Ordinal Hyperplanes Ranker with Cost Sensitivities for Age Estimation

Kuang-Yu Chang [1,3], Chu-Song Chen [1,2,4], and Yi-Ping Hung [1,3,4]
[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan.
[2] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.
[3] Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.
[4] Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan.
{kuangyu, song}@iis.sinica.edu.tw, hung@csie.ntu.edu.tw

Abstract

In this paper, we propose an ordinal hyperplane ranking algorithm called OHRank, which estimates human ages via facial images. The design of the algorithm is based on the relative order information among the age labels in a database. Each ordinal hyperplane separates all the facial images into two groups according to the relative order, and a cost-sensitive property is exploited to find better hyperplanes based on the classification costs. Human ages are inferred by aggregating a set of preferences from the ordinal hyperplanes with their cost sensitivities. Our experimental results demonstrate that the proposed approach outperforms conventional multiclass-based and regression-based approaches as well as recently developed ranking-based age estimation approaches.

1. Introduction

In recent years, there has been growing interest in age estimation based on facial images. Age information is useful in a variety of applications, such as human-computer interaction, surveillance monitoring, and video content analysis. The objective of age estimation is to evaluate a person's exact age or age-group based on features derived from a facial image. In this paper, we focus on predicting a person's exact age, but the proposed method can also be applied to age-group estimation.

Existing approaches formulate the age estimation problem as a multi-class classification problem [14, 30, 25, 10] or a regression problem [9, 18, 12, 11, 31]. Given a set of N training examples $\{(x_i, y_i)|i = 1, \ldots, N\}$, where x_i is the i-th training facial image and y_i is the age label of x_i, multi-class classification approaches simply treat y_i as a set of discrete labels and learn a classifier to infer the person's age. Regression approaches basically learn a function that best fits the mapping from x_i to y_i with appropriate regularization. First, we discuss the two approaches based on the nature of the problem.

In a multi-class problem, the class labels are basically uncorrelated. Hence, multi-class classification approaches may ignore some characteristics for age estimation, because the labels are assumed to be independent or have no inherent relationship to each other. However, the age labels themselves are ordinal by nature, i.e., they have strong interrelationships, since they form a well-ordered set. For example, if a child is 10-years old, the age label is more likely to be related to a label for 9- or 11-year-old children than to a label for 8 or 12-years old. Typical multi-class approaches cannot reflect this property because the labels are treated unrelated.

Regression approaches, on the other hand, consider the labels as numerical values that utilize the ordering information for age estimation. It has been shown that they [11, 26, 31, 12] achieve a better performance than classification approaches. Many nonlinear regression approaches, such as quadratic regression [11], Gaussian Process [26, 31] and Support Vector Regression (SVR) [12, 11], have been used to solve the age estimation problem. However, the human face matures in different ways depending on the person's age, e.g., bone growth in childhood and skin wrinkles in adulthood [20]. This property makes the random process formed by human aging patterns non-stationary in the feature space; thus, the kernel functions used to measure the pair-wise similarities between ages could be shift- or time-varying. Nevertheless, learning non-stationary kernels is difficult for a regressor because it is apt to overfit the training data.

In recent years, there has been growing interest in learning to rank models in the machine learning community. The concept of ordinal ranking has attracted increasing attention for age estimation because human aging processes show diversity in different age ranges. For example, the difference in the aging process between 50- and 55-years old is not equivalent to that between 5- and 10-years old. Facial aging effects appear as changes in the shape of the face during childhood and changes in skin texture during adulthood. Hence, given two labels k_1 and k_2, the "larger than" information ($k_2 > k_1$ or $k_1 > k_2$) could be a more reliable property for age estimation than the differences between the labels.

In this paper, we employ the *relative order* of age labels because it provides more stable information than exact age values. Ranking is often used in information retrieval and formalized as a "learning to rank" problem that maps the given documents into ordered ranks. Some early schemes simply performed ranking based on regression [6] or classification [16]. Other popular schemes include pointwise ordinal regression and pairwise preferences. Herbrich et al. [13] proposed a ranking approach called Ranking SVM, which is based on hinge loss and SVM formulation. Ranking SVM (and several variations of the approach) takes the difference between two feature vectors as input for learning, and maps the higher ranked vector to higher scores during testing. Similarly, RankBoost [8] and RankNet [2] employ exponential loss and cross entropy loss respectively for learning pairwise ranking algorithms.

Ranking SVM only uses a single hyperplane to infer the order, which may not be enough to represent the different class boundaries of distinct labels. In [22], Shashua and Levin proposed utilizing a set of parallel hyperplanes as the ranking model, and constructed an SVM-based formulation to solve the common normal vector and the shifting constants (i.e., thresholds) of the parallel hyperplanes. Subsequently, Lin and Li [15] extended the approach in [22] and developed RED-SVM, which uses the costs that are sensitive to labels to further improve the performance. Unlike Ranking SVM, the ordinal regression approach formulated by parallel hyperplanes yields well ranked results directly and infers the threshold for each rank automatically.

However, single or parallel hyperplanes could still be too restrictive to reflect the diverse distributions of different classes. To resolve this problem, Qin et al. [19] proposed a multiple-hyperplanes approach called Multiple Hyperplanes Ranker, which employs several (possibly non-parallel) hyperlanes for pairwise comparison and then aggregates the results. Multiple Hyperplanes Ranker uses Ranking SVM as base ranker and aggregates rank information between two classes. The approach enables a single ranker to handle the relationship between instances from different ranks. If there are K classes, Multiple Hyperplanes Ranker constructs $K(K-1)/2$ rankers for the class pairs. Although the approach improves the performance of single hyperplanes because their form is less restrictive, it does not make complete use of the ordering information among the labels. Basically, the Multiple Hyperplanes Ranker only uses part of ordering information and reduces the ranking problem to a one-versus-one classification problem.

In this paper, we propose an approach that fully utilizes a well-ordered set of relationship labels. Given a well-ordered set of labels, such as human ages, the approach separates the labels into two groups based on the following ordering property: *larger than* and *no larger than* a label k. By exploiting this property, we can form $K-1$ subproblems, which are considerably less complex than the $K(K-1)/2$ setting. Although we could use the conventional one-versus-all strategy to form $K-1$ subproblems, it increases the number of imbalanced subproblems. Moreover, it does not exploit the order of the relationships among labels.

We also formulate a cost sensitive property to deal with each subproblem. Recently, the cost sensitive property concept has been discussed in the machine learning field as an effective way to reflect the severity of misclassification problems. The goal of cost-sensitive learning is to minimize the *total cost* rather than the *total error* as the cost of misclassification typically varies among different pairs of labels. We propose a cost-sensitive ordinal ranking framework for age estimation in this paper.

Although not a hyperplane-based ranker, an early approach [7] also integrates *K* classification probabilities according to the orders for ranking. However, their base classifier is built by decision tree that could not be discriminating enough, and the cost-sensitive property has not been considered. Experimental results demonstrate that our algorithm outperforms several state-of-the-art methods on two popular age estimation databases. The contribution of this paper is two-fold.

1) We formulate a series of proper subproblems for age estimation based on the ordering information which is a distinctive feature of age labels.

2) We utilize the cost sensitivity of labels for age estimation, which has not been well studied before.

The reminder of this paper is organized as follows. In the next section, we review related works. In Section 3, we describe the framework of OHRank, the proposed cost-sensitive ranking-based age estimation approach; and in Section 4 we present the experimental results. Section 5 contains some concluding remarks.

2. Related Work

In recent years, a number of age estimation approaches have been proposed. Lanitis et al. [14] were the first to use Active Appearance Models (AAMs) [5], which combine shape and intensity variation in facial images. Age estimation is regarded as a classification problem that can be solved by the shortest distance classifier and neural networks. The approach also differentiates between age-specific estimation and appearance-specific estimation. The former assumes that the aging process is the same for everyone; while the latter is based on the assumption that people who look similar tend to have similar aging processes. Personalized age estimation used in the specialty of aging processes is then introduced to cluster similar faces before classification.

Geng et al. [10] proposed a personalized age estimation method that describes the long-term aging process of a person and estimates his/her age by minimizing the reconstruction error. However, person's facial features could still be similar in different age ranges. Zhang et al. [31] developed a more accurate model that considers both common and person-independent information in a multi-task learning framework, and uses a warped Gaussian process to model a person's face. How the long-term aging process will affect a person may not be easy to estimate. Suo et al. [23] addressed the problem of the lack of long-term dense aging sequences by building long-term aging patterns from several short-term patterns because the latter are easier to obtain.

However, there may still be insufficient long-term or short-term aging patterns when there are not enough samples of a person's face. Hence, many studies have focused on non-personalized approaches. For example, Yang and Ai [30] used a real AdaBoost algorithm to train a strong classifier by composing a sequence of local binary pattern (LBP) histogram features. Then, they conducted experiments on gender, ethnicity and age classifications. Ni et al. [18] presented a multi-instance regression method that estimates the ages of faces in images with noisy labels collected from Web image resources. Guo et al. [12] investigated biologically inspired features comprised of a pyramid of Gabor filters in all positions in faccial images, and used either Support Vector Machine (SVM) or SVR with Radial Basis Function (RBF) kernels for evalution.

Some approaches use a manifold learning scheme for accurate modeling and age prediction. Fu and Huang [9] applied discriminative manifold learning and quadratic regression to age estimation, and claimed that their framework is more suitable than linear and cubic regressions. Guo et al. [11] introduced an age manifold scheme and combined SVR and SVM to learn and predict human ages. First, robust regression is employed to approximate the data, after which classification is used for local adjustment.

Existing approaches, both personalized and non-personalized, treat age estimation as either a classification or a regression problem. In this paper, we investigate a new direction by treating the age labels as ranking orders instead of exact values or independent tags. To the best of our knowledge, very few approaches employ the ranking principle for age estimation. Yang et al. [29] employ the RankBoost algorithm, which is a single hyperplane ranker in the feature space, for Harr feature selection and age estimation. The drawback of this approach is that it uses a single hyperplane model, which can not reflect the multiple thresholds of different classes properly. Chang et al. [4] employ the parallel hyperplanes model RED-SVM [15], but it is relatively restricted for age estimation. As showen in Figure 1, RED-SVM constructs $K-1$ parallel hyperplanes that simultaneously maximize the $K-1$ margins. The obtained parallel hyperplanes separate the hyperspace into K ranks by using $K-1$ thresholds. The underlying assumption is that the K classes are well ordered in a unique direction and separable by hyperspaces. This algorithm is uaually applied on databases containing a small number of classes (typically less than 10). When predicting a person's exact age, the above assumption may not always hold because K is often set at 80.

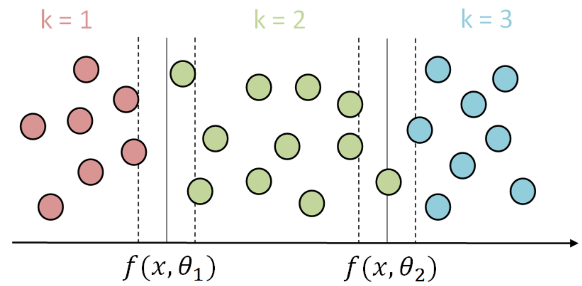


Figure 1. Ordinal ranking age estimation with a thresholds model in the kernel space of RED-SVM.

When K is large, it is more difficult to separate data by using parallel hyperplanes. Multiple Hyperplanes Ranker [19] can be used to capture more scattered data when K is large; however, this approach overlooks some potentially useful cues provided by a well ordered set of labels, as discussed in Section 1. To resolve this problem, we propose an *ordinal hyperplanes ranker*, called OHRank, which aggregates $K-1$ binary classifiers based on the order of the labels. Our experiment results demonstrate that the proposed approach outperforms RED-SVM, RankBoost, and Multiple Hyperplanes Ranker on the standard FG-NET database and the MORPH Album 2 database.

3. Ordinal Hyperplanes Ranker

For humans, it is easier to distinguish who is the older of two people than to determine the person's actual age. When inferring a person's age, we may compare the input face with the fact of many people whose ages are known, resulting in a series of comparisons, and then estimate the person's age by integrating the results. This process involves numerous pairwise preferences, each of which is obtained by comparing the input face to the faces in the database. However, exhaustive comparison of all faces is time consuming.

Since our approach only employs the relative order of labels, we treat the age labels y_i as a rank order, $y_i \in$

$\{1, \ldots, K\}$, where K is the number of labels (typically set as 80 in our approach). Then, for a given age k, we separate the dataset into two subsets, X_k^+ and X_k^-, as follows:

$$\begin{aligned} X_k^+ &= \{(x_i, +1) | y_i > k\} \\ X_k^- &= \{(x_i, -1) | y_i \leqslant k\}. \end{aligned} \tag{1}$$

Next, we use the two subsets to learn a binary classifier and conduct the query: "Is the face older than age k?" Each query reduces the age estimation task to a simple binary classification problem that determines which face is older. A series of query results imply the ordinal relationships between the age labels. Each query forms an exact binary classification problem that identifies the preferred classes. Actually, each problem has more data available for training (i.e., the datasets X_k^+ and X_k^-) than just the data for age k, and could utilize its own feature space for classification. After a series of such classifications, a set of preferences is derived and integrated for age estimation.

Let us consider on each subproblem. Given the training sets X_k^+ and X_k^- for an age label k, we introduce a cost sensitive setting to the subproblem. Let the cost of misclassifying the data for age label l in subproblem k be defined as $cost_k(l)$, where $k, l = 1...K$. We explain how the costs are set later in the following.

In the above scenario, the k-th subproblem is constructed from the age label k. In this subproblem, the cost of misclassifying data could differ according to the application. For example, when a person's exact age is very close to k, we do not care much whether he/she is more than k years old. However, when the age is far from k, we do care about the correctness of the inference of "being older than k." We use some popular performance evaluation measurements in age estimation applications as examples. In the *mean absolute error* (MAE) [10, 11, 12, 9, 4, 31, 26, 23, 14], large differences between the predicted and the exact ages contribute more to the total error; and in the *cumulative score* (CS) [10, 11, 12, 9, 4, 31, 26], the misclassification error in a tolerable range is set at zero. Hence, wrong inferences of an age to different ages contribute to different levels of errors. These measures reveal that the cost of misclassifying data in the subproblem k varies with the data labels. Recently, other measures that consider more practical issues have been proposed. For example, in [24], Ueki et al. introduced a measure that reflects the impact of misclassification that is adaptive to age and gender. In this paper, we focus on the cost sensitive settings for the MAE and CS measures; however, our approach can also be used for other measures.

For the i-th sample x_i, we define the cost of misclassifying x_i in the k-th subproblem as $c_k[i] = cost_k[y_i]$, where y_i is the age label of x_i and $c_k[i] \geqslant 0$. In studies of cost-sensitive learning, rescaling (or rebalancing) is one of the most popular techniques. The technique, which resets the importance of data according to the class costs, can be implemented by various approaches, e.g., data reweighting, data resampling, and moving decision thresholds. Note that applying the rescaling technique to multi-class data directly is not appropriate because it causes the inconsistency of costs problem [33]. However, for binary classification, the rescaling technique is often applicable [33, 32]. Because our subproblem is a simple binary classification problem, we use this technique to find a single hyperplane ranker for the k-th subproblem. We then use the data reweighted (or biased penalties) SVM to solve the k-th binary classification problem with cost sensitivities as follows:

$$\begin{aligned} \min_{w_k, b_k, \xi} \quad & \frac{1}{2}\langle w_k, w_k \rangle + C\left[\sum_i c_k[i]\, \xi_i\right] \\ \text{s.t.} \quad & z_k[i]\,(w_k^T \phi_k(x_i) + b_k) \geqslant 1 - \xi_i \\ & \xi_i \geqslant 0, \forall i, \end{aligned} \tag{2}$$

where $z_k[i] = +1$ if $x_i \in X_k^+$ and $z_k[i] = -1$ if $x_i \in X_k^-$, ϕ_k is an implicit mapping in the Hilbert space with a reproducible kernel function for its inner product evaluation; and (w_k, b_k) are the hyperplane parameters in the implicit feature space defined by ϕ_k.

As a single kernel can not fit all subproblems, we use multiple kernels. Hence, the kernel selected varies with the subproblem k. A cost-transformation technique called training set expansion and weighting [17] is applied to transform the biased-penalties SVM defined in Equation 2 into a standard SVM. To select the kernels for each subproblem, multiple kernel learning could be used. However, we simply apply cross validation to select a single kernel for each subproblem, and use LIBSVM [3] in our implementation. Since a kernel is selected for each subproblem, the kernels used for different subproblems may vary; thus, each subproblem can find its own feature space for casting.

Then, the discriminating function $f_k(x)$ used to model the confidence of "larger than k" is defined as

$$f_k(x) = \langle w_k, \phi_k(x) \rangle + b_k. \tag{3}$$

In the following, we introduce the costs set for MAE and CS in our implementation based on the rescaling technique. The performance measurement MAE is typically defined as

$$\text{MAE} = \sum_{j=1}^{N} |\overline{y_j} - y_j| / M, \tag{4}$$

where $\overline{y_j}$ is the estimated age, y_j is the ground truth age and M is the number of test images. In our work, the *absolute cost* associated with the MAE serves as the cost sensitivity function:

$$cost_k(l) = |l - k|, \text{ for } k = 1, ..., K. \tag{5}$$

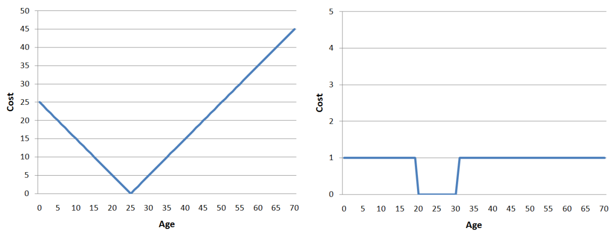


Figure 2. (a) Absolute cost function and (b) truncated cost function (L=5) for a 25-years-old person.

The CS performance measurement proposed by Geng et al. [10] is defined as

$$\text{CS}(L) = (M_{e<L}/M) \times 100\%, \tag{6}$$

where $M_{e<L}$ is the number of test images with the absolute error e less than the error level L. CS calculates the percentage of test data whose predicted error is less than the tolerance error level L.

The *truncated cost* associated with the CS serves as the cost sensitivity function:

$$cost_k(l) = \begin{cases} 0, \text{ if } (l - L) \leqslant k \leqslant (l + L) \\ 1, \text{ otherwise.} \end{cases} \tag{7}$$

Figures 2(a) and 2(b) illustrate the absolute cost function and the truncated cost function (with L=5), respectively, when the ground truth age is 25-years. In our experiments, the two cost functions are used in association with the corresponding performance measurements for age estimation.

The steps of the OHRank algorithm are as follows:
Algorithm (OHRank: Ordinal Hyperplanes Ranker):

1. For each k where $1 \leqslant k < K$,

   (a) Divide the original training data into two sets

   $$Z_k^+ = \{X_k^+, c_k[i]\} \tag{8}$$
   $$Z_k^- = \{X_k^-, c_k[i]\}, \tag{9}$$

   (b) Use a cost-sensitive binary classifier A_k to obtain a decision function f_k based on Z_k^+ and Z_k^-.

2. Construct a ranking rule r by collecting all the preference information

   $$r(x) = 1 + \sum_{k=1}^{K-1} [\![f_k(x) > 0]\!], \tag{10}$$

   where $[\![\cdot]\!]$ is 1 if the inner condition is true, and 0 otherwise.

Table 1. Age range distribution of face images in the FG-NET and MORPH databases.

Age Range	FG-NET (%)	MORPH (%)
0-9	37.03	0
10-19	33.83	8.94
20-29	14.37	26.04
30-39	7.88	32.16
40-49	4.59	24.58
50-59	1.50	7.37
60-69	0.80	0.82
70-77	0	0.09

In the OHRank algorithm, A_k could be any binary classifier. Without loss of generality, we use the biased penalties SVM defined in Equations 2 and 3.

To summarize, the proposed OHRank method constructs the subproblems based on the ordinal property. Compared to the Multiple Hyperplanes Ranker [19], the proposed algorithm reduces the number of comparisons from $K(K-1)/2$ to $(K-1)$ and utilizes all the data to solve every subproblem. For example, when $K = 80$, OHRank performs 40 times faster than the Multiple Hyperplanes Ranker. Moreover, in contrast to the one-versus-all strategy, which always generates data-imbalanced subproblems, OHRank reduces the likelihood of imbalanced subproblems and employs the ordinal information to construct more effective subproblems.

4. Experiments

4.1. Data Sets

We performed age estimation experiments on two benchmark age databases: FG-NET [1] and MORPH Album 2 [21]. FG-NET contains 1,002 color or gray facial images of 82 individuals with large variations in pose, expression and lighting. For each subject, there are more than ten images ranging from age 0 to age 69.

There are two scales of MORPH databases. Since MORPH Album 1 only contains a similar number of images of FG-NET, we use the MORPH Album 2 that is a larger-scale database in our experiments. MORPH Album 2 contains 55,608 facial images with about three images per person ranging from 16 to 77 years old. To reduce the variation between ethnic groups, we selected 5,492 images of people of Caucasian descent, so that cross-race influence can be avoided.

Table 1 details the age range distributions of the face images in the two databases; and Figures 3 and 4 show some examples taken from the databases. Note that the distributions of the two databases are quite different. We use AAM [5] as the feature extraction method in the experiments because it is capable of extracting both the shape and the ap-

Figure 3. Images of a person from childhood to adulthood in the FG-NET database.

Figure 4. Images of two people in the MORPH Album 2 database.

pearance features of human images. A number of methods [14, 10, 11] also use AAM for primitive or baseline feature extraction. Although other feature extraction tools have been proposed, AAM is still one of the most powerful age estimation techniques. For both databases, AAM extracts the number of features that preserve 95 percent of the variability.

4.2. Experiment Setup

We compare the performance of Multiple Hyperplanes Ranker (MHR) [19], RED-SVM [4], RankBoost [29], WAS [10], AGES [10], RUN1 [27], RUN2 [28], LARR [11], GP [31], and MTWGP [31] on the FG-NET database by using leave-one-person-out (LOPO), a popular test strategy, as suggested in [10, 11, 27, 28, 29, 31]. The parameters are determined via cross validation and a random search of the nearby parameter combinations. Among the above approaches, RankBoost, RED-SVM and MHR are ranking-based, and the others are classification- or regression-based approaches. For ranking, RankBoost uses a single hyperplane, RED-SVM employs parallel hyperplanes, and MHR employs arbitrary hyperplanes in a one-versus-one setting.

As a small database, recent studies on the improvement of FG-NET tend to be saturate. This can be reflected in the results (shown in Section 4.3) that many methods have closed errors. To evaluate the accuracy of our algorithm, we also make comparisons on the MORPH Album 2 database. In this database, we randomly split the data into 80% for training and 20% for testing over 30 trails. We then used five-fold cross validation to select the parameters from the training data. In the experiments, we compare the proposed approach OHRank with standard age estimation SVR, SVM, k-Nearest Neighbors (kNN), Back Propagation neural networks (BP), Binary Tree (BT), and the RED-SVM and RankBoost ranking approaches. The parametric configurations of the above methods are as follows. In SVR and SVM learning, LIBSVM [3] is used to evaluate the approaches. The RBF kernel function is used for evaluation and the associated parameters, C and γ, are selected by five-fold cross validation. The k in kNN is 30. BP is composed of a single layer with 100 neurons, and the number of output neurons is the same as that of the classes.

4.3. Experimental Results

Table 2 and Table 3 show the MAE results derived on the FG-NET database and the MORPH database respectively. The results demonstrate that the ranking-based approaches consistently outperform the regression-based and classification-based approaches. Among the ranking-based approaches, RankBoost and RED-SVM have been used in recent age estimation studies [29] and [4] respectively, and we implemented MHR for comparison. As RED-SVM is a cost-sensitive enhancement version of the approach in [22], it already has a cost sensitive property, although this point is not made clear in [4]. To be fair, we also added a cost-sensitive setting in the implementation of MHR. It is worth mentioning that the CSOVO approach proposed by Lin [17] can be regarded as a very similar approach to MHR. The difference lies in the former uses Ranking SVM as base ranker and the latter uses binary SVM with cost-sensitive property. Among all the compared methods, the proposed OHRank algorithm achieves the lowest MAE on both databases, as shown in Table 2 and Table 3.

Figures 5(a) and 5(b) show the CS curves for different error levels on both databases. When the CS is fixed, the smaller error level is more accurate. For all error levels, the proposed OHRank yileds the highest accuracy among the compared methods. Since FG-NET is a smaller database, we believe that the improvement on the MORPH Album 2 could be more meaningful and worth for further study.

Table 4 and Table 5 show the impact of different cost sensitive settings. As mentioned earlier, the absolute cost function (Equation 5) is more suitable for the MAE measure than other cost functions, such as the truncated function (Equation 7). However, the truncated cost function is more suitable for the CS measure than the absolute cost function. We employ both cost functions, and evaluate their performance based on MAE and CS. Table 4 shows the MAEs and CS at error level L=5 on both databases. As expected, the truncated cost function is better for CS and the absolute

Table 2. MAEs of the compared age estimation algorithms on the FG-NET database.

Method	MAE
OHRank (Absolute cost) [Ours]	**4.48**
MTWGP [31]	4.83
LARR [11]	5.07
MHR [19] with cost sensitivities	4.87
RED-SVM [4]	5.24
RankBoost [29]	5.67
RUN1 [27]	5.78
RUN2 [28]	5.33
GP [31]	5.39
SVR	5.91
AGES [10]	6.77
WAS [10]	8.06
SVM	7.25

Table 3. MAEs of the comapred age estimation algorithms on the MORPH Album 2 database.

Method	MAE
OHRank (Absolute cost) [Ours]	**6.07** $\pm$ 0.14
RED-SVM [4]	6.49 $\pm$ 0.17
SVR	6.99 $\pm$ 0.07
SVM	7.55 $\pm$ 0.08
KNN	9.39 $\pm$ 0.28
BP	10.03 $\pm$ 1.00
BT	11.97 $\pm$ 0.24

Table 4. The MAEs and CSs ($L = 5$) of ranking frameworks on (a) FG-NET and (b) MORPH.

FG-NET	Cost function	MAE	CS(%)
OHRank	Absolute	**4.48**	74.4
OHRank	Truncated	4.56	**74.7**

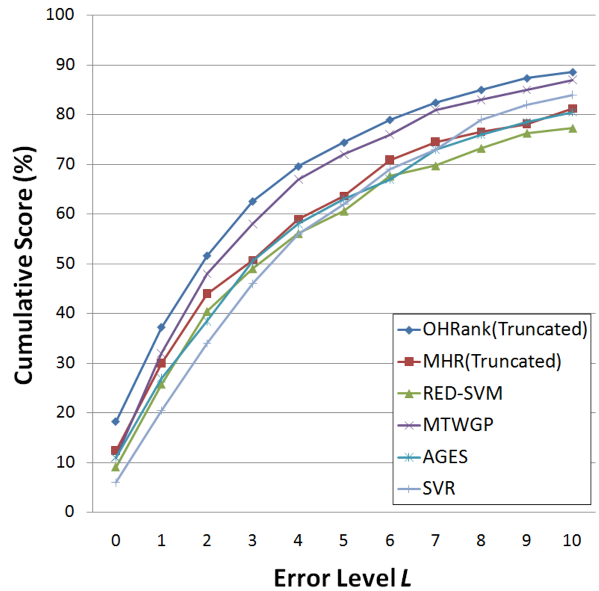
(a)

MORPH	Cost function	MAE	CS(%)
OHRank	Absolute	**6.07**	56.3
OHRank	Truncated	6.12	**56.5**

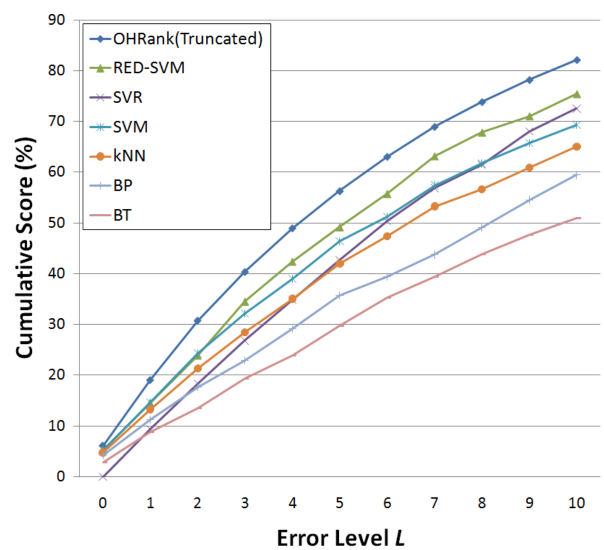
(b)

cost is better for MAE. The differences between the performance of employing different cost functions are not significantly high, and we owe this to the reason that common cost functions employed for age estimations are correlated to each other.

We also investigate the performance of OHRank with and without using cost sensitivities in an experiment based



(a)



(b)

Figure 5. CS curves of the error levels from 0 to 10 years of different age estimation algorithms on (a) FG-NET and (b) MORPH Album 2 databases.

on the MAE measure. Specifically, we conducted 10 trials of random splits the FG-NET database. The results are shown in Table 5, where the "Equal" cost function is defined as $cost_k(l) = 1$; in other words, it is equivalent to not using any cost functions. As shown in the table, both the absolute and truncated costs functions perform better than the Equal case. We can see from Table 5 that, although not extensively high, the associated cost-sensitive settings suggested in this paper can consistently enhance the performance for all cases.

Table 5. MAEs of OHRank using different cost functions with 10 trails of random splits on FG-NET.

FG-NET	Cost function	MAE
OHRank	Absolute	4.68 ± 0.41
OHRank	Truncated	4.78 ± 0.43
OHRank	Equal	4.82 ± 0.44

5. Conclusion

In this paper, we have proposed an ordinal hyperplanes ranker for age estimation based on information about the relative order of ages. The information of relative order between ages is more reliably employed than conventional ways of using it in our ranking framework. The age estimation problem is converted into a series of K subproblems of binary classifications according to the ordering property. The cost-sensitive property is introduced to each subproblem to further improve the performance. Our experimental results demonstrate that, for age estimation based on human faces, the proposed OHRank method outperforms other ranking-based methods, as well as multi-class-based and regression-based methods.

Acknowledgment

This work was supported in part by the National Science Council, Taiwan, under the grants NSC100-2631-H-001-013.

References

[1] The FG-NET aging Database, available at `http://www.fgnet.rsunit.com/`.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, 2005.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. A ranking approach for human age estimation based on face images. In *ICPR*, 2010.

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE TPAMI*, 2001.

[6] D. Cossock and T. Zhang. Subset ranking using regression. In *COLT*, 2006.

[7] E. Frank and M. Hall. A simple approach to ordinal classification. In *ECML*, 2001.

[8] Y. Freund, R. Iyer, R. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *JMLR*, 2003.

[9] Y. Fu and T. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE TMM*, 2008.

[10] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE TPAMI*, 2007.

[11] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE TIP*, 2008.

[12] G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bio-inspired features. In *CVPR*, 2009.

[13] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *NIPS*, 1999.

[14] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE TSMC-Part B*, 2004.

[15] L. Li and H.-T. Lin. Ordinal regression by extended binary classification. In *NIPS*, 2007.

[16] P. Li, C. Burges, Q. Wu, J. Platt, D. Koller, Y. Singer, and S. Roweis. Mcrank: Learning to rank using multiple classification and gradient boosting. In *NIPS*, 2007.

[17] H.-T. Lin. *From ordinal ranking to binary classification*. PhD thesis, 2008.

[18] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *ACM MM*, 2009.

[19] T. Qin, X. Zhang, D. Wang, T. Liu, W. Lai, and H. Li. Ranking with multiple hyperplanes. In *SIGIR*, 2007.

[20] N. Ramanathan, R. Chellappa, and S. Biswas. Age progression in human faces: A survey. *JVLC*, 2009.

[21] K. Ricanek Jr. and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *AFGR*, 2006.

[22] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *NIPS*, 2003.

[23] J. Suo, X. Chen, S. Shan, and W. Gao. Learning long term face aging patterns from partially dense aging databases. In *ICCV*, 2009.

[24] K. Ueki, M. Sugiyama, and Y. Ihara. Perceived Age Estimation under Lighting Condition Change by Covariate Shift Adaptation. In *ICPR*, 2010.

[25] C. Wang, Y. Su, C. Hsu, C. Lin, and H. Liao. Bayesian age estimation on face images. In *ICME*, 2009.

[26] B. Xiao, X. Yang, H. Zha, Y. Xu, and T. Huang. Metric learning for regression problems and human age estimation. In *PCM*, 2009.

[27] S. Yan, H. Wang, T. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *ICME*, 2007.

[28] S. Yan, H. Wang, X. Tang, and T. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *ICCV*, 2007.

[29] P. Yang, L. Zhong, and D. Metaxas. Ranking model for facial age estimation. In *ICPR*, 2010.

[30] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *ICB*, 2007.

[31] Y. Zhang and D. Yeung. Multi-task warped Gaussian process for personalized age estimation. In *CVPR*, 2010.

[32] Y. Zhang and Z. Zhou. Cost-sensitive face recognition. *IEEE TPAMI*, 2010.

[33] Z. Zhou and X. Liu. On multi-class cost-sensitive learning. In *AAAI*, 2006.