scatter matrix S_W^c are computed as follows:

$$S_B^c = \sum_{i=1}^{n} (\mu_i^c - \mu^c)(\mu_i^c - \mu^c)^T \quad (2)$$

$$S_W^c = \sum_{i=1}^{n} (\sum_{j=1}^{N_i^N} (x_{i,j}^N(c) - \mu_i^c)(x_{i,j}^N(c) - \mu_i^c)^T \quad (3)$$

$$+ \sum_{k=1}^{N_i^H} (x_{i,k}^H(c) - \mu_i^c)(x_{i,k}^H(c) - \mu_i^c)^T)$$

According to the optimization criteria in Fisher linear discriminant, we can solve the generalized eigenvalue problem for the two expressions, e.g., Neutral and Happy, for age estimation.

$$S_B^c W^c = \lambda^c S_W^c W^c \quad (4)$$

The solution of the generalized eigenvalue problem, i.e., the eigenvector matrix W^c is used to project the aging patterns of two expressions for age estimation. We call the optimal solution cross-expression discriminant analysis, or CEDA.

Based on the projections on W^c, the aging patterns are expected to be transformed so that the aging differences between two different expressions can be minimized and the aging patterns of the same age in different expressions can be "pulled" towards similar distributions. Then the aging functions can be learned on the transformed aging patterns. In our work, we use the support vector machines (SVMs) [17] for aging function learning.

2.3. Graph-based decision fusion

The CEDA is used to extract the discriminative information within each feature subset, while at the same time reducing the influence of expression changes. The age estimation results are determined by the overall performance of multiple classifiers. In other words, the improvement of a single base classifier may not guarantee better estimation results. As a popular fusion scheme, the majority voting strategy is often applied to improve the overall classification performance. In our study, we name it Voting-based Random Subspace (V-RS). Suppose the age label ranges from l to u. For the i-th aging pattern, the estimated age of V-RS is given by

$$\hat{y}_i = \underset{l \le k \le u}{\operatorname{argmax}} \sum_{s=1}^{NC} I(h_i^s = k). \quad (5)$$

However, the majority voting scheme often substantially deteriorates if quite a few base classifiers perform poorly. Intuitively, the base classifiers with better classification accuracy should be assigned larger weights. In our problem, the feature subsets are randomly chosen from the original feature space, it is natural to state that some feature subsets are robust to expression changes and they are more suitable for making the estimations. Therefore, we further study how to determine the optimal weights to assign to each subspace classifier.

Let $H_i^S = [h_i^1 \dots h_i^{NC}]$ be the $1 \times NC$ vector of predicted ages of NC random subspace classifiers for the i-th sample from a new expression. Let $\alpha = [\alpha_1 \dots \alpha_{NC}]^T$ be the $NC \times 1$ weight vector, where α_s is the weight corresponding to the s-th base classifier. Then, the estimated age of the i-th sample from a new expression is

$$\hat{y}_i = \sum_{s=1}^{NC} \alpha_s h_i^s = H_i^S \alpha. \quad (6)$$

The key to get good estimation results is the optimal weight vector. In order to estimate the weights α, we incorporate the smoothness assumption on the probability distribution of the aging patterns for a specific expression. Consider N_t points $x_1, x_2, \dots, x_{N_t}$ in $\mathbb{R}^d$, we can use a p-nearest neighbor graph G to model the relationship between nearby data points. Notice here the graph G is obtained by computing the distance of each pair of nodes in the original feature space rather than in the CEDA projected space. Specifically, we put an edge between nodes i and j if x_i and x_j are "close", i.e., x_i and x_j are among p nearest neighbors of each other. We call the matrix S the adjacency matrix, defined by

$$S_{ij} = \begin{cases} 1, & \text{if } x_i \in N_p(x_j) \text{ or } x_j \in N_p(x_i) \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

The underlying assumption is that if two aging patterns are similar in the feature space, they should also have similar age labels. Then, we can write the objective function as follows:

$$\min_{\alpha:\alpha \ge 0} \sum_{i,j=1}^{N_t} (H_i^S \alpha - H_j^S \alpha)^2 S_{ij} \quad (8)$$

where H_i^S and H_j^S are the predicted ages for the i-th and j-th samples in the new expression data. S_{ij} is the edge weight between the two samples. N_t is the number of samples from the new expression.

This problem could be rewritten as follows:

$$\min_{\alpha:\alpha \ge 0} \alpha' H^{S'} L H^S \alpha \quad (9)$$

where L is the graph Laplacian associated with the new expression data, given by $L = D - S$, where D is the diagonal matrix given by $D_{ii} = \sum_{j=1}^{n} S_{ij}$. The minimization problem is a standard quadratic problem (QP) and can be solved by applying existing solvers.

After solving the optimal weights α, the estimated age is represented as the linear combination of all base predictions. We call this Weighted Random Subspace (W-RS).

2.4. Using Expressions from Different Subjects

The state-of-the-art approach to age estimation under expression changes is in [9], where two expressions of the same subject at the same age are needed in order to learn the correlation mapping between two expressions. This is required for the training data in order to deal with cross-expression age estimation. In practice, however, there are aging patterns of subjects without presenting the expression pairs, e.g., neutral and happy. The correlation-based method in [9] cannot utilize these aging patterns for learning. Considering the difficulty of aging pattern data collection (both the aging faces and the corresponding age labels), it is of great value if a method can be developed to utilize the aging patterns without requiring the expression pairs for learning.

In our proposed weighted subspaces age estimation framework, there is no need to use the training examples with exact expression pairs, e.g., neutral and happy, from the same subjects. The new formulation uses the aging patterns as a whole in each expression, without requesting the one-to-one correspondence in learning. Note that in order to compare with the performance of the method in [9], we used the same data for training and testing where expression pairs of the same subjects do exist. Now, we want to show that our new framework can do even better, by utilizing the aging patterns without expression pairs for each subject. This is a nice property of our new approach. To understand this clearly, we will add extra training data that do not present the expression pairs for each subject.

Without loss of generality, given two datasets D_{pri} and D_{aux}, where D_{pri} is a primary dataset and D_{aux} is considered as auxiliary dataset. The primary dataset gives both training data and test data, while the auxiliary dataset only provides training data. Here, we take "Neutral" to "Happy" as a cross-expression age estimation case to explain the process in details. During the training stage, "Neutral" data $D_{aux}^{Neutral}$ and "Happy" data D_{aux}^{Happy} are added up to the training data to learn the random subspace classifiers, while in testing, the test samples in the primary dataset are projected using the learned CEDA subspaces, followed by the age estimator. This experiment design is different from semi-supervised learning which relies on the use of the large number of unlabeled data; it also differs from the domain adaptation which tries to adapt the model learned from one dataset to another new dataset where few labeled samples are available. In general, the aim of domain adaptation is to improve the classification accuracy on a dataset B using the model learned from dataset A. To make the difference clearly, we show the ideas of different methods in Fig. 2.

Table 1. Two aging datasets (F.: FACES and L.: Lifespan) with facial expressions and image number of each expression. (Neu.=Neutral, Hap.=Happy, Dis.=Disgust, Ang.=Angry)

D.	# of Images						Total
	Neu.	Hap.	Dis.	Fear	Sad	Ang.	
F.	171	171	171	171	171	171	1026
L.	203	203	–	–	–	–	406

We also notice that the inclusion of aging samples from another database may cause a new problem. Different databases may contain people from different populations and the aging patterns may also be different. In addition, cross-database age estimation is itself a challenging problem [16]. In this paper, we assume that the samples of different databases have similar aging patterns and we do not consider the issues involved in cross-database age estimation.

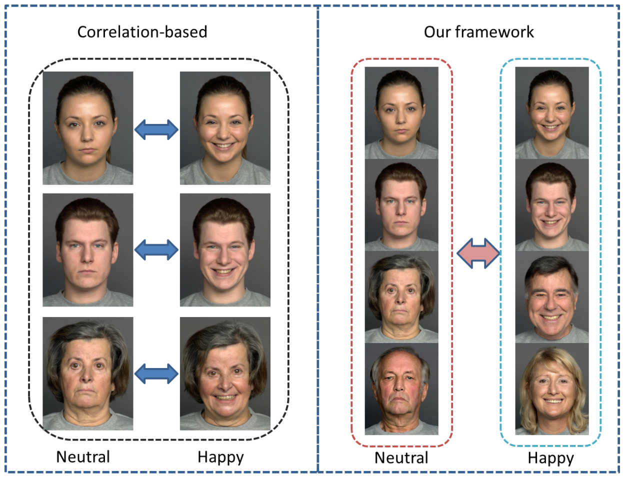

Figure 2. Cross-expression age estimation framework. One characteristic of our framework is that the correspondence are not required, thus it enables us to make use of training data from different datasets to improve the estimation accuracy.

3. Experiments

In this section, we will evaluate the performance of the proposed weighed random subspace method for human age estimation using two datasets: FACES [2] and Lifespan[14]. These two datasets were first used for age estimation in [9]. We show the age estimation results in different cases quantitatively, measured by the mean absolute error (MAE). The MAE is defined by $\frac{1}{N}\sum_{k=1}^{N}|\hat{y}_k - y_k|$, where y_k is the ground truth age for the test image k, $\hat{y}_k$ is the estimated age, and N is the number of test images. We use the same five-fold cross validation in our experiments to make the results comparable.

Table 2. Age estimation under facial expression changes: MAE (in yrs.). "CEDA" indicates BIF feature vectors are processed by CEDA. "[9]" indicates the best results reported in [9]. "V-RS" indicates voting-based random subspace using cross-expression discriminant analysis. "W-RS" means graph-based weighting random subspace using cross-expression discriminant analysis. The percentages in last column are the MAE reduction rate using random subspace methods(Col. 6 or 7) compared to other methods (Col. 4 or 5).

Data	Train	Test	[9]	CEDA	V-RS	W-RS	MAE Red. Rate
FACES	Neutral	Happy	8.66	10.91	9.68	**7.89**	8.9%
		Disgust	10.86	11.60	**10.50**	10.63	2.1%
		Fearful	8.57	12.88	8.90	**8.89**	-3.7%
		Sad	10.54	9.06	**8.32**	9.24	8.2%
		Angry	9.75	11.20	**8.78**	8.86	9.9%
	Happy	Neutral	8.11	9.97	8.53	**7.76**	4.3%
	Disgust		8.57	10.21	8.14	**7.56**	11.8%
	Fearful		9.25	9.79	8.12	**7.90**	14.6%
	Sad		8.66	11.06	**7.35**	7.68	15.1%
	Angry		8.26	8.60	7.34	**6.94**	16.0%
	MEAN		9.12	10.53	8.57	**8.33**	8.6%
Lifespan	Neutral	Happy	7.08	15.40	8.84	**6.37**	10.0%
	Happy	Neutral	6.19	13.10	7.76	**6.09**	1.6%
	MEAN		6.63	14.25	8.30	**6.23**	6.0%

The Lifespan database [14] contains 844 frontal face images with neutral and happy expressions. The FACES database [2] contains 171 individuals with 6 expressions for each person in a frontal view. The Lifespan contains more individuals than FACES, but less number of expressions. We will use both databases for our study. More details about these two databases are given in Table 1.

We use the BIF (Bio-Inspired Features) feature for aging pattern representation [8]. In our experiments, the optimal age estimation results are obtained by cross-validation within range of values for the two parameters: the number of subspaces (NC) and the dimensionality of selected features (t). The parameter ranges of NC and t are [10, 90] and [100, 900], respectively. To verify the effectiveness of our proposed method, we compare it with the state-of-the-art method [9].

The SVMs are used to learn the aging function. The method in [9] used the Partial Least Square (PLS) technique to learn the correlation bases between two different expressions, and then the aging patterns of different expressions can be projected into a common space robust to expression changes. After correlation mapping, a discriminative mapping is used to make the aging patterns more discriminative to improve the age estimation with expression changes.

Table 2 shows the cross-expression age estimation results. We can see that the method using only one subspace determined by CEDA cannot get good results. When the random subspaces are used, the performance is much better than the single subspace. Comparing with the state-of-the-art method, our proposed random subspace framework performs better than [9] in almost all cases, except the case from "Neutral" to "Fearful" in the FACES database. This improves that random subspace based approach can be utilized to reduce the expression changes in age estimation. Furthermore, in 7 out of all 12 cases, our proposed graph-based weighted random subspace method performs better than the voting-based random subspace method. On average, the MAE of our graph-bases fusion is 8.33 years and 6.23 years, in FACES and Lifespan, respectively. Considering the MAE reduction rate, our random subspace based method can improve the result by 8.6% and 6.0%, in FACES and Lifespan, respectively.

Table 3 shows the results before and after using extra data from another dataset. FACES(Lifespan) means using FACES as primary and Lifespan as auxiliary. Lifespan(FACES) means using Lifespan as primary and FACES as auxiliary. In all the 12 cases of two schemes, our proposed weighted random subspace method performs best with lowest MAE value. In two cases, the voting-based method performs better than the weighted version, with quite similar MAE values. Therefore, using extra data, even from a different dataset, can enhance the estimation performance.

4. Conclusions

We have studied the problem of age estimation across different expressions. A weighted random subspace framework is proposed after the construction of cross-expression discriminative subspaces. We have presented a graph-based weighted combination method to obtain the optimal weights for multiple classifiers, which results in the significant improvement of the age estimation performance.

Table 3. Age estimation under facial expression changes using extra data : MAE (in yrs.). The first column indicates the usage of two datasets: "Pri" as primary, and "Aux" as auxiliary. "Pri only" indicates that both training and test data come from "Pri" dataset; "Pri+Aux" means during training, both "Pri" and "Aux" provide training data. "V-RS" indicates voting-based random subspace using cross-expression discriminant analysis. "W-RS" means graph-based weighted random subspace using cross-expression discriminant analysis. The last column "MAE Red. Rate" is the MAE reduction rate before and after using extra data in training. We used the best results in "Pri only" and "Pri+Aux" for the calculation of reduction rates.

Pri (Aux)	Train	Test	Pri only		Pri+Aux		MAE Red. Rate
			V-RS	W-RS	V-RS	W-RS	
FACES (Lifespan)	Neutral	Happy	9.68	7.89	9.22	**6.43**	18.5%
		Disgust	10.50	10.63	11.18	**9.75**	7.1%
		Fearful	8.90	8.89	9.29	**7.67**	13.7%
		Sad	8.32	9.24	8.14	**7.26**	12.7%
		Angry	8.78	8.86	**7.73**	7.92	11.9%
	Happy	Neutral	8.53	7.76	7.49	**6.30**	18.8%
	Disgust		8.14	7.56	7.28	**6.96**	7.9%
	Fearful		8.12	7.90	**6.15**	6.17	22.1%
	Sad		7.35	7.68	7.29	**6.19**	15.8%
	Angry		7.34	6.94	6.92	**6.61**	4.7%
Lifespan (FACES)	Neutral	Happy	8.84	6.37	7.04	**6.26**	1.7%
	Happy	Neutral	7.76	6.09	6.69	**5.99**	1.6%

References

[1] C. Bishop et al. Neural networks for pattern recognition. 1995.

[2] N. Ebner, M. Riediger, and U. Lindenberger. Facesa database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1):351–362, 2010.

[3] Y. Fu, G. Guo, and T. Huang. Age synthesis and estimation via faces: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1955–1976, 2010.

[4] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2234–2240, 2007.

[5] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *Image Processing, IEEE Transactions on*, 17(7):1178–1188, 2008.

[6] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 657–664. IEEE, 2011.

[7] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1986–1991. IEEE, 2009.

[8] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE, 2009.

[9] G. Guo and X. Wang. A study on human age estimation under facial expression changes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2547–2553. IEEE, 2012.

[10] M. Harandi, M. Ahmadabadi, B. Araabi, A. Bigdeli, and B. Lovell. Directed random subspace method for face recognition. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2688–2691. IEEE, 2010.

[11] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18:507, 2006.

[12] T. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[13] G. McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. Wiley-Interscience, 2004.

[14] M. Minear and D. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods*, 36(4):630–633, 2004.

[15] N. Ramanathan, R. Chellappa, S. Biswas, et al. Age progression in human faces: A survey. *Visual Languages and Computing*, 2009.

[16] Y. Su, Y. Fu, Q. Tian, and X. Gao. Cross-database age estimation based on transfer learning. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1270–1273. IEEE, 2010.

[17] V. N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

[18] X. Wang and X. Tang. Random sampling for subspace face recognition. *International Journal of Computer Vision*, 70(1):91–104, 2006.

[19] S. Yan, D. Xu, B. Zhang, and H.-J. Zhang. Graph embedding: A general framework for dimensionality reduction. In *Computer Vision and Pattern Recognition (CVPR), 2005. IEEE Conference on*, volume 2, pages 830–837. IEEE, 2005.