

A Probabilistic Fusion Approach to Human Age Prediction

Guodong Guo
Computer Science
NCCU
Durham, NC, 27707
gdguo@nccu.edu

Yun Fu
Beckman Institute
UIUC
Urbana, IL 61801
yunfu2@uiuc.edu

Charles R. Dyer
Computer Sciences
UW-Madison
Madison, WI, 53706
dyer@cs.wisc.edu

Thomas S. Huang
Beckman Institute
UIUC
Urbana, IL 61801
t-huang1@uiuc.edu

Abstract

Human age prediction is useful for many applications. The age information could be used as a kind of semantic knowledge for multimedia content analysis and understanding. In this paper we propose a Probabilistic Fusion Approach (PFA) that produces a high performance estimator for human age prediction. The PFA framework fuses a regressor and a classifier. We derive the predictor based on Bayes' rule without the mutual independence assumption that is very common for traditional classifier combination methods. Using a sequential fusion strategy, the predictor reduces age estimation errors significantly. Experiments on the large UIUC-IFP-Y aging database and the FG-NET aging database show the merit of the proposed approach to human age prediction.

1. Introduction

Human faces convey a significant amount of semantic information for human-to-human communication. People have the ability to accurately recognize and interpret faces in real time. Facial attributes play a crucial role in applications including multimedia communication and Human-Computer Interaction (HCI). For example, if the user's age can be estimated by a computer, an Age-Specific Human Computer Interaction (ASHCI) system may be developed. Such a system could be used for secure Internet access control in order to ensure that young children do not have access to Internet pages containing adult materials. Or, a vending machine could refuse to sell alcohol or cigarettes to minors [16] [7]. In image and video retrieval, retrieving photographs or videos could be restricted to a required age range [16]. As a kind of semantic knowledge, the age information could be very helpful for multimedia content analysis and understanding.

Image-based automatic age estimation is challenging. The main difficulty is that people age quite differently. The

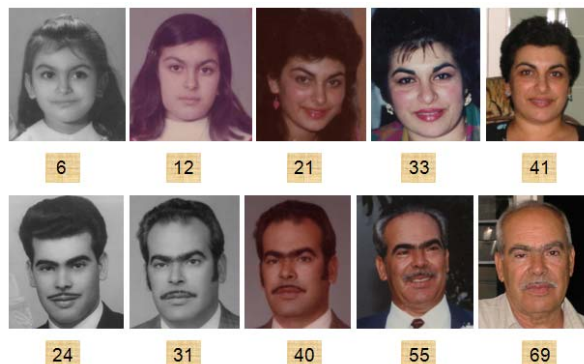


Figure 1. Images of individuals at different ages in FG-NET.

aging process is determined by not only the person's genes but also many external factors such as health, life style, living location and weather conditions. The age progression displayed on faces is uncontrollable and personalized [18, 19, 7]. Males and females may also age differently. Figure 1 shows face images of two people at different ages.

Face-image-based age prediction can be viewed as a constrained pattern recognition problem involving two general steps: feature extraction and recognition. It is constrained because of age continuity, *e.g.*, 0, 1, 2, \dots years, for an individual. Thus a regression model could be used for age prediction. For feature extraction, there are three main approaches: 1) An anthropometric model based on cranio-facial development theory and facial skin wrinkle analysis [14] [19]. The growth related changes of face shape and texture patterns are measured to categorize a face into several age groups. Such methods are suitable for coarse age estimation rather than continuous or refined classification [14] or modelling ages just for young people [19]. 2) Active Appearance Models (AAMs) [3] that incorporate shape and appearance information together [16]. An input face image is represented by a set of fitted model parameters. 3) Sub-space analysis. The AGing pattern Subspace (AGES) method [7] models a sequence of personal aging face im-

ages by learning a subspace representation, handling incomplete data such as missing ages in the sequence. A test face image is projected into the subspace and the corresponding coefficients that best reconstruct the face image are used as features. The Orthogonal Locality Preserving Projections (OLPP) method [2] could also be used to extract age patterns [5]. The OLPP method is appropriate for a large database where an age manifold exists.

Given a set of extracted features, the recognition step can be solved using a classification or regression process. For regression, the Quadratic Model (QM) [15] is usually adopted [5]. A drawback of the QM method is that it may be sensitive to outliers that could come from incorrectly labelled ages. A semi-definite programming formulation was also used for age regression [21] or an EM-based algorithm for regression [20]. For classification, different classifiers can be used for age estimation [16], such as the nearest neighbor classifier, Multilayer Perceptrons (MLPs), or unsupervised Neural Networks.

1.1. Motivation

As mentioned above, age estimation can be viewed as a “special” pattern recognition problem. Each age label can be viewed as a class, thus the age estimation can be thought of as a classification problem. On the other hand, age numbers are a sequence of values, age estimation can be viewed as a regression problem, too. For the purpose of improving the age estimation performance, a natural thought is to combine the classifier and regressor. Very recently, a Locally Adjusted Robust Regressor (LARR) was proposed by Guo *et al.* in [8] to explore this combination. They showed that the combination improves the age estimation performance significantly. However, the LARR method [8] cannot determine the range of local search for the classifier. It has to heuristically try different ranges, such as 4, 8, 16, 32, and 64, and requires the user to choose a best solution among those results. For practical use of the age information, *e.g.*, in multimedia content analysis and understanding, it is important to develop an age information extractor automatically without the user involvement. In other words, the system has to determine the combination parameters *automatically* in a *data-driven* manner. Towards this goal, we interpret the regression and classification results probabilistically in order to fuse them automatically.

In this paper our focus is on developing a probabilistic framework that fuses the regression and classification approaches for age estimation. In Section 2 we present a theoretical framework to derive the decision rule, and in Section 3 we develop a specific fusion strategy. Experiments using two aging databases are reported in Section 4. Section 5 concludes the paper.

2. Theoretical Framework

Consider a pattern recognition problem [4] where pattern Z is to be assigned to one of the m possible labels $L = \{l_1, l_2, \dots, l_m\}$. For the age estimation problem, the labels are human ages (in years), such as $0, 1, \dots$. Assume we have a regressor R and a classifier C , each representing the given pattern by a distinct measurement vector, denoted by \mathbf{x}_R and \mathbf{x}_C , respectively. In the measurement space each label or class l_k is modeled by the probability density function $p(\mathbf{x}_R|l_k)$ or $p(\mathbf{x}_C|l_k)$, and the prior probability of occurrence of each label is denoted by $P(l_k)$.

According to the Bayesian theory, given measurements \mathbf{x}_R and \mathbf{x}_C , the pattern, Z , should be assigned label l_j when the posterior probability of that interpretation is maximum, *i.e.*,

$$l_j = \operatorname{argmax}_{l_k \in L} P(l_k | \mathbf{x}_R, \mathbf{x}_C) \quad (1)$$

The Bayesian decision rule (1) states that all the measurements should be considered simultaneously in order to make a decision utilizing all the available information correctly. The computation of the posterior probability functions in (1) depends on knowledge of high-order measurement statistics described in terms of joint probability density functions $p(\mathbf{x}_R, \mathbf{x}_C | l_k)$, which are generally difficult to obtain. A classical approach to deal with these kinds of joint probabilities is to assume that all the measurements are independent for a given pattern. For example, the mutual independence assumption was used in combining different classifiers in [13].

Here we build a “causal” relation between R and C . Specifically, the classifier C makes decision based on the output of the regressor R , but the regressor R works on the input data directly. Therefore

$$P(\mathbf{x}_R | \mathbf{x}_C) = P(\mathbf{x}_R). \quad (2)$$

There are two reasons to have this causal relation assumption: 1) To reduce the measurement space sequentially – the decisions of the first learner could impact or reduce the measurement space of the second learner. This “early” influence might simplify the the original complex decision problem into a simpler one, and therefore improve the recognition accuracy of the second learner. As a result, the performance of the whole system can be improved. 2) To consider the internal structure of the learners – a regressor usually takes into account all data points, computing in a “global” style, while some modern classifiers, [22] use a pairwise classification scheme, working in a “local” style. Therefore it might be easier to change the measurement space of the classifiers instead of the regressors.

Now let us go back to the Bayesian decision rule (1) and rewrite it. Based on the conditioned Bayes’ rule (*i.e.*, Bayes’ rule conditioned on another variable; see page 10 in

[11]), we have

$$P(l_k | \mathbf{x}_R, \mathbf{x}_C) = \frac{P(\mathbf{x}_R | l_k, \mathbf{x}_C) P(l_k | \mathbf{x}_C)}{P(\mathbf{x}_R | \mathbf{x}_C)} \quad (3)$$

which holds in general. Substituting (2) into (3) we obtain

$$P(l_k | \mathbf{x}_R, \mathbf{x}_C) = \frac{P(\mathbf{x}_R | l_k) P(l_k | \mathbf{x}_C)}{P(\mathbf{x}_R)}. \quad (4)$$

By Bayes' rule, we have

$$P(\mathbf{x}_R | l_k) = \frac{P(\mathbf{x}_R) P(l_k | \mathbf{x}_R)}{P(l_k)}. \quad (5)$$

Plugging (5) into (4), we get

$$P(l_k | \mathbf{x}_R, \mathbf{x}_C) = \frac{P(l_k | \mathbf{x}_R) P(l_k | \mathbf{x}_C)}{P(l_k)}. \quad (6)$$

Now, the decision rule (1) becomes:

$$l_j = \operatorname{argmax}_{l_k \in L} \frac{P(l_k | \mathbf{x}_R) P(l_k | \mathbf{x}_C)}{P(l_k)} \quad (7)$$

subject to constraints (2). Decision rule (7) fuses the posterior probabilities computed by the regressor and the classifier sequentially. We call this a *Probabilistic Fusion Approach* (PFA).

In practice, the denominator of (7), i.e., the prior probabilities $P(l_k)$, will have equal values if no strong prior knowledge is given for a recognition problem. In this case, the decision rule becomes

$$l_j = \operatorname{argmax}_{l_k \in L} P(l_k | \mathbf{x}_R) P(l_k | \mathbf{x}_C) \quad (8)$$

3. Fusion Strategy

Decision rules (7) and (8) constitute the basic scheme for combining a regression measurement with a classification result in a probabilistic way. Now we develop a specific combination strategy based on decision rule (8).

In our sequential probabilistic fusion scheme, the regressor R and classifier C work sequentially so that the output of the regressor, $P(l_k | \mathbf{x}_R)$, is used as an intermediate decision which is then fed to the classifier C to affect the measurement or decision space of the classifier, \mathbf{x}_C . The classifier C has no effect on the regression measurement, \mathbf{x}_R . This causal relation can be depicted by the decision graph in Figure 2.

To realize the decision process shown in Figure 2, several issues have to be addressed, including 1) which methods to use for the regression and classification modules, 2) how to produce the probabilistic output for each method, and 3) how to alter the measurement space of the classifier based on the regression output.

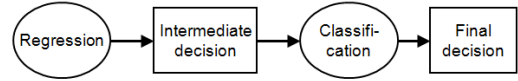


Figure 2. The decision graph of the PFA approach.

3.1. Selection of the Regressor and Classifier

For the regressor, it should have high performance, since its results will influence the decision of the classifier in our sequential fusion strategy. A low performance regressor might “drift” the measurement space badly for the following classifier. The requirement for the classifier is that its measurement space should be able to change (*e.g.*, shrink or expand) easily.

Guided by the above consideration, we chose to use a Support Vector Machine (SVM) [22] as the classifier, and the Support Vector Regression (SVR) method [22] as the regressor, which were also chosen in [8]. The difference is that there is no probabilistic computation for the SVM and SVR in the LARR method [8], while here we transform the results of the SVM and SVR into probabilities and then fuse them automatically without trying different local ranges and requiring users' selection as in [8].

3.2. Probabilistic Output for SVMs

Standard support vector machines provide only an estimated target value, *e.g.*, a category label for classification or a real value for regression. In order to combine the regression and classification measurements probabilistically, probabilities need to be extracted from the standard SVM and SVR results.

For the SVM, some methods have been proposed, mainly in the machine learning literature, to produce probabilistic outputs. For example, Platt [17] proposed a sigmoid training method to post-process standard SVM output, focusing on a two-class classification problem. But it is not clear how to extend this method to a multi-class scenario. In [10], an MAP rule was used on the estimate of the overall posterior probabilities obtained from the outputs of the pairwise classifiers.

Here we adopt a simple yet efficient method to generate a probability estimate for the SVM in a multi-class classification problem, using the counts of occurrences in pairwise comparisons. This simple idea has been used successfully for face recognition [9], for example.

For an n -class classification problem (n could be less than the total number of classes m in the original measurement space), the total number of pairwise comparisons is $n(n-1)/2$. The output of the $n(n-1)/2$ classifiers is used

to construct a matrix as shown below:

$$\begin{pmatrix} 0 & \phi_{1,2} & \phi_{1,3} & \cdots & \phi_{1,n} \\ \phi_{2,1} & 0 & \phi_{2,3} & \cdots & \phi_{2,n} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \phi_{n,1} & \phi_{n,2} & \phi_{n,3} & \cdots & 0 \end{pmatrix}.$$

Each element in the matrix is equal to 1 or 0. $\phi_{i,j} = 1$ if pattern Z is classified as class i in the pairwise competition between classes i and j ; otherwise $\phi_{i,j} = 0$. All elements in the main diagonal are zeros. Based on the measurement matrix, we can create a probability measure for the SVM classifier output as

$$P(l_k|\mathbf{x}_C) = \frac{\sum_{j=1}^n \phi_{k,j}}{\sum_{i=1}^n \sum_{j=1}^n \phi_{i,j}} \quad (9)$$

3.3. Probabilistic Output for the SVR

For the SVR, several methods have been proposed to produce a probabilistic output, but many of them involve either complex computations or modification of the SVR formulation. For example, a Gaussian process is integrated into the SVR to formulate a Gaussian SVM regression model in [6]. A Gaussian (or Laplace with fatter tails) distribution, could be used to approximate the probabilistic outputs for SVRs. However, the Gaussian approximation may encounter problems in practice, especially in human age prediction, because of the diversity of aging variations. Each individual may age in his or her own way and be affected by many different external factors.

As pointed out in [8], the ages estimated by the SVR method could be far away from the true age labels. Consequently, a small probability value (possibly close to zero) could be generated for a true age label when a Gaussian model is used for transforming the SVR target values into probabilistic outputs. This would inhibit a correct decision when multiplying the two probabilities in the decision rule (7) or (8). In order to avoid such undesirable effects, we propose to use a uniform distribution centered at the estimated target value, l_0 , obtained from a regressor, i.e., $\mu = l_0$. In fact, we found that the Gaussian model gave much worse results than the uniform distribution in our initial experiment on age estimation which is not shown here.

The uniform distribution model assumes that only a finite range of age labels is possible, each with equal probability. The Probability Density Function (PDF) of the uniform distribution $U(\mu - \Delta, \mu + \Delta)$ is given by

$$p(x) = \begin{cases} \frac{1}{2\Delta} & \text{for } \mu - \Delta \leq x \leq \mu + \Delta, \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where $[\mu - \Delta, \mu + \Delta]$ is the function support. Now the question is how to estimate the range of support for the uniform distribution.

Let us look at the SVR prediction error or residual, ζ_i , with $\zeta_i = l_i - \hat{f}(Z_i)$, where l_i is the true age label for pattern Z_i , and $\hat{f}(Z_i)$ is the regression estimate. Recall that the variance of the uniform distribution satisfies $\sigma^2 = \frac{1}{12}(2\Delta)^2$, i.e., $\sigma^2 = \frac{1}{3}\Delta^2$, so we have $\Delta = \sqrt{3}\sigma$. Thus the function support can be estimated by the sample standard deviation. To compute the sample standard deviation, σ , we can collect the residuals, ζ_i , on a validation data set, and then compute the standard deviation of these residuals. Finally, we have

$$P(l_k|\mathbf{x}_R) \leftarrow U(l_0 - \sqrt{3}\sigma, l_0 + \sqrt{3}\sigma). \quad (11)$$

The uniform distribution (11) is simple but works well in our experiments. To our knowledge, no previous work uses it to model the probabilistic output of a regressor such as the SVR.

3.4. Decision Space Deduction

Given the probabilistic outputs, $P(l_k|\mathbf{x}_R)$ and $P(l_k|\mathbf{x}_C)$, for the regressor and classifier, respectively, the next step is to combine the two probabilities together to make a final decision for a given pattern. According to the decision rule (8), the two probabilities are multiplied and the label l_j corresponding to the maximum product is selected as the final decision.

Our serial probabilistic fusion approach can also be interpreted as a decision space deduction process. The uniform distribution modeling of the probabilistic output of the regressor reduces the original label space (all possible ages) into a smaller decision space, $[l_0 - \sqrt{3}\sigma, l_0 + \sqrt{3}\sigma]$, by using the cut-off boundaries. The reduced decision space is refined by the classifier to obtain the final decision, l_j . As a result, the probabilistic output of the SVR plays the role of an intermediate decision, as shown in Figure 2, reducing the search space (i.e., less number of classes to compare) for the classifier SVM. The LARR method [8] shares the same spirit as the PFA in terms of decision space deduction, however, it does not address the probabilities for automatic local range determination.

4. Experiments

The age prediction experiments were performed on two age databases. One is the public available FG-NET aging database [1]. It contains 1,002 face images of 82 individuals with ages ranging from 0 to 69 years. Each face image has 68 labelled points characterizing shape features. The AAM features with 200 parameters is used to model each face [7] [21] [20] [8]. To evaluate our PFA method on FG-NET, we

used the same AAMs for feature extraction as in [7] [21] [20] [8].

The other is the UIUC-IFP-Y aging database. It contains 8,000 high-resolution RGB color face images captured from 1,600 different Asian subjects, 800 females and 800 males, in the age range from 0 to 93 years. Each subject has 5 near frontal images with ground truth ages. To our knowledge, this human age database is the largest one ever reported. Faces were detected, cropped and resized to 60×60 grayscale images, with eye corner locations labeled for each subject. The gray-level values were normalized to reduce the effect of outdoor illumination changes. The OLPP method [2] was applied to the training face patches to learn an age manifold. Then each face image was projected onto the age manifold to extract a feature vector. We used the first 150 features for each face image as in [5] [8].

In SVR and SVM learning, related parameters were determined using a tuning or validation set, which is part of the learning data. The SVM-light package [12] was used for SVM and SVR evaluation. For classification, pairwise linear SVM classifiers were used to deal with the multi-class classification problem and learn the probabilistic outputs based on Eqn. (9). For fusion, the standard deviations of the residuals were estimated using the same tuning set as in choosing the SVR and SVM parameters, and then used to determine the support of the uniform distribution given in (11). A Gaussian distribution was also tried for modeling the residuals of the SVR outputs, but the age estimation results (not reported here) are much worse than using the uniform distribution. Given the determined uniform distribution, the classifiers were then used on a reduced decision space to make the final decision.

Two measures can be used to measure the performance of age estimation: the Mean Absolute Error (MAE) and the Cumulative Score (CS). The MAE [16] [15] [7] [21] [8] is defined as the average of the absolute errors between the estimated ages and the ground truth ages, $MAE = \sum_{k=1}^N |\hat{l}_k - l_k|/N$, where l_k is the ground truth age for the test image k , \hat{l}_k is the estimated age, and N is the total number of test images. The cumulative score [7] is defined as $CS(i) = N_{e \leq i}/N \times 100\%$, where $N_{e \leq i}$ is the number of test images on which the age estimation makes an absolute error no greater than i years.

To evaluate the accuracy of our PFA method on the UIUC-IFP-Y age database, we performed a 4-fold cross-validation test. The test was executed on the female and male subsets separately. The reason for this is that females and males aged quite differently in this database [5] [8]. Women tend to use anti-aging makeup or medication to cover up their true ages. For the FG-NET aging database, a test strategy called Leave-One-Person-Out (LOPO), has often been used [7] [21] [20] [8]. We followed the same strategy and compared our results with previous ones.

Table 1. MAE comparisons with others.

Method	U-I-Y/F	U-I-Y/M	FG-NET
WAS [7]	–	–	8.06
AGES [7]	–	–	6.77
QM [16]	9.96	10.51	6.55
MLPs [16]	10.99	12.00	6.98
RUN 1 [21]	9.79	10.36	5.78
RUN 2 [20]	6.95	6.95	5.33
SVR [8]	7.00	7.47	5.16
SVM [8]	5.55	5.52	7.16
LARR [8]	5.25	5.30	5.07
PFA(Ours)	5.11	5.12	4.97

Experimental results are shown in Table 1 and Figure 3. The first and second columns in Table 1 show the MAEs for females and males in the UIUC-IFP-Y aging database, respectively. The last column shows the MAEs on the FG-NET aging database. Results using pure SVR learning and SVM classification [8] are also shown in the table to demonstrate the significant improvement of our PFA.

The PFA method produced an MAE of 5.11 years for females and 5.12 years for males in the UIUC-IFP-Y database, smaller than the 5.55 and 5.52 years produced by the SVM method, and much smaller than the 7.00 and 7.47 years by the SVR method [8]. For the FG-NET aging database, the PFA method gives an MAE of 4.97 years, smaller than the 5.16 years of the SVR, and much smaller than the 7.16 years of the SVM method. This demonstrates that our PFA is a powerful tool to combine a classifier and a regressor for human age prediction.

We also compared PFA with all other previous methods reported on the UIUC-IFP-Y aging database in Table 1. The MAEs of our PFA method are significantly smaller than most of previous approaches. For example, our method achieves about a 26% reduction of MAEs over the RUN 2 method [20], and is lower than the LARR method [8], too. On the FG-NET aging database, our PFA method has an MAE of 4.97 years, which is a reduction of about 7% over the 5.33 of the Run 2 method [20], and is lower than the 5.07 of the LARR method [8]. Overall, our PFA method gives much better results than the Run 1 [21], Run 2 [20], QM [16], and MLPs [16], and slightly better than the LARR method [8]. In terms of the LARR and PFA, remember that here the PFA presents a theoretical framework for automatic fusion based on probabilities, while the LARR requires the user’s selection of local ranges which is not trivial in practice.

Comparisons of cumulative scores are shown in Figure 3. Our PFA method performs much better than the QM, MLP, Run 1 and Run 2 methods on the UIUC-IFP-Y database shown in Figure 3 (a) and (b), and on the FG-NET aging

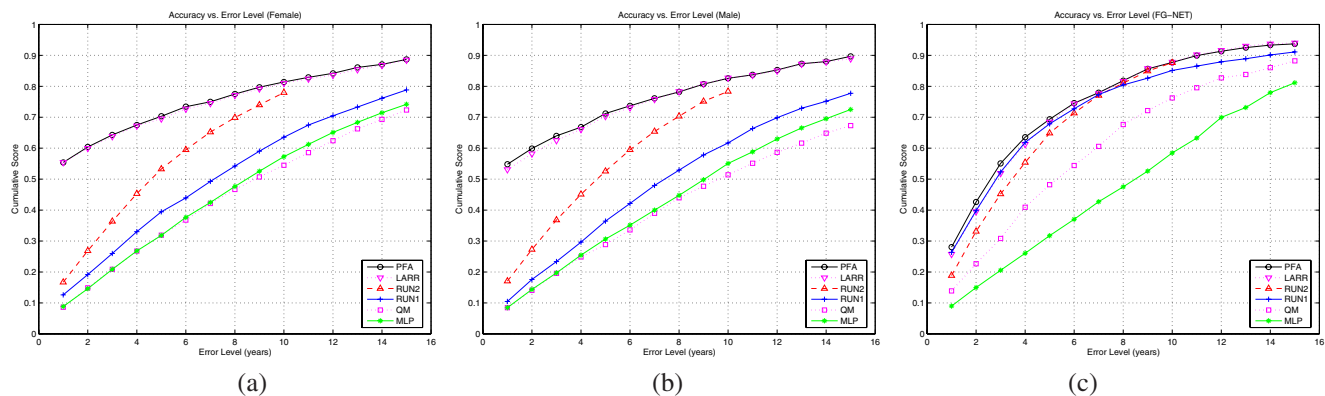


Figure 3. Comparisons between our PFA method and other state-of-the-art methods in terms of the cumulative scores for (a) female age estimation, (b) male age estimation on the UIUC-IFP-Y database, and (c) age estimation on the FG-NET aging database.

database shown in Figure 3 (c). The cumulative scores of the PFA are slightly better than the LARR especially at low error levels.

From the experiments we conclude that the probabilistic fusion approach gives better age estimation than previous approaches. Our PFA method combines the merits of each model probabilistically in a unified and consistent way. In contrast to the LARR method [8], our PFA method fuses the classifier and regressor *probabilistically* and *automatically*.

5. Conclusion

We have presented a novel procedure for human age prediction. The theoretical framework is derived based on the conditioned Bayes' rule. By transforming the SVR and SVM outputs into probabilities, a serial probabilistic fusion scheme has been developed for combining a regressor and a classifier systematically. Experiments on the UIUC-IFP-Y and FG-NET aging databases show that our PFA method performs better than all previous approaches.

References

- [1] The FG-NET Aging Database, <http://sting.cycollege.ac.cy/~alanitis/fgnetaging/index.htm> 4
- [2] D. Cai, X.F. He, J.W. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Trans. on Image Processing* 15(11), pp. 3608-3614, 2006. 2, 5
- [3] T.F.Cootes, G.J. Edwards and C.J.Taylor. "Active appearance models", in *Proc. ECCV* Vol. 2, pp. 484-498, 1998. 1
- [4] R.O. Duda, P.E. Hart, and D. Stork, *Pattern Classification*, 2nd ed., Wiley Interscience, 2000. 2
- [5] Y. Fu, Y. Xu, and T. S. Huang, "Estimating human ages by manifold analysis of face pictures and regression on aging features," *Proc. ICME*, pp. 1383-1386, 2007. 2, 5
- [6] J. B. Gao, S. R. Gunn, C. J. Harris, and M. Brown, "A probabilistic framework for SVM regression and error bar estimation," *Machine Learning* 46, pp. 71-89, 2002. 4
- [7] X. Geng, Z.-H. Zhou and K.S.-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. PAMI*, 29(12), pp. 2234-2240, 2007. 1, 4, 5
- [8] G. Guo, Y. Fu, T. Huang and C. Dyer, "Locally adjusted robust regression for human age estimation," *IEEE Workshop on Application of Computer Vision*, 2008. 2, 3, 4, 5, 6
- [9] G. Guo, H. Zhang, and S. Li, "Pairwise face recognition," *Proc. ICCV*, Vol. 2, pp. 282-287, 2001. 3
- [10] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Advances in NIPS* 10, pp. 507-513, 1998. 3
- [11] F. V. Jensen and T. D. Nielsen, *Bayesian networks and decision graphs*, 2nd ed., Springer, 2007. 3
- [12] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges and A. Smola, eds., 1999. 5
- [13] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *PAMI*, 20(3), pp. 226-239, 1998. 2
- [14] Y. Kwon and N. Lobo, "Age classification from facial images," *CVIU* 74(1), pp. 1-21, 1999. 1
- [15] A. Lanitis, C. Taylor, and T. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. PAMI* 24(4), pp. 442-455, 2002. 2, 5
- [16] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. SMC-B* 34(1), pp. 621-628, 2004. 1, 2, 5
- [17] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, pp. 61-74, 1999. 3
- [18] N. Ramanathan and R. Chellappa, "Face verification across age progression," *IEEE TIP* 15(11), pp. 3349-3361, 2006. 1
- [19] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," *Proc. CVPR*, pp. 387-394, 2006. 1
- [20] S. Yan, H. Wang, T. S. Huang, and X. Tang, "Ranking with uncertain labels", *Proc. ICME*, 2007. 2, 4, 5
- [21] S. Yan et al., "Learning auto-structured regressor from uncertain nonnegative labels", *Proc. ICCV*, 2007. 2, 4, 5
- [22] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998. 2, 3