Chapter 8
Soft Biometrics from Face Images Using Support Vector Machines

Guodong Guo

Abstract Soft biometrics, such as age, gender, and ethnicity, are useful for many applications in practice. For instance, in business intelligence, it is helpful to automatically extract and compute the statistics of potential customers, such as the number of males and females; the number of young, adult, and senior people; or the number of Caucasian, African American, or Asian people. It is also helpful to use soft biometrics to improve the performance of traditional biometrics for human identification, such as face recognition. Different methods can be developed to recognize the soft biometric characteristics from face images. In this chapter, we present the application of the support vector machines (SVM) to learn an estimator or recognizer to extract these soft biometrics. We will mainly focus on age estimation, while the gender and ethnicity classification will also be discussed. Both classification and regression will be considered. The combination of regression and classifiers based on the SVM will also be described which is useful especially for age estimation.

8.1 Introduction

Support vector machines (SVM) [41] have shown many successful applications in a variety of areas, including computer vision, pattern recognition, image analysis, biometrics, bioinformatics, etc. The SVMs are graceful in theory (e.g., the large margin optimization and mathematical programming solver) and have good performance in

G. Guo (✉)
Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA
e-mail: guodong.guo@mail.wvu.edu

Y. Ma and G. Guo (eds.), *Support Vector Machines Applications*,
DOI 10.1007/978-3-319-02300-7_8,

practice (e.g., good generalization capability and high discriminative power). In this chapter, we show the use of the SVMs and their extensions to extract soft biometric characteristics from face images.

The typical soft biometric characteristics include age, gender, and ethnicity. These measures cannot be used to identify a person uniquely. For instance, different subjects can share the same age, gender, or even the same ethnicity. However, when two subjects are confused by a face recognizer, the age, gender, and/or ethnicity might be used to help the discrimination, assuming the two subjects have some differences in those measures. For example, we have shown that the soft biometrics, such as gender, ethnicity, weight, and height can help to improve the face recognition performance [27]. So those characteristics are called "soft biometrics," while the traditional biometric cues, such as face, iris, and fingerprints, are assumed to be unique for each individual.

In addition to helping human identification, the soft biometrics themselves are also useful for other applications. A typical case is business intelligence, where there is no need to know the identities of the customers. The real care is the statistics of the group of customers, such as the number of males and females, young, adult or senior people, and Caucasian or Asian. These soft biometric characteristics can help the business owners or managers to know more about the potential customers, do a better advertisement to the related customers, or introduce commercial products to the appropriate customers who might be interested in those products.

Among the three soft biometric characteristics, age estimation is probably the most challenging problem. Our primary focus here is the age estimation, while we will also consider gender and ethnicity classification. Further, age estimation is a very special problem. The age labels, e.g., 1, 2, 3 in years, can be considered as regression values, thus age estimation can be taken as a regression problem. On the other hand, each age label can also be considered as a separate class, thus age estimation can also be taken as a classification problem [21, 23]. We study the performance of the SVM-based classification and regression for age estimation on different databases. We also present a scheme to combine the regression and classifiers for an improved performance on age estimation [21, 23]. Further, a probabilistic fusion is also presented to make the combination automatic without much parameter adjustment [22].

For gender and ethnicity classification, we show the performance of the SVM classifiers on large databases. We also present a study of whether the gender and ethnicity classification is affected by age or not [15, 24].

Soft biometric characteristics have other measures, in addition to age, gender, and ethnicity. For instance, we have recently developed a computational approach to body mass index (BMI) prediction in face images [43]. We believe that more and more soft biometric cues can be extracted along with practical applications. In this chapter, we just study the most popular soft biometrics, i.e., age, gender, and ethnicity.

In the following, we briefly introduce the support vector regression (SVR) in Sect. 8.2 and the SVM in Sect. 8.3. Then in Sect. 8.4 we present a method, called locally adjusted robust regression (LARR), to combine the SVR and SVM for an improved age estimation. In Sect. 8.5 we describe a probabilistic fusion to combine the SVM and SVR. Some simple introduction of the face image representation is presented in Sect. 8.6. The experiments are conducted in Sect. 8.7, and finally, we draw conclusions.

8.2 Support Vector Regression

The basic idea of SVR is to find a function $f(\mathbf{y})$ that has most ε deviation from the actually obtained target z_i for the training data $\mathbf{y}_i$, and at the same time is as flat as possible [41]. In other words, we do not care about the errors as long as they are less than ε. This property determines the SVR to be less sensitive to outliers than the quadratic loss function. In comparison with the conventional quadratic loss function shown in Fig. 8.1a, the ε-insensitive loss function of SVR is shown in Fig. 8.1b. Given the same input, the ε-insensitive loss function is more robust than the quadratic function in dealing with outliers.

8.2.1 *Linear SVR*

Consider the problem of approximating the set of data $\mathscr{D} = \{(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_n, z_n)\}$, $\mathbf{y}_i \in \mathbb{R}^d, z_i \in \mathbb{R}$, with a linear function,

$$f(\mathbf{y}) = \langle \mathbf{w}, \mathbf{y} \rangle + b. \tag{8.1}$$

The optimal regression function [41] is given by

$$\min_{\mathbf{w},\xi} \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{n} (\xi_i^+ + \xi_i^-)$$

$$z_i - \langle \mathbf{w}, \mathbf{y}_i \rangle - b \leq \varepsilon + \xi_i^+$$

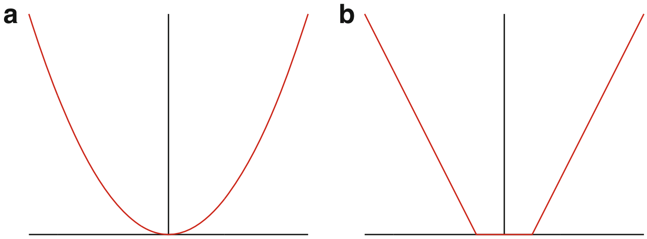

Fig. 8.1 Regression criteria. (**a**) Quadratic regression loss function. (**b**) ε-insensitive loss function which is less sensitive to outliers than the quadratic loss function. Another benefit from this function is a sparse set of support vectors to represent the regression function, i.e., only points outside the ε zone contribute to the regression function. The horizontal and vertical axes are $\mathbf{y}$ and $f(\mathbf{y})$, respectively

$$\text{subject to} \quad \langle \mathbf{w}, \mathbf{y}_i \rangle + b - z_i \leq \varepsilon + \xi_i^-$$
$$\xi_i^+, \xi_i^- \leq 0 \tag{8.2}$$

where constant $C > 0$ determines the trade-off between the flatness of f and data deviations, and ξ_i^+, ξ_i^- are slack variables to cope with otherwise infeasible constraints on the optimization problem of (8.2). The ε-insensitive loss function as shown in Fig. 8.1b is

$$L_\varepsilon(\mathbf{y}, z) = \begin{cases} 0, & \text{if } |f(\mathbf{y}) - z| < \varepsilon \\ |f(\mathbf{y}) - z| - \varepsilon, & \text{otherwise} \end{cases} \tag{8.3}$$

The *primal* problem of (8.2) can be solved more efficiently in its *dual* formulation [41] resulting in the final solution given by

$$\mathbf{w} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \mathbf{y}_i, \tag{8.4}$$

and

$$f(\mathbf{y}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \mathbf{y}_i, \mathbf{y} \rangle + b, \tag{8.5}$$

where α_i, α_i^* are Lagrange multipliers. The value of b in Eq. (8.1) can be determined by plugging Eq. (8.4) into Eq. (8.1) [12].

8.2.2 A Toy Example

To illustrate the SVR idea and see the importance of proper setting of the parameter ε, we use a toy example that contains 30 points in 2D with 10 in a line and the remaining 20 being outliers distributed on both sides of the line [20]. Hence the data contains 67 % outliers. Using the SVR algorithm implemented by Gunn [12] (which provides a user interface) and a linear kernel with $\varepsilon = 0.02$, the result is shown in Fig. 8.2a. Observe that the line was correctly estimated despite the high percentage of outliers.

On the other hand, observe that SVR returns 27 support vectors (90 % of the input data) and seven of them are very close to the boundaries (two dashed lines), but there are actually 20 outliers in the original data. So we cannot simply classify the support vectors (SVs) as outliers. Increasing the ε value might "drag" the seven closest support vectors inside the dashed boundaries, and then only the outliers in the data would be returned as support vectors. However, when we increase ε gradually up to 0.09, there are still 26 SVs returned which are still not the true outliers, as

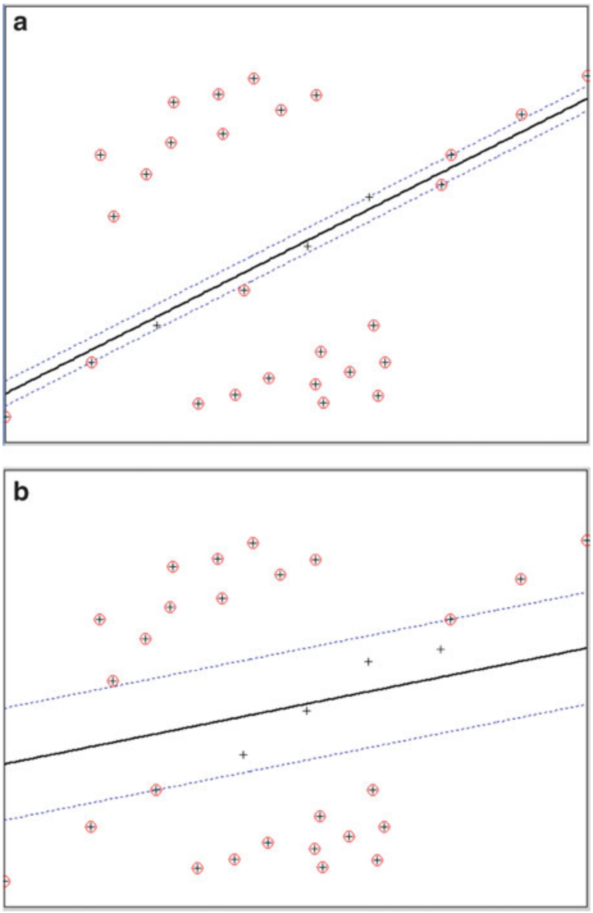

Fig. 8.2 SVR on real 2D data with $\varepsilon = 0.02$ in (**a**) and $\varepsilon = 0.09$ in (**b**). Note that the support vectors (marked by *circles*) are not the true outliers in either case

shown in Fig. 8.2b. And even worse, the slope of the line has changed significantly. This demonstrates that using a large ε is not a good idea because it may degrade the model structure.

Based on this experiment, we observe: (1) the SVR technique can potentially deal with data containing a high percentage of outliers; (2) classifying support vectors as outliers is not workable; (3) using a large value for ε is not a good idea for SVR; and (4) using small ε is preferable, especially when a large number of outliers are present.

This toy example and the above observations were first presented by Guo et al. in [20]. The robust regressor, SVR, was applied successfully for outlier detection and removal in affine motion tracking with the setting of a small ε. Here we adopt the same idea but use it for another application—robust age regression. Instead of using the simple linear regression, we need a nonlinear SVR for the complex aging patterns.

8.2.3 Nonlinear SVR

A nonlinear regression function may be required in practice to adequately model the data. It can be obtained by using kernels, in the same manner as a nonlinear SVM for classification [41]. A nonlinear mapping can be used to map the data into a high dimensional feature space where a linear regression is performed. Different kernels, such as polynomials, sigmoid, or Gaussian radial basis functions, can be used depending on the tasks. For our robust age regression, we found that the Gaussian radial basis function kernel performs much better than the linear regression [21, 23]. The reason is that the linear regression cannot model the complex aging process. A radial basis function is of the form,

$$k(\mathbf{y},\mathbf{y}') = e^{-\gamma\|\mathbf{y}-\mathbf{y}'\|^2}, \tag{8.6}$$

where γ is a constant to adjust the width of the Gaussian function. Given the kernel mapping, the solution of the nonlinear SVR is obtained as [41],

$$\langle \mathbf{w},\mathbf{y}\rangle = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(\mathbf{y}_i,\mathbf{y}), \tag{8.7}$$

and

$$f(\mathbf{y}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(\mathbf{y}_i,\mathbf{y}) + b. \tag{8.8}$$

The difference to the linear regression is that $\mathbf{w}$ is no longer given explicitly. Also note that in the nonlinear case, the optimization problem corresponds to finding the *flattest*, or linear regression function in the higher dimensional *feature* space,[1] not in the input space.

[1] Note that the feature space means a higher dimensional space in SVR, which is different from the feature extracted from data in image processing. Actually the extracted features from images are the input data for SVR in our age modelling.

8.3 Support Vector Machine

SVM [41] are a class of classifiers that can learn an optimal separating hyperplane based on the maximum margin criterion. It can use different kernels to make the linear SVM work on a higher dimensional space to improve the separability between two classes. The kernel extension is similar to the SVR learning. In the following, we only briefly introduce the linear SVM. More details on the kernel SVMs can be referred to [41].

8.3.1 Linear SVM

Given a set of training vectors belong to two separate classes, $(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_n, z_n)$, where $\mathbf{y}_i \in \mathbb{R}^D$, $z_i \in \{-1, +1\}$, the linear SVM learns an optimal separating hyperplane, $\mathbf{wy} + b = 0$, that maximizes the margin [41]. The SVM learning is to find the saddle point of the Lagrange functional,

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \parallel \mathbf{w} \parallel^2 - \sum_{i=1}^{n} \alpha_i \left\{ z_i \left[(\mathbf{w} \cdot \mathbf{y}_i) + b \right] - 1 \right\} \tag{8.9}$$

where α_i are the Lagrange multipliers. The Lagrangian has to be minimized with respect to $\mathbf{w}$, b and maximized with respect to $\alpha_i \geq 0$. The optimization is usually transformed to its *dual* problem,

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left\{ \min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha) \right\}, \tag{8.10}$$

and the optimal hyperplane is represented by the dual solution, α,

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i z_i \mathbf{y}_i \tag{8.11}$$

The value of b can be estimated by plugging $\mathbf{w}$ into the original equation, $\mathbf{wy} + b = 0$.

In testing, the classification is given by

$$f(\mathbf{y}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{y} + b), \tag{8.12}$$

for any new data point $\mathbf{y}$. If the training data are non-separable, slack variables ξ_i can be introduced. See [41] for more details.

8.4 Locally Adjusted Robust Regression

Age estimation can be considered as a regression problem. Now, a question may be asked, is it "good" enough to use the SVR as a robust regressor for human age prediction? To answer this question, let us look at an estimation result using the SVR [21]. Figure 8.3 shows the predicted ages (red squares) with respect to the ground truth ages (black circles). Note that this is not a regression curve. One thousand data points are sorted in ascending order of the ground truth ages, i.e., from 0 to 91 years for females. The predicted ages are obtained from the SVR method. From this figure, we observe that the SVR method can estimate the global age trend, but cannot predict the ages precisely. By inspecting the result carefully, we find that the SVR predictions give bigger age values for many younger people, and smaller age values for some older people. In some cases, the estimated age values could be far away from the true ages, e.g., more than 40 years. This result was based on a database used in [21].

Why the SVR method cannot show better performance than we expect for age prediction? The reason can be in two aspects: First, the problem of age prediction is really challenging because of the diversity of aging variation. Each individual may age in his/her own way and be affected by external factors, such as health, living condition, and exposure to weather conditions. Second, the SVR method attempts to find a flat curve to approximate the data in order to obtain good generalization capability. As shown in Fig. 8.4, the SVR computes a flat curve within a small ε tube. But the age data may distribute like the (green) irregular curve. One cannot expect the SVR to estimate an irregular curve like this because of the over-fitting

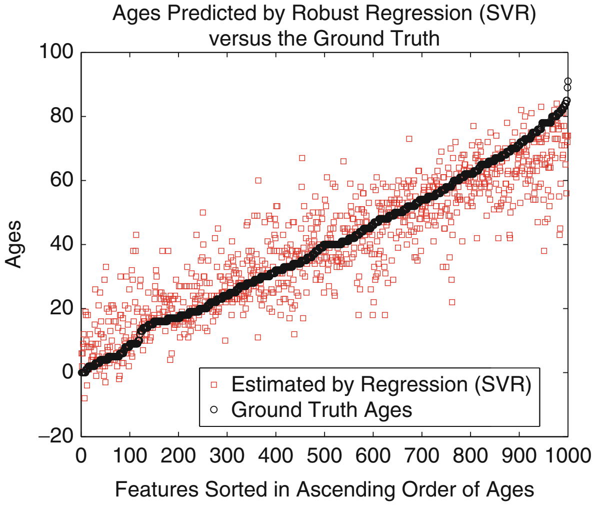

Fig. 8.3 A plot of the true ages (*black circles*) versus the estimated ages (*red squares*) for one thousand female face images. The ages are predicted by the nonlinear SVR with a Gaussian kernel

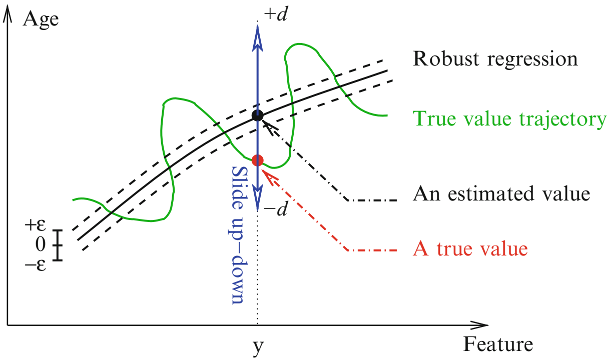

Fig. 8.4 Illustration of the idea of locally adjusted robust regression (LARR)

problem. Further, one cannot assign a large ε to enclose all true data points inside the ε tube, as demonstrated in the toy example in Sect. 8.2.2. So how to model the aging function by allowing the irregular distribution of true ages?

8.4.1 Local Adjustment of the Regression Result

One feasible solution is to adjust the age regression values locally so that the estimated age values can be "dragged" towards the true ages. We call it a locally adjusted robust regression (LARR) [21, 23]. The idea of LARR is illustrated in Fig. 8.4. Suppose the predicted age value by SVR is $f(\mathbf{y})$, corresponding to the input data $\mathbf{y}$. The point $f(\mathbf{y})$ is displayed by the black dot on the regression curve. The estimated age, $f(\mathbf{x})$, may be far away from the true age value, L, shown as the red dot on the true age trajectory curve. The idea of the LARR method is to slide the estimated value, $f(\mathbf{y})$, up and down (corresponding to greater and smaller age values) by checking different age values, $t \in [f(\mathbf{y}) - d, f(\mathbf{y}) + d]$, to see if it can come up with a better age estimation. The value d indicates the range of ages for local search. Hopefully the true age value, L, is also within this range, i.e., $L \in [f(\mathbf{y}) - d, f(\mathbf{y}) + d]$.

Therefore the LARR method is a two-step procedure: (1) a robust regression over all ages of the training data by using the SVR method. This step can be considered as a global regression process; (2) a local adjustment within a limited range of ages centered at the regression result.

Now the key issue is how to verify different age values within a specified range for the purpose of local adjustment. Remember our goal is to "drag" the initially estimated age value, $f(\mathbf{y})$, by the global regressor, towards the true age, L, as close as possible. We take a classification approach to locally adjust or verify different ages, considering each age label as one class. Because only a small number of age labels are used for each local adjustment, regression methods cannot work properly.

For our classification-based local adjustment, there are many possible choices of classifiers, but here we adopt a linear SVM for our local age adjustment. The main reason is that the SVM can learn a classifier given a small number of training examples. This has been demonstrated by the author previously for learning in the small sample case, such as face recognition [16, 18], image retrieval [19], audio classification and retrieval [14], and face expression recognition [13]. The capability of learning a classifier in the small sample case is also important for human age prediction. Usually the number of training examples, e.g., 50, is smaller than the feature dimension, e.g., 150, in age estimation, even though we perform experiments on a large database (see Sect. 8.7 for details).

8.4.2 Binary Tree Search with Limited Range

The classical SVMs are designed to deal with the two-class classification problems. There are three typical ways to extend it to a multi-class classification application. (1) Learning classifiers for each pair of classes, and taking a binary tree search in testing; (2) training SVMs for each class against all the remaining classes; and (3) training SVMs for all classes simultaneously. The last two schemes are not appropriate for our purpose, because in the local adjustment only partial classes of age data are involved. If the last two schemes are used, the SVMs have to be re-trained dynamically for each adjustment, which is computationally expensive. The first scheme is feasible to fulfill our task since there is no need to retrain the SVMs online. Therefore, all pair-wise SVM classifiers can be trained off-line. Only a limited number of classes are involved in the binary tree search for test.

The binary tree structure for multi-class SVM classification has been used successfully in previous research, e.g., face recognition [16]. In general, the number of pair-wise comparisons is $n_c - 1$ for each test in an n_c-class classification problem. Here the number of pair-wise comparisons is limited to $m_c - 1$ when only m_c classes are involved in each local adjustment, and $m_c < n_c$. Each age corresponds to one class label.

8.4.3 Local Search Range Determination

The local search range, m_c, is determined by several factors, such as the scale of the data (large versus small scale) and the performance of the robust regressor (here the kernel SVR).

It is not trivial to determine the local search range. There are some guidelines for choosing local search ranges. The larger the search range, the bigger the chance to contain the true ages within that range. If the search range is too small, the true age label might not be reached and the local search may find an arbitrary age label.

On the other hand, if the search range is too big, it also increases the possibility to obtain an adjusted age that is far away from the true age, because the local classification is just a locally optimal search.

In our experiments, we specify different ranges and demonstrate the effects of different local search ranges on the results [21]. The main goal is to show that the local adjustment can really improve the performance over the robust regressor for human age estimation.

8.5 Probabilistic Fusion of the SVR and SVM

As presented above, the combination of the SVR and SVM can take advantage of both classifiers and regression for age estimation. Our first scheme is a LARR proposed by Guo et al. in [21, 23]. It has been shown that the age estimation performance can be improved significantly by using the LARR method.

However, the LARR method 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 [22].

8.5.1 Theoretical Framework

Consider a pattern recognition problem [42] where pattern Z is to be assigned to one of the m possible labels $L = \{l_1, l_2, \cdots, l_m\}$. For the age estimation problem, the labels are human ages (in years), such as $0, 1, \cdots$. 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 (PDF) $p(\mathbf{x}_R|l_k)$ 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.,

$$l_j = \arg\max_{l_k \in L} P\left(l_k | \mathbf{x}_R, \mathbf{x}_C\right) \tag{8.13}$$

The Bayesian decision rule (8.13) 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 (8.13) depends on knowledge of high-order measurement statistics described in terms of joint PDFs $p\left(\mathbf{x}_R, \mathbf{x}_C | l_k\right)$, 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 [31].

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

$$P\left(\mathbf{x}_R | \mathbf{x}_C\right) = P\left(\mathbf{x}_R\right). \tag{8.14}$$

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 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 [41] 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 (8.13) and rewrite it. Based on the conditioned Bayes' rule (i.e., Bayes' rule conditioned on another variable; see page 10 in [30]), we have

$$P\left(l_k | \mathbf{x}_R, \mathbf{x}_C\right) = \frac{P\left(\mathbf{x}_R | l_k, \mathbf{x}_C\right) P\left(l_k | \mathbf{x}_C\right)}{P\left(\mathbf{x}_R | \mathbf{x}_C\right)} \tag{8.15}$$

which holds in general. Substituting (8.14) into (8.15) we obtain

$$P\left(l_k | \mathbf{x}_R, \mathbf{x}_C\right) = \frac{P\left(\mathbf{x}_R | l_k\right) P\left(l_k | \mathbf{x}_C\right)}{P\left(\mathbf{x}_R\right)}. \tag{8.16}$$

By Bayes' rule, we have

$$P\left(\mathbf{x}_R | l_k\right) = \frac{P\left(\mathbf{x}_R\right) P\left(l_k | \mathbf{x}_R\right)}{P\left(l_k\right)}. \tag{8.17}$$

Plugging (8.17) into (8.16), we get

$$P\left(l_k | \mathbf{x}_R, \mathbf{x}_C\right) = \frac{P\left(l_k | \mathbf{x}_R\right) P\left(l_k | \mathbf{x}_C\right)}{P\left(l_k\right)}. \tag{8.18}$$

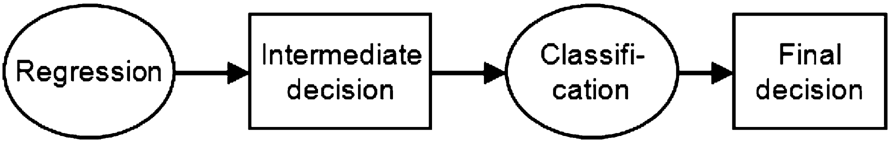

Fig. 8.5 The decision graph of the PFA approach

Now, the decision rule (8.13) becomes:

$$l_j = \arg\max_{l_k \in L} \frac{P(l_k|\mathbf{x}_R)\,P(l_k|\mathbf{x}_C)}{P(l_k)} \tag{8.19}$$

subject to constraints (8.14). Decision rule (8.19) 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 (8.19), 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

$$l_j = \arg\max_{l_k \in L} P(l_k|\mathbf{x}_R)\,P(l_k|\mathbf{x}_C) \tag{8.20}$$

8.5.2 Fusion Strategy

Decision rules (8.19) and (8.20) 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 (8.20).

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 Fig. 8.5.

To realize the decision process shown in Fig. 8.5, 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.

8.5.2.1 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 SVM [41] as the classifier, and the SVR method [41] as the regressor, which were also chosen in [21, 23]. The difference is that there is no probabilistic computation for the SVM and SVR in the LARR method [21, 23], while here the results of the SVM and SVR are transformed into probabilities and then fuse them automatically [22] without trying different local ranges and requiring users' selection as in [23].

8.5.2.2 Probabilistic Output for SVMs

Standard SVM 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 [38] 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 [29], 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 [17], 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

$$P(l_k|\mathbf{x}_C) = \frac{\sum_{j=1}^{n} \phi_{k,j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{i,j}} \tag{8.21}$$

8.5.2.3 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 [10]. 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 [23], 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 (8.19) or (8.20). 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 PDF of the uniform distribution $\mathrm{U}(\mu - \Delta, \mu + \Delta)$ is given by

$$p(x) = \begin{cases} \frac{1}{2\Delta} & \text{for } \mu - \Delta \leq x \leq \mu + \Delta, \\ 0 & \text{otherwise,} \end{cases} \tag{8.22}$$

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

$$P(l_k|\mathbf{x}_R) \leftarrow U\left(l_0 - \sqrt{3}\sigma,\ \ l_0 + \sqrt{3}\sigma\right). \tag{8.23}$$

The uniform distribution (8.23) 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.

8.5.2.4 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 (8.20), the two probabilities are multiplied and the label l_j corresponding to the maximum product is selected as the final decision.

Our serial PFA 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, $\left[l_0 - \sqrt{3}\sigma, l_0 + \sqrt{3}\sigma\right]$, by using the cutoff 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 Fig. 8.5, reducing the search space (i.e., less number of classes to compare) for the classifier SVM. The LARR method [21, 23] 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.

8.6 Soft Biometrics Computation

We have presented the methods of SVM, SVR, and the combinations of them. These methods will be used for age estimation on different databases. For gender and ethnicity classifications, only the SVMs are used, since these problems are typically considered as classifications.

We only use face images for soft biometrics computation. The face images are usually detected, aligned, cropped, and resized into the same size. Various features can be extracted from the face images to characterize the facial appearance. The specific methods for feature extraction will be briefly introduced in the experiments.

8.7 Experiments

We conduct experiments for age, gender, and ethnicity estimation, separately. Different databases might be used for each of the soft biometric measure. Not all databases are proper to study all of the three soft biometric characteristics.

8.7.1 Age Estimation Results

Age estimation experiments are conducted on the FG-NET and Yamaha Aging Databases. The FG-NET Aging Database [7] is a publicly available age database that we adopt for the experiment. The database contains 1,002 color or grayscale face images with variations of lighting, pose, and expression. There are 82 subjects (multiple races) in total with the age ranges from 0 to 69 years, and each face image has 68 labeled points characterizing shape features. The shape features can be combined with appearance features to form a face representation, called active appearance models (AAMs) [5]. The AAMs use 200 parameters to model each face for the purpose of age estimation [11, 46, 47].

The Yamaha Aging database contains 8,000 high-resolution RGB color face images captured from 1,600 different voluntary Asian subjects in an outdoor environment, 800 females and 800 males, in the age range from 0 to 93 years. Each subject has five near frontal images with provided ground truth ages. It has been used in some previous studies, e.g., [8, 9, 46, 47]. The Yamaha database is much larger than the FG-NET.

To evaluate the age estimation performance on Yamaha, a face detector was used to find the face area in each image, and the eye corner locations are labeled for each face subject. Based on the face and eye corner locations, the face images are cropped, scaled, and transformed to 60×60 gray-level patches [21, 23]. The images have significant variances in illumination since the photographs were taken in the outdoor environment. The gray-level values of each face image are normalized to a normal distribution with zero-mean and one standard deviation in order to reduce the effect of out-door illumination changes. The database also contains some facial expression variations and makeup.

The face image patches with the same size of 60×60 are fed into the manifold learning module. The age manifold can be embedded in a low dimensional subspace using different techniques [21]. Some manifold visualizations can be found in [21]. It has been shown in [21] that: (1) The principal component analysis (PCA) method does not show clear manifold trend of ages. The reason is that the PCA is purely unsupervised without using any age label information, which seems to be important for learning the embedded manifold from the complex aging patterns; (2) The manifold learned by the local linear embedding (LLE, a nonlinear embedding method) is approximately an ellipsoid with higher ages in the center and lower ages at periphery; and (3) The OLPP algorithm [4] achieves good visualization of the age manifold with a distinct aging trend. Therefore, we used the OLPP method in our age manifold learning module for age estimation [21].

After the age manifold was learned, each face image can be projected onto the age manifold to extract a feature vector. We used the first 150 features for each face image [21, 23]. The system then learns a robust regression function using the kernel SVR method for females and males separately. Actually the manifold was learned for the female and male independently. As demonstrated in the toy example in Sect. 8.2.2, a small ε value should be chosen for the ε-insensitive loss function in Eq. (8.3). We set $\varepsilon = 0.02$ for our age estimation task. In SVR learning, parameters C and γ are determined on a validation set. Experimentally we found that a good choice is $C = 40$ and $\gamma = 12$, separately. To locally adjust the global regression results, we tried different local search ranges as powers of two, e.g., 4, 8, 16, 32, and 64 classes, and the results from different search ranges are compared to see the effect of local adjustment. The purpose of choosing the powers of two is to simplify the binary search structure. One can observe that the local search range does influence the age estimation results. The pair-wise linear SVM classifiers were used for the local adjustment, centered at the age value (or label) obtained from the global regressor.

We perform a standard fourfold cross validation test to evaluate the accuracy of our algorithms for age estimation on the Yamaha age database. The test was executed on the female and male subsets separately. The females and males age quite differently in the database. For each experiment, about 1/3 of the training data are used as a validation subset to determine the optimal parameter setting such as C and γ. Then the parameters are fixed and the whole training data set is used to learn the robust regression function. The pair-wise linear SVM classifiers are learned using the same training data and used for local adjustment in testing. Finally all performance measures are reported on the unseen test data.

The performance of age estimation can be measured by two different measures: the mean absolute error (MAE) and the cumulative score (CS). The MAE is defined as the average of the absolute errors between the estimated ages and the ground truth ages, $\mathrm{MAE} = \sum_{k=1}^{N} |\widehat{l_k} - l_k|/N$, where l_k is the ground truth age for the test image k, $\widehat{l_k}$ is the estimated age, and N is the total number of test images. The cumulative score [11] is defined as $\mathrm{CS}(j) = N_{e\leq j}/N \times 100\,\%$, where $N_{e\leq j}$ is the number of test images on which the age estimation makes an absolute error no higher than j years.

Table 8.1 shows the experimental results. The first and second columns in Table 8.1 show the MAEs for females and males in the Yamaha aging database, separately. Different ranges, e.g., 4, 8, 16, 32, and 64, were tried for local adjustment of the global regression results. One can see that the local adjustment truly reduces the errors of the global regression. For example, the MAE of the SVR is 7 years for the female (column 1 in Table 8.1), but is reduced to 5.86 (column 1, row 5) when 16 local age classes are used for the LARR method, and so on. Different ranges of adjustment do have different MAEs. For comparison, we also show the results using purely the SVM classifiers in the first row. One can see that the classification scheme has lower errors than the pure regression method for both females and males, but it

Table 8.1 MAEs of the methods: SVM, SVR, and LARR with different settings [21, 23]

Various setup	Yamaha (Female)	Yamaha (Male)	FG-NET
SVM	5.55	5.52	7.16
SVR	7.00	7.47	5.16
LARR4	6.83	7.21	**5.07**
LARR8	6.48	6.81	**5.07**
LARR16	5.86	5.95	5.12
LARR32	5.29	**5.30**	6.03
LARR64	**5.25**	5.38	–

The bold fonts indicate the lowest errors in each case.

has higher error rates than some of the locally adjusted results. The best LARR result in terms of MAE is 5.25 years for females when the local search range is 64 classes, while it is 5.30 years for males when the adjust range is 32 classes. The ranges of local adjustment depend on the data and the global regression results. To illustrate the MAEs at each age, two pictures for female and male results are displayed in Fig. 8.6, respectively.

Figure 8.7a, b show the CS measures for females and males separately. We can observe that the LARR methods (with different ranges for local adjustment) improve the score significantly over the pure regression method for lower error levels, e.g., $m_c < 10$ years. For example, in one year error level, most LARRs with proper ranges of local adjustment could improve the accuracy by 175 % and 267 % for females and males separately. This improvement is significant. We also notice that large ranges are required for local adjustment on the Yamaha aging database. For instance, when 16 age classes are used for local adjustment, the CS curve is explicitly lower than 32 or 64 classes. We do not show the cumulative scores for four and eight classes here in order to not mess up the figures. Those two CS curves are even lower than 16 classes. One may also notice that the CS curve of SVM classifiers is close to the LARR32 and LARR64 for both females and males, but the MAEs of the SVM are higher than the LARR16 or LARR32 as shown in Table 8.1. This indicates that we need both MAE and CS measures complementarily to measure the performance of an algorithm in age estimation.

As shown in Table 8.2, we also compare our results with all previous methods reported on the Yamaha aging database. It turns out our LARR method has the MAEs of 5.25 and 5.30 years for females and males separately, which are explicitly smaller than the previous results under the same experimental protocol. Our method brings about 24 % deduction of MAEs over the best result of previous approaches, given in [46].

For age estimation on the FG-NET database, we used the same AAM features as in [11, 46, 47] to evaluate our LARR method [21, 23]. Since the FG-NET database has small size, we do not learn any age manifold but use the AAM features directly. Our focus is then to evaluate the performance of the LARR method for age estimation on the FG-NET database. The popular test strategy, namely leave-one-

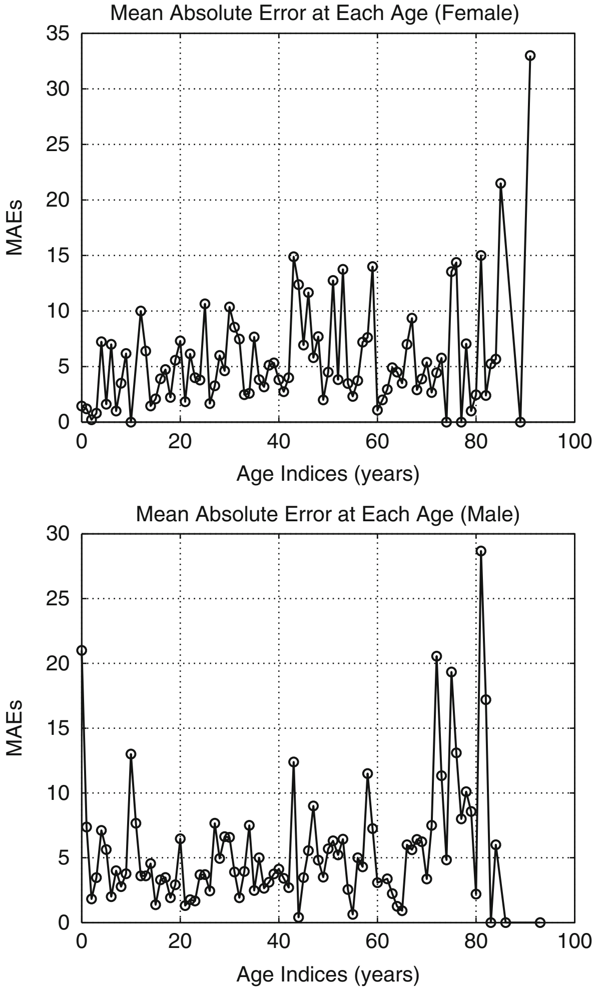

Fig. 8.6 MAEs at each age for females and males on the Yamaha Aging database, obtained by the LARR method [21]

person-out (LOPO), was usually taken for the FG-NET age database, as suggested by the existing work [11, 46, 47]. We follow the same strategy and compare our results with the state-of-the-art methods. The experimental results are shown in the third column of Tables 8.1 and 8.2. One can see that the LARR method has an MAE of 5.07 years which is lower than the previous methods listed in Table 8.2 [21]. The best MAE was obtained using either four or eight classes for local adjustment as shown in Table 8.1. Increasing the local search ranges for the LARR method will make the errors larger. For example, the MAE will be 6.03 years when 32 classes are used for local adjustment. We cannot get the result for 64 classes since there are

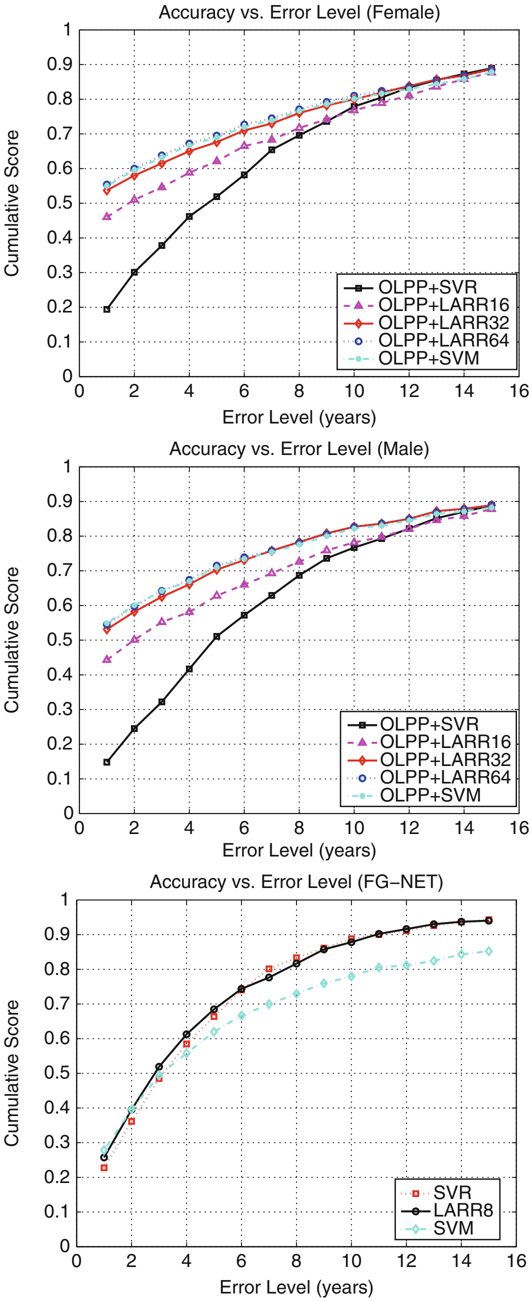

Fig. 8.7 Cumulative scores of the algorithms with different settings for (**a**) *Top*: female age estimation, (**b**) *Middle*: male age estimation on the Yamaha Aging database, and (**c**) *Bottom*: age estimation on the FG-NET database, at error levels from 1 to 15 years [21]

Table 8.2 MAE comparisons of different algorithms [21, 22]

Method	Yamaha (Female)	Yamaha (Male)	FG-NET
WAS [11]	–	–	8.06
AGES [11]	–	–	6.77
QM [32]	9.96	10.51	6.55
MLPs [32]	10.99	12.00	6.98
RUN1 [47]	9.79	10.36	5.78
RUN2 [46]	6.95	6.95	5.33
LARR [21]	5.25	5.30	5.07
PFA [22]	**5.11**	**5.12**	**4.97**

at most 63 or 61 age labels in the LOPO test. In other words, there are missing ages in the FG-NET database. When the pure classifiers, SVMs, are used, the MAE is 7.16, which is much higher than the 5.16 years of the pure regression. One possible reason is that there is not sufficient data for pair-wise SVM training, while the global SVR uses all the data in the model. Another observation is that the robust regression itself (without local adjustment) has an MAE of 5.16 years, which is still lower than all previous methods shown in Table 8.2. The LARR method further reduces the MAE to 5.07 years [21].

Figure 8.7c shows the cumulative scores of the LARR method on the FG-NET database. LARR8 means using eight classes for local adjustment. We do not show LARR4, LARR16, and LARR32 in order to avoid messing up the display. The cumulative scores of those ranges are close to LARR8 with slight differences. LARR8 has higher accuracy than the pure regression by SVR at lower error levels (1–6), but close to it at higher error levels. The cumulative scores of the pure SVM are much lower than the pure SVR for most error levels, which indirectly indicates the significance of constraining the SVM search in a local range. The LARR method performs much better than the QM and MLP methods. The method of RUN1 [47] is close to our LARR in low age error levels, but worse than LARR in high levels. In contrast, the method of RUN2 [46] is close to our LARR in high age error levels, but worse than the LARR in low error levels. Overall, the LARR method has higher accuracy than both the RUN1 and RUN2 on the FG-NET database.

Following the idea of combining the SVR with SVMs in the LARR method [21, 23], we proposed a PFA to combine the classifiers with regression in a probabilistic manner [22]. The PFA method can avoid the search range selection in LARR. To validate the PFA method, we performed age estimation experiments [22] on the Yamaha and FG-NET databases using the same protocol and data. The experimental results are shown in the last row in Table 8.2. The first and second columns in Table 8.2 show the MAEs for females and males in the Yamaha aging database, respectively. The last column shows the MAEs on the FG-NET aging database. From the table, we can see that the PFA method can improve the age estimation accuracies over the LARR method, in addition to the determination of local adjustment automatically.

Table 8.3 Numbers of male and female faces in the three age groups of the Yamaha aging database: young, adult, and senior

	Young (0–19)	Adult (20–60)	Senior (61–93)	All ages (0–93)
Male	1,000	2,050	950	4,000
Female	1,000	2,050	950	4,000
Both	2,000	4,100	1,900	8,000

In summary, we have shown that the SVM and the regression formulations can do well for age estimation. Age estimation can be considered as either a regression or a classification problem. Different results might be obtained in different databases, when classification or regression is applied. We proposed two methods to combine the SVR and SVM in order to improve the performance in age estimation. Both the LARR and PFA methods can take advantage of the SVM classifiers and the SVR for an improved performance.

8.7.2 Gender Classification

Gender classification is an interesting topic in both psychology [3, 44, 45] and computer vision [2, 28, 34, 48]. In computational approaches, various methods have been proposed for gender classification based on different facial image representations and classifier learning. Some typical approaches were listed in [24]. Among the different methods, an earlier work [34] applied the SVM to raw face images for gender classification.

We have studied the influence of age on gender classification in [24], based on several face image representations. The Yamaha database was used for the study. The number of males and females in each age group is shown in Table 8.3. One can see that it is very balanced for males and females in the database.

In addition to the raw face images, the LBP, HOG, and BIF features were used for gender recognition experimentally [24]. The goal was to evaluate the influence of age on gender recognition using several facial representations.

Both LBP and HOG features were extracted from each face image at various patch positions for the three age groups. Face images are of size 60×60, and the patch size is 16×16 with an interval of eight pixels between neighboring patches. The HOG operator has eight directions as in [6], and the LBP operator uses the uniform pattern as in [1]. Since HOG features were initially used with a linear SVM for pedestrian detection [6], we also show gender classification results based on linear SVMs (labeled as L-SVM) in addition to nonlinear SVMs with the RBF kernel (denoted as N-SVM). The LBP operator can be applied to the whole face image (denoted as LBP(W)) or applied to small patches on the face (denoted as LBP(P)). The patch-based LBP is much better than the whole face-based LBP in

Table 8.4 Gender recognition with different representations: raw pixels, LBP, HOG, and BIF, using the linear SVM (L-SVM) or nonlinear SVM (N-SVM) with the RBF kernel as classifiers [24]

Methods	Young (0–19)	Adult (20–60)	Senior (61–93)
Raw + L-SVM	78.59 %	89.91 %	81.17 %
Raw + N-SVM	84.38 %	94.56 %	85.32 %
LBP(W) + L-SVM	68.17 %	72.33 %	63.40 %
LBP(W) + N-SVM	69.65 %	77.08 %	68.40 %
LBP(P) + L-SVM	79.76 %	92.65 %	87.55 %
LBP(P) + N-SVM	81.93 %	94.96 %	90.64 %
HOG + L-SVM	75.83 %	88.00 %	77.13 %
HOG + N-SVM	86.44 %	94.03 %	89.04 %
BIF + L-SVM	83.01 %	94.22 %	91.81 %
BIF + N-SVM	87.13 %	96.03 %	92.34 %
Average(L-SVM)	80.52 %	91.20 %	84.42 %
Average(N-SVM)	84.97 %	94.90 %	89.34 %

gender recognition, as shown in Table 8.4. In each case, the parameters of the SVM are adjusted to optimal values on a tuning set (part of the training data).

8.7.2.1 HOG Feature

When the HOG features are used with nonlinear SVMs, gender recognition accuracies were 86.44 %, 94.03 %, and 89.04 %, for the young, adult, and senior groups, respectively, as shown in row 8 of Table 8.4. When compared with the "Raw+SVM" approach, the accuracies improved from 84.38 % to 86.44 % for the young faces, and improved from 85.32 % to 89.04 % for seniors, while the accuracy of 94.03 % for adults is slightly lower than the 94.56 % based on raw pixel representation. These results demonstrate that the HOG operator can characterize shape and improve recognition accuracies for young and senior faces. However, the two improved accuracies are still much lower than the 94.03 % accuracy for adult faces. On the other hand, the results indicate that the "Raw+SVM" approach is still good for gender recognition on adult faces.

We further explain the result [24] as: (1) adult male and female faces have local shape differences that can be described by the HOG operator and (2) shape changes in young faces and wrinkles in senior faces result in gradient variations that can be encoded by the HOG operator to some extent. However, the HOG performs much better for gender recognition on adult faces than on young and senior faces.

Also notice that linear SVMs performed much worse than kernel SVMs for each age group, as shown in rows 7 and 8 of Table 8.4.

8.7.2.2 LBP Feature

When LBP features were used with kernel SVMs, gender recognition accuracies were 81.93 %, 94.96 %, and 90.64 % for the young, adult, and senior groups, respectively, as shown in row 6 of Table 8.4. Here "P" represents patch-based LBP. When compared with the "Raw+SVM" approach, LBP features improved gender recognition accuracy for seniors (from 85.32 % to 90.64 %), but this is still lower than the accuracy of 94.96 % for adult faces. More interestingly, the accuracy reduced to 81.93 % for young faces, which is even lower than the 84.38 % accuracy of the "Raw+SVM" approach, and much lower than the 94.96 % accuracy for adult faces. Again, gender recognition performance is very different for the three age groups using LBP features: high performance for adult faces, lower performance for senior faces, and very low performance for young faces. Possible reasons for this phenomenon are: (1) adult male and female faces have local texture differences that can be described well by the LBP operator and (2) complex textures (e.g., wrinkles) on senior faces can also be described well by the LBP operator. For young faces, facial textures are not very rich and the main changes are facial shapes where the LBP operator does not work well [24].

It should be mentioned that linear SVMs with LBP features did not perform well for gender as shown in row 5 of Table 8.4. In addition, the LBP operator performed much worse when applied to whole faces, as shown in rows 3 and 4 of Table 8.4, no matter what classifier was used.

8.7.2.3 BIF Feature

For the biologically inspired features [26], we need to find the best structure and setting. To simplify the process, the gender recognition is performed over all ages first. A twofold cross validation was used as the test scheme. The same divisions of training and test data are used for all algorithms here, either over all ages or at separate age groups.

First, we evaluated the C2 features with a nonlinear SVM for gender classification over all ages. The feature extraction process is almost the same as that in [40]. The only difference is the number of prototypes to represent the gender. Since we have 8,000 images for the two-class classification problem, a small number of prototypes cannot work well (not shown here). We let the algorithm randomly select 2,000 prototypes from the female faces for S2 and C2 feature calculations. An accuracy of 81.05 % was obtained. This result is much worse than the 89.28 % using the raw pixel representation, the 88.65 % accuracy of the HOG method, and the 90.53 % of the LBP, shown in Table 8.4. We also randomly selected 2,000 prototypes from the male faces, and the result was 81.00 %—almost the same. Finally, we also let the algorithm randomly select 4,000 prototypes from both males and females, and got an accuracy of 83.00 %—still very low. From this experiment, we believe that C2 features do not work well for gender recognition. We notice that Meyers and Wolf [33] did not use C2 features in their face recognition problem, but they did not show any results when C2 features were used for face recognition. Based on our experience, C2 features are not a good choice for face-based gender classification, although these features demonstrated super performance on object category recognition [35, 40].

When the proper structure is determined for the BIF features, a better result can be obtained. More details about the BIF can be found in [24]. The results of BIF with both linear and nonlinear SVMs are given in Table 8.4. One can see that the kernel SVM performs better than the linear SVM in each age group. The BIF features combined with the kernel SVM can perform better than all other approaches in our comparisons.

8.7.2.4 Summary

We have shown the performance of the SVM for gender classification on a large database. The nonlinear SVM with the RBF kernel can perform significantly better than the linear SVM in all cases. Different methods have been used for facial image representation in the context of gender classification. The BIF features are better than the LBP and HOG features. More interestingly, we have shown that the gender classification is affected by ages. The adult faces can provide a much higher accuracy for gender classification than on young or senior faces. This was discovered quantitatively for the first time [24].

8.7.3 Ethnicity Estimation

We study ethnicity classification under variations of gender and age [15], using the SVM [41] as the classifier. We investigate whether the ethnicity estimation performance is affected by other human attributes, such as gender and age. Towards this goal, we designed experiments under two situations: (1) using female faces to learn an ethnicity classifier and then apply to males, and vice versa, and (2) learning ethnicity classifiers using faces from three age groups, and testing with different age groups.

The data were selected from the MORPH database [39] for this study [15]. The distribution of the selected data is shown in Table 8.5. The BIF features [26] were used for facial image characterization combined with manifold learning techniques [15].

8.7.3.1 Ethnicity w.r.t. Gender

To study whether the performance of ethnicity classification is affected by gender, we learn the ethnic classifiers using female and male faces, separately. Then we test the performance on the same and different gender to observe the difference.

Table 8.5 The distribution of the data selected from MORPH for the study

	Female	Male	Female and male
White	2,570	7,960	10,530
Black	2,570	7,960	10,530
White and black	5,140	15,920	21,060

Table 8.6 A study of ethnicity estimation with respect to gender [15]

		Ethnicity classification concerning gender						
		BIF		BIF+PCA		BIF+OLPP		
Train.	Test	Accuracy	Accuracy decrease	Accuracy	Accuracy decrease	Accuracy	Accuracy decrease	Comments
F1	F2	98.7 %	–	98.9 %	–	99.1 %	–	Same gender
	M1	94.0 %	4.8 %	93.7 %	5.3 %	90.3 %	8.9 %	Female → Male
	M2	94.1 %	4.7 %	93.8 %	5.2 %	90.4 %	8.8 %	Female → Male
F2	F1	98.6 %	–	98.9 %	–	99.3 %	–	Same gender
	M1	91.4 %	7.3 %	90.9 %	8.1 %	92.3 %	7.1 %	Female → Male
	M2	91.4 %	7.3 %	91.1 %	7.9 %	92.2 %	7.2 %	Female → Male
M1	M2	98.8 %	–	98.8 %	–	99.1 %	–	Same gender
	F1	96.8 %	2.0 %	97.2 %	1.6 %	97.7 %	1.4 %	Male → Female
	F2	96.5 %	2.3 %	97.1 %	1.7 %	97.3 %	1.8 %	Male → Female
M2	M1	98.7 %	–	98.7 %	–	98.8 %	–	Same gender
	F1	97.5 %	1.2 %	97.6 %	1.1 %	98.3 %	0.5 %	Male → Female
	F2	97.1 %	1.6 %	97.3 %	1.4 %	97.6 %	1.2 %	Male → Female
F1	F2	98.7 %	–	98.9 %	–	99.1 %	–	Same gender
	M1$_S$	93.9 %	4.9 %	93.1 %	5.9 %	89.6 %	9.6 %	Female → Male$_S$
	M2$_S$	94.1 %	4.7 %	94.1 %	4.9 %	90.5 %	8.7 %	Female → Male$_S$
F2	F1	98.6 %	–	98.9 %	–	99.3 %	–	Same gender
	M1$_S$	90.6 %	8.1 %	90.4 %	8.6 %	91.9 %	7.5 %	Female → Male$_S$
	M2$_S$	91.7 %	7.0 %	91.7 %	7.3 %	92.5 %	6.8 %	Female → Male$_S$
M1$_S$	M2$_S$	98.4 %	–	98.6 %	–	99.0 %	–	Same gender$_S$
	F1	96.2 %	2.2 %	96.8 %	1.8 %	96.9 %	2.1 %	Male$_S$ → Female
	F2	95.8 %	2.6 %	96.3 %	2.3 %	96.6 %	2.4 %	Male$_S$ → Female
M2$_S$	M1$_S$	98.3 %	–	98.7 %	–	98.8 %	–	Same gender$_S$
	F1	96.4 %	1.9 %	97.3 %	1.4 %	97.6 %	1.2 %	Male$_S$ → Female
	F2	96.3 %	2.0 %	97.0 %	1.7 %	97.2 %	1.6 %	Male$_S$ → Female

There are two ethnic groups, white (W) and black (B). So we have four groups with gender: white female (WF), black female (BF), white male (WM), and black male (BM). Each of the four groups is randomly divided into two subgroups for cross validations.

Through comparisons, we can infer the effect of gender difference on ethnicity estimation [15].

For the selected data shown in Table 8.5, we have four groups, black female (BF), white female (WF), black male (BM), and white male (WM). Within each group, the data are randomly divided into two subgroups, labeled as 1 and 2, in order to do cross validations. Suppose we choose one subgroup from the BF and another subgroup from WF to learn the ethnic classifier, labeled as F1 = BF1 + WF1, without any loss of generality. Then we can test the performance on female or male faces. Remember that we have another female data set, denoted as F2 = BF2 + WF2, and two subsets for male faces, M1 = BM1 + WM1, and M2 = BM2 + WM2. For ethnicity estimation with the same gender, the subset F2 is tested, denoted as F1 $\rightarrow$ F2. For different gender evaluation, we use M1 and M2 for testing, denoted as F1 $\rightarrow$ M1 and F1 $\rightarrow$ M2. Similarly, we can use F2, M1, or M2 for training, and use the remaining data for testing.

The experimental results are shown in Table 8.6. We have 16 ethnicity classification experiments for each face representation. So there are 48 experiments in total, using the three face representations. The original dimensionality of the BIF is 4,376. It is reduced to about 500 using PCA, and reduced to about 100 using OLPP. These numbers are kept the same throughout the experiments.

The 48 experiments can be categorized into three kinds of ethnicity classifications: same gender, female $\rightarrow$ male, and male $\rightarrow$ female. From Table 8.6, we can observe that (1) for ethnicity estimation using the same gender, the classification accuracies are very high (from 98.6% to 99.3%) for all three face representations. This demonstrates that our face representations have very good performance for ethnicity estimation; (2) for ethnicity estimation of male $\rightarrow$ female (using male faces to learn and females to test), the classification accuracies are slightly lower than using the same gender, ranging from 96.5 % to 98.3 %, but the accuracy decreases (accuracy difference between the cases of cross-gender and the same gender using the same training data, divided by the accuracy in the same gender case) are relatively small, e.g., from 0.5 % to 2.3 %; and (3) for ethnicity estimation of female $\rightarrow$ male, the classification accuracies range from 90.3 % to 94.1 %, with quite large accuracy decreases, e.g., from 4.7 % to 8.9 %, corresponding to different face representations.

One might notice that the number of female faces is smaller than males in Table 8.5. Do the accuracy differences come from the different sample sizes? To check this issue, we reduce the number of males in Table 8.5 to make the number of males equal to females. Specifically, we randomly chose partial males from M1 (i.e., 1,285 faces) and from M2 (1,285 faces), denoted as $M1_S$ and $M2_S$, respectively. Now, F1, F2, $M1_S$, and $M2_S$ have the same number of faces. Then we use the reduced data set to re-learn the ethnicity classifiers, and re-perform the 48 ethnic classification experiments, with the results shown in the lower part of Table 8.6. One can see that almost the same accuracy decreases can be observed from the equal-sized-data experiments.

As a result, our study demonstrates that ethnicity estimation is influenced by gender significantly when the female faces are used for training while males for testing, i.e., female $\rightarrow$ male. However, the reversed process (male $\rightarrow$ female) has some influence but not very significant. This *unsymmetric* influence is interesting. We are not very clear about how to interpret this phenomenon yet; however, we hope the computational results inspire more psychological studies [37,49,50] to get a reasonable interpretation.

8.7.3.2 Ethnicity w.r.t. Age

To study whether ethnicity estimation is affected by age [15], we divided the data set into three age groups, labeled as A, B, and C. The partition considers the number of face images in different age groups to make them comparable, since the original data in MORPH do not have balanced number of faces at each age. Based on this and the age range (from 16 to 67 years), we determined that age group A contains ages less than or equal to 25 years, group B has ages greater than 25 but less than or equal to 40, and group C contains ages above 40. Remember that we still need two subgroups (1 and 2) within each age group for the purpose of cross validations, and each subgroup has both black and white faces to learn the ethnic classifier. The final distribution of the age groups is that A1 (2,756 faces, $16 \leq age \leq 25$), B1 (4,508 faces, $25 < age \leq 40$), C1 (3,266 faces, $40 < age \leq 67$), A2 (2,756 faces, $16 \leq age \leq 25$), B2 (4,508 faces, $25 < age \leq 40$), and C2 (3,266 faces, $40 < age \leq 67$). Not strictly, we name groups A, B, and C as young, middle, and old to make it easier to interpret the results.

Then we use one age group to train the ethnicity classifier, and the remaining age groups for testing. There are 30 ethnicity estimation experiments for each face representation, and there are 120 experiments in total given the three face representations. The experimental results are given in Table 8.7.

From the table, we can observe that (1) for ethnicity estimation within the same age group, i.e., A1 $\leftrightarrow$ A2, B1 $\leftrightarrow$ B2, C1 $\leftrightarrow$ C2, the ethnic classification accuracies can be very high, ranging from 98.3% to 99.1%, using the three face representations. (2) for ethnicity estimation with different age groups for training and testing, most of the results still have high accuracies, e.g., from 97.6% to 98.7%, for young $\leftrightarrow$ middle, middle $\leftrightarrow$ old, and old $\rightarrow$ young. In comparison with the same age group results, the accuracy decreases are relatively small, ranging from 0.0% to 1.3%, using three face representations. In the case of young $\rightarrow$ old, the accuracy decreases are slightly larger, e.g., from 2.0% to 2.7%, but not so significant as the gender influence on ethnicity estimation in the case of female $\rightarrow$ male.

8.7.3.3 Summary

We can reorganize the above experimental results by averaging over the subcases, so that one can observe the performance more directly. The new results are shown in Table 8.8. From the results, we can easily observe that (1) ethnicity estimation can have very high accuracies if it is performed within the same gender and age groups; (2) our face representations based on biologically inspired features with or without manifold learning show high performance in ethnicity classification; (3) ethnicity estimation can be affected in the cross-gender case of female $\rightarrow$ male, with accuracy decreases of 6$\sim$8 % in average, which is significantly different from the situations of the same gender and male $\rightarrow$ female; (4) ethnicity estimation is not affected very much under the situation of cross-age.

Table 8.7 A study of ethnicity estimation with respect to age [15]

		Ethnicity classification concerning age						
		BIF		BIF+PCA		BIF+OLPP		
Train.	Test	Accuracy	Accuracy decrease	Accuracy	Accuracy decrease	Accuracy	Accuracy decrease	Comments
A1	A2	98.4 %	–	98.5 %	–	99.0 %	–	Same age group
	B1	97.7 %	0.7 %	97.8 %	0.7 %	98.2 %	0.8 %	Young → Middle
	B2	97.7 %	0.7 %	97.8 %	0.7 %	98.4 %	0.6 %	Young → Middle
	C1	95.9 %	2.5 %	96.0 %	2.5 %	96.8 %	2.2 %	Young → Old
	C2	95.7 %	2.7 %	95.8 %	2.7 %	96.3 %	2.7 %	Young → Old
A2	A1	98.5 %	–	98.5 %	–	98.8 %	–	Same age group
	B1	97.7 %	0.8 %	97.9 %	0.6 %	98.3 %	0.5 %	Young → Middle
	B2	98.3 %	0.2 %	98.2 %	0.3 %	98.5 %	0.3 %	Young → Middle
	C1	96.3 %	2.2 %	96.3 %	2.2 %	96.8 %	2.0 %	Young → Old
	C2	95.8 %	2.7 %	95.8 %	2.7 %	96.2 %	2.6 %	Young → Old
B1	B2	98.9 %	–	98.7 %	–	99.1 %	–	Same age group
	A1	98.1 %	0.8 %	98.3 %	0.4 %	98.6 %	0.5 %	Middle → Young
	A2	98.5 %	0.4 %	98.2 %	0.5 %	98.7 %	0.4 %	Middle → Young
	C1	97.7 %	1.2 %	97.9 %	0.8 %	98.1 %	1.0 %	Middle → Old
	C2	97.6 %	1.3 %	97.8 %	0.9 %	98.2 %	0.9 %	Middle → Old
B2	B1	98.7 %	–	98.8 %	–	99.0 %	–	Same age group
	A1	98.3 %	0.4 %	98.3 %	0.5 %	98.4 %	0.6 %	Middle → Young
	A2	98.5 %	0.2 %	98.4 %	0.4 %	98.7 %	0.3 %	Middle → Young
	C1	97.9 %	0.8 %	97.8 %	1.0 %	97.7 %	1.3 %	Middle → Old
	C2	98.0 %	0.7 %	97.8 %	1.0 %	98.2 %	0.8 %	Middle → Old
C1	C2	98.7 %	–	98.7 %	–	98.8 %	–	Same age group
	A1	98.1 %	0.6 %	98.1 %	0.6 %	98.1 %	0.7 %	Old → Young
	A2	98.1 %	0.6 %	98.3 %	0.4 %	98.2 %	0.6 %	Old → Young
	B1	98.4 %	0.3 %	98.3 %	0.4 %	98.6 %	0.2 %	Old → Middle
	B2	98.4 %	0.3 %	98.4 %	0.3 %	98.6 %	0.2 %	Old → Middle
C2	C1	98.3 %	–	98.3 %	–	98.7 %	–	Same age group
	A1	97.8 %	0.5 %	97.7 %	0.6 %	98.0 %	0.7 %	Old → Young
	A2	97.6 %	0.7 %	98.0 %	0.3 %	97.9 %	0.8 %	Old → Young
	B1	98.2 %	0.1 %	98.0 %	0.3 %	98.5 %	0.2 %	Old → Middle
	B2	98.3 %	0.0 %	98.3 %	0.0 %	98.6 %	0.1 %	Old → Middle

The data set is divided into three age groups: Young or A ($age \leq 25$ years), Middle or B ($age \leq 40$), and Old or C ($age > 40$). Each age group is randomly divided into two subgroups for cross validations.

8.7.3.4 Usefulness of the Study

Our study results have applications in many real problems. For example, for a large database containing multiple ethnic groups, one may categorize the ethnic groups before age estimation [25, 36], since the ethnicity estimation is not very sensitive to age variations from our studies. Categorizing into different ethnicity groups may reduce the age estimation errors [36] significantly. For the problem of gender

Table 8.8 A summary of our studies on ethnicity classification versus the changes of gender and age groups [15]

	Ethnicity classification					
	BIF		BIF+PCA		BIF+OLPP	
Versus Gender or age	Average accuracy	Accuracy decrease	Average accuracy	Accuracy decrease	Average accuracy	Accuracy decrease
Same gender	98.7 %	–	98.8 %	–	99.1 %	–
Female → Male	92.7 %	6.1 %	92.4 %	6.5 %	91.3 %	7.9 %
Male → Female	97.0 %	1.7 %	97.3 %	1.5 %	97.7 %	1.4 %
Same gender$_S$	98.5 %	–	98.8 %	–	99.1 %	–
Female → Male$_S$	92.6 %	6.0 %	92.3 %	6.7 %	91.1 %	8.1 %
Male$_S$ → Female	96.2 %	2.3 %	96.9 %	1.9 %	97.1 %	2.0 %
Same age group	98.6 %	–	98.6 %	–	98.9 %	–
Young → Middle	97.9 %	0.7 %	97.9 %	0.7 %	98.4 %	0.5 %
Young → Old	95.9 %	2.7 %	96.0 %	2.6 %	96.5 %	2.4 %
Middle → Young	98.4 %	0.2 %	98.3 %	0.3 %	98.6 %	0.3 %
Middle → Old	97.7 %	0.9 %	97.8 %	0.8 %	98.1 %	0.8 %
Old → Young	97.9 %	0.7 %	98.0 %	0.6 %	98.1 %	0.8 %
Old → Middle	98.3 %	0.3 %	98.3 %	0.3 %	98.6 %	0.3 %

classification [24,48] on a large database with multiple ethnic groups, one may also perform ethnic classification first, and then gender recognition is performed within each single ethnic group, since in most cases, the ethnicity estimation is not very sensitive to gender variations based on our studies [15]. We believe that multi-ethnic databases will be more and more popular in computer vision research, considering more databases are collected from the Internet, such as in [36]. We expect more research work will be reported on multi-ethnic face image databases in the near future.

On the other hand, our study based on computational analysis may inspire more psychological studies on ethnic grouping [37, 49, 50] related to age and gender variations. Interpretations about our results could be derived from further psychological studies.

8.8 Conclusions

We have presented the applications of the SVM to soft biometrics recognition in face images. The SVM can have very good performance for gender and ethnicity classification, when combined with appropriate features to characterize the facial appearance. For age estimation, we showed the performance of the SVM and SVR on two databases, since age estimation can be considered either a classification or a regression problem. We found that the two approaches can perform quite differently on different databases. A better way is to combine them to take advantage of both.

Two schemes, called LARR and PFA, have been proposed to integrate the SVM with SVR and validated for age estimation. The performance can be improved significantly when these schemes are used for age estimation. Further, we studied the influence of age on gender classification, and also the influence of age and gender on ethnicity estimation, based on the SVM classifiers. Overall, the SVM and their extensions are very useful for learning soft biometric characteristics from face images.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: European Conference on Computer Vision, pp. 469–481 (2004)
2. Baluja, S., Rowley, H.A.: Boosting sex identification performance. Int. J. Comput. Vision **71**(1), 111–119 (2007)
3. Bruce, V., Burton, A., Hanna, E., Healey, P., Mason, O.: Sex discrimination: how do we tell the difference between male and female faces? Perception **22**, 131–152 (1993)
4. Cai, D., He, X., Han, J., Zhang, H.: Orthogonal laplacianfaces for face recognition. IEEE Trans. Image Process. **15**, 3608–3614 (2006)
5. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. In: European Conference on Computer Vision, pp. 484–498 (1998)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on CVPR, pp. 886–893 (2005)
7. FGNET: The fg-net aging database. http://www.fgnet.rsunit.com/ (2002)
8. Fu, Y., Huang, T.S.: Human age estimation with regression on discriminative aging manifold. IEEE Trans. Multimedia **10**(4), 578–584 (2008)
9. Fu, Y., Xu, Y., Huang, T.S.: Estimating human ages by manifold analysis of face pictures and regression on aging features. In: IEEE Conference on Multimedia and Expo, pp. 1383–1386 (2007)
10. Gao, J.B., Gunn, S.R., Harris, C.J., Brown, M.: A probabilistic framework for svm regression and error bar estimation. Mach. Learn. **46**(1–3), 71–89 (2002)
11. Geng, X., Zhou, Z.H., Zhang, Y., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: ACM Conference on Multimedia, pp. 307–316 (2006)
12. Gunn, S.R.: Support vector machines for classification and regression. ISIS Technical Report 14 (1998)
13. Guo, G.D., Dyer, C.: Learning from examples in the small sample case: face expression recognition. IEEE Trans. Syst. Man Cybern. Part B **35**(3), 447–488 (2005)
14. Guo, G.D., Li, S.: Content-based audio classification and retrieval by support vector machines. IEEE Trans. Neural Netw. **14**(1), 209–215 (2003)
15. Guo, G.D., Mu, G.: A study of large-scale ethnicity estimation with gender and age variations. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures (2010)
16. Guo, G.D., Li, S., Chan, K.: Face recognition by support vector machines. In: Proceedings of Fourth IEEE International Conference Automatic Face and Gesture Recognition (2000)
17. Guo, G.D., Zhang, H., Li, S.: Pairwise face recognition. In: Proceedings of Eighth International Conference on Computer Vision, vol. 2, pp. 282–287 (2001)
18. Guo, G.D., Li, S., Chan, K.: Support vector machines for face recognition. Image Vis. Comput. **19**(9–10), 631–638 (2001)
19. Guo, G.D., Jain, A., Ma, W., Zhang, H.: Learning similarity measure for natural image retrieval with relevance feedback. IEEE Trans. Neural Netw. **13**(4), 811–820 (2002)

20. Guo, G.D., Dyer, C., Zhang, Z.: Linear combination representation for outlier detection in motion tracking. In: Proceedings of IEEE Conference Computer on Vision and Pattern Recognition, vol. 2, pp. 274–281 (2005)
21. Guo, G.D., Fu, Y., Dyer, C., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. Image Process. **17**(7), 1178–1188 (2008)
22. Guo, G.D., Fu, Y., Dyer, C., Huang, T.S.: A probabilistic fusion approach to human age prediction. In: International Workshop on Semantic Learning Applications in Multimedia (2008)
23. Guo, G.D., Fu, Y., Huang, T., Dyer, C.: Locally adjusted robust regression for human age estimation. In: IEEE Workshop on Application of Computer Vision (2008)
24. Guo, G.D., Dyer, C., Fu, Y., Huang, T.S.: Is gender recognition affected by age? In: IEEE International Workshop on Human-Computer Interaction, pp. 2032–2039 (2009)
25. Guo, G.D., Mu, G., Fu, Y., Dyer, C., Huang, T.S.: A study on automatic age estimation on a large database. In: IEEE International Conference on Computer Vision, pp. 1986–1991 (2009)
26. Guo, G.D., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 112–119 (2009)
27. Guo, G.D., Mu, G., Ricanek, K.: Cross-age face recognition on a very large database: the performance versus age intervals and improvement using soft biometric traits. In: International Conference on Pattern Recognition (2010)
28. Gutta, S., Wechsler, H.: Gender and ethnic classification of face images. In: International Conference on Automatic Face and Gesture Recognition, pp. 194–199 (1998)
29. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. Ann. Stat. **26**(2), 451–471 (1998)
30. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Springer, New York (2007)
31. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
32. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE Trans. SMC-B **24**(4), 621–628 (2002)
33. Meyers, E., Wolf, L.: Using biologically inspired features for face processing. Int. J. Comput. Vis. **76**, 93–104 (2008)
34. Moghaddam, B., Yang, M.H.: Learning gender with support faces. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 707–711 (2002)
35. Mutch, J., Lowe, D.: Object class recognition and localization using sparse features with limited receptive fields. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 11–18 (2006)
36. Ni, B., Song, Z., Yan, S.: Web image mining towards universal age estimator. In: ACM Multimedia (2009)
37. Okazaki, S., Sue, S.: Methodological issues in assessment research with ethnic minorities. Psychol. Assess. **7**(3), 367–375 (1995)
38. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classifiers **10**(3), 61–74 (1999)
39. Ricanek, K., Tesafaye, T.: Morph: a longitudinal image database of normal adult age-progression. In: IEEE Conference on AFGR, pp. 341–345 (2006)
40. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 411–426 (2007)
41. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
42. Webb, A.R.: Statistical Pattern Recognition, 2nd edn. Wiley, New York (2002)
43. Wen, L., Guo, G.: A computational approach to body mass index prediction from face images. Image Vis. Comput. **31**(5), 392–400 (2013)
44. Wild, H.A., Barrett, S.E., Spence, M.J., O'Toole, A.J., Cheng, Y.D., Brooke, J.: Recognition and sex categorization of adults' and children's faces: examining performance in the absence of sex-stereotyped cues. J. Exp. Child Psychol. **77**, 269–291 (2000)

45. Yamaguchi, M.K., Hirukawa, T., Kanazawa, S.: Judgment of sex through facial parts. Perception **24**, 563–575 (1995)
46. Yan, S., Wang, H., Huang, T.S., Tang, X.: Ranking with uncertain labels. In: IEEE Conference on Multimedia and Expo, pp. 96–99 (2007)
47. Yan, S., Wang, H., Tang, X., Huang, T.: Learning auto-structured regressor from uncertain nonnegative labels. In: IEEE Conference on ICCV (2007)
48. Yang, Z., Ai, H.: Demographic classification with local binary patterns. In: International Conference on Biometrics, pp. 464–473 (2007)
49. Yee, A.H.: Ethnicity and race: psychological perspectives. Educ. Psychol. **18**(1), 14–24 (1983)
50. Zuckerman, M.: Some dubious premises in research and theory on racial differences: scientific, social, and ethical issues. Am. Psychol. **45**(12), 1297–1303 (1990)