Sense Beauty by Label Distribution Learning

Yi Ren, Xin Geng*
MOE Key Laboratory of Computer Network and Information Integration,
School of computer Science and Engineering,
Southeast University, Nanjing, China
{y.ren, xgeng}@seu.edu.cn

Abstract

Beauty is always an attractive topic in the human society, not only artists and psychologists, but also scientists have been searching for an answer – what is beautiful. This paper presents an approach to learning the human sense toward facial beauty. Different from previous study, the human sense is represented by a label distribution, which covers the full range of beauty ratings and indicates the degree to which each beauty rating describes the face. The motivation is that the human sense of beauty is generally quite subjective, thus it might be inappropriate to represent it with a single scalar, as most previous work does. Therefore, we propose a method called *Beauty Distribution Transformation* (BDT) to covert the k-wise ratings to label distributions and propose a learning method called *Structural Label Distribution Learning* (SLDL) based on structural Support Vector Machine to learn the human sense of facial beauty.

1 Introduction

For centuries, artists and psychologists have been fascinated by the secret of beauty [Cross and Cross, 1971; Dion *et al.*, 1972; Alley and Cunningham, 1991; Pallett *et al.*, 2010], and recently, computer scientists took part in as well. They produced human-like machines to predict facial attractiveness, which can help to learn the elements affect human sense towards facial attractiveness [Eisenthal *et al.*, 2006; Kagian *et al.*, 2006; Gray *et al.*, 2010; Nguyen *et al.*, 2012].

There are some well known theories, such as "beauty is in the eye of the beholder", which suggests individual attractiveness varies from person to person, because of the difference of age, sex, culture, historical era, or personal history. However, several studies demonstrate high cross-cultural agreement in facial attractiveness [Cunningham *et al.*, 1995], which suggests that some specific features can lead to an attractive face, especially with the opinion that 'golden ratio' and 'symmetry' can optimize the attractiveness [Pallett *et al.*, 2010; Jones *et al.*, 2001].

In essence, most of the previous work by scientists, can be described as geometric or landmark feature methods, which mainly focus on landmark selection and features extracting. Aarabi et al. [2001] built a classification system based on 8 landmark ratios and collected a dataset of 80 images. Eisenthal et al. [2006] built two datasets of 92 images, and used an ensemble of features that include landmark distances and ratios. Later, Gray et al. [2010] presented a regression method using a hierarchical feed-forward model without the landmarks. Nguyen et al. [2012] considered this problem by multiple modalities like face, dressing and voice. They proposed a Dual-supervised Feature-Attribute-Task network and collected a Multi-Modality Beauty dataset, containing 1240 female instances. Xie et al.[2015] collect a novel face dataset with attractiveness ratings, namely SCUT-FBP dataset. Moreover, some specific work on portraits or selfies has been explored [Kalayeh *et al.*, 2015; Redi *et al.*, 2015], which considers not only the beauty of face, but also the aesthetic value.

To get the ratings, most previous work uses absolute ratings where some raters are presented with one image at a time and asked to give a score in certain range with a rule that the higher score means more attractive. As mentioned above, the range is 1-4 in [Aarabi *et al.*, 2001] and 1-7 in [Eisenthal *et al.*, 2006]. This form of ratings requires a number of raters to rate each image, in order to gather their opinions that is close to the fact. Then the median or the average of these ratings is regarded as the label of the image.

However, as human sense toward beauty is a subjective property, a single scalar value is insufficient to capture the true nature. Take two instances in SCUT-FBP dataset as the example, which is shown in Figure. 1. Though the two pictures have the equal average and median value, the raters do not perceive them equally. They rate the first picture more inconsistently than the second one, and some of them even rate two extremes. If the average or the median value is used to label these two pictures, we will lose the detail of the human sense, and regard the two pictures as the same. So, in this paper, we use the label distribution[Geng, 2016] to represent the human sense. The label distribution covers the full range of labels and indicates the degree to which each label describes the instance. Taking Figure 1 as an example, the five levels of attractiveness can be regarded as the labels, and the ratio of the raters labeling a certain level over the whole raters can

*Corresponding author.

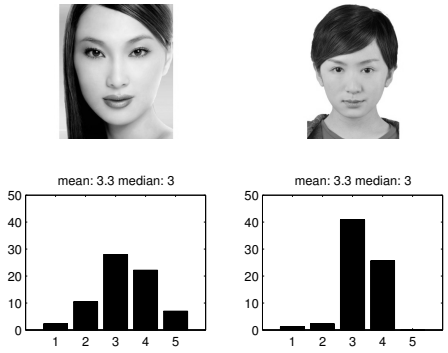

Figure 1: Two images with ratings. The histograms show the number of the raters giving the corresponding ratings.

be seen as the degree of that label describes the image. Representing facial beauty by label distribution has three advantages: firstly, it contains more information of the ratings, thus the description is closer to the true nature; secondly, with the fully description, it can well reflect the subjective property of human sense, and records the consistent and inconsistent among different individuals; last but not least, because of the closer to the nature and the fully reflection of the subjective property of human sense, it helps build a model to learn the human sense.

There are some shortages of the absolute ratings, i.e.,the ratings given by too few raters maybe highly biased, which cannot represent the general opinion; the order of presenting the images may affect the raters, e.g., when a more attractive image is showed before a less attractive one, the latter might get a lower score than showed after an equal or less attractive one; the raters may be affected by aesthetic fatigue when rate a large number of images. So, Gray et al. [2010] applied pairwise comparison for attractiveness study, which presents two pictures at one time and asks raters to choose the prefer one. Nguyen et al. [2012] proposed a k-wise comparison method, which presents k images each time instead of two for a quicker collection. The raters are asked to rank these k images from the most attractive to the least attractive. Then they convert the comparisons into absolute scores by some methods. The scores are regarded as the label of the image.

In this paper, for the datasets containing the whole ratings of the raters, we regard each beauty level as a label in the label distribution, and the ratio of the raters giving a specific rating over the total number of the raters as the descriptor degree of the labels. As for the datasets where only k-wise comparison results are available, we propose an algorithm called *Beauty Distribution Transformation* (BDT) to convert the comparisons to label distribution. Then we build a model of *label distribution learning* (LDL), which is a learning process on the instances labeled by label distribution. Different from the previous LDL algorithms where a label distribution is always regarded as a conditional probability, we regard the label distributions as a structure and propose a learning method called *Structural Label Distribution Learning* (SLDL) based on the structural SVM to learn a label distribution model that considers the correlation among labels.

The main contributions of this paper can be summarized as follows.

- Apply label distribution to represent the human sense of beauty, which matches the nature of the subjective human sense better.
- Propose a novel method to covert the k-wise comparisons to label distributions.
- Propose Structural Label Distribution Learning (SLDL) based on structural SVM to predict human sense toward facial beauty.

The rest of this paper is organized as follows. Section 2 introduces the label distribution representation for facial beauty and the BDT algorithm of converting the comparisons to label distributions. Section 3 proposes the SLDL algorithm for the prediction task. Section 4 reports the experimental results. Finally, a conclusion of this work is made in section 5.

2 Label Distribution Representation

Different from the previous studies, we use the label distribution to represent the human sense of facial image. Because the human sense is a quite subjective property, and a scalar can not fully describe the inconsistent among the individuals. The label distribution covers the whole possible labels, and assigns a real number to each label called *description degree*, representing the degree to which the label describes the instance.

In this work, the image is denoted by $\boldsymbol{x}$, and the particular n-th image is denoted by $\boldsymbol{x}_n$. The label $\boldsymbol{y}$ can be regarded as the degree of attractiveness, which is assumed to be five levels, i.e., "excellent", "good", "fair", "ordinary", and "poor", corresponding to the numerical beauty labels of 5, 4, 3, 2, and 1, respectively. The particular l-th label is denoted by y_l, and its description degree to the instance $\boldsymbol{x}_n$ is denoted by $d_{\boldsymbol{x}_n}^{y_l}$. Assume that $d_{\boldsymbol{x}}^{y} \in [0, 1]$, and all the labels in set can always fully describe the instance , thus $\sum_l d_{\boldsymbol{x}_n}^{y_l} = 1$. The label distribution of $\boldsymbol{x}_n$ is denoted by $\boldsymbol{d}_n = \{d_{\boldsymbol{x}_n}^{y_1}, d_{\boldsymbol{x}_n}^{y_2}, \cdots, d_{\boldsymbol{x}_n}^{y_c}\}$, where c is the number of possible label values and equals to 5 in this work.

As mentioned before, the labels of the images are generally collected by two rating methods, the absolute methods and the comparison methods. For the datasets collected by the absolute methods, which contain the whole ratings given by the raters, we compute the label distributions by Eq. (1), assuming that the number of raters is large enough to be representative.

$$d_{\boldsymbol{x}_n}^{y_l} = \frac{\sum_{i=1}^{m^{(n)}} \mathbb{I}(r_i^{(n)} = y_l)}{m^{(n)}}. \tag{1}$$

where $m^{(n)}$ denotes the total number of raters giving ratings to $\boldsymbol{x}_n$, $r_i^{(n)}$ indicate the label of $\boldsymbol{x}_n$ given by the i-th rater, $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if $r_i^{(n)}$ equals to y_l, returns 0 otherwise.

As for the dataset collected by the comparison methods, if the dataset contains k-wise comparisons, we first convert

them into $\binom{k}{2}$ pairwise preferences by listing all the potential pairs, e.g., when k is 10, we can obtain 45 pairwise preferences. Since label distribution shares the same constraints with probability distribution, $d_{\boldsymbol{x}}^{y}$ can be regarded as a form of conditional probability, i.e., $d_{\boldsymbol{x}}^{y} = P(y|\boldsymbol{x})$. Then the problem of converting pairwise preferences to label distributions can be seen as a joint probability problem. We apply the Log Maximum Likelihood estimation to solve this problem as follow,

$$\begin{aligned}
&\arg\max_{\boldsymbol{d}_n, n=1,\cdots,N} \quad \sum_{(i,j)\in S} \log P(\boldsymbol{x}_i \succ \boldsymbol{x}_j) \qquad (2)\\
&= \sum_{(i,j)\in S} \log \sum_{y_p > y_q} P(y_p|\boldsymbol{x}_i) P(y_q|\boldsymbol{x}_j)\\
&= \sum_{(i,j)\in S} \log \sum_{y_p > y_q} d_{\boldsymbol{x}_i}^{y_p} d_{\boldsymbol{x}_j}^{y_q}\\
&\text{s.t.} \quad \forall l = 1,\cdots,c, \forall n = 1,\cdots,N\\
&\quad 0 < d_{\boldsymbol{x}_n}^{y_l} < 1,\\
&\quad \sum_l d_{\boldsymbol{x}_n}^{y_l} = 1.
\end{aligned}$$

Suppose the total number of instances is N, and the pairwise preferences are contained in the set S. Each pair in S can be recorded as a form of (i, j), which means the rater prefer $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$. $P(\boldsymbol{x}_i \succ \boldsymbol{x}_j)$ stands for the probability of the human sense that prefer $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$. It can be computed as a joint probability that considers all the conditions, where $\boldsymbol{x}_i$ is rated a higher level/label y_p than the level/label y_q that $\boldsymbol{x}_j$ is rated. With the definition of the description degree, we limit $d_{\boldsymbol{x}}^{y}$ in the range $[0, 1]$ and sum up to 1.

Considering the theory "high cross-cultural agreement in facial attractiveness" [Cunningham *et al.*, 1995] , we add a regularization term of the variance values of the label distributions of beauty, to punish too much inconsistency of the human sense while learning from the pairwise preferences,

$$\begin{aligned}
&\arg\max_{\boldsymbol{d}_n, n=1,\cdots,N} \quad \sum_{(i,j)\in S} \log P(\boldsymbol{x}_i \succ \boldsymbol{x}_j) - \lambda \sum_n v(\boldsymbol{d}_n) \qquad (3)\\
&= \sum_{(i,j)\in S} \log \sum_{y_p > y_q} P(y_p|\boldsymbol{x}_i) P(y_q|\boldsymbol{x}_j)\\
&\quad -\lambda \sum_n \sum_l P(y_l|\boldsymbol{x}_n)(y_l - \sum_{l'} y_{l'} P(y_{l'}|\boldsymbol{x}_n))^2\\
&= \sum_{(i,j)\in S} \log \sum_{y_p > y_q} d_{\boldsymbol{x}_i}^{y_p} d_{\boldsymbol{x}_j}^{y_q}\\
&\quad -\lambda \sum_n \sum_l d_{\boldsymbol{x}_n}^{y_l} (y_l - \sum_{l'} y_{l'} d_{\boldsymbol{x}_n}^{y_{l'}})^2\\
&\text{s.t.} \quad \forall l = 1,\cdots,c, \forall n = 1,\cdots,N\\
&\quad 0 < d_{\boldsymbol{x}_n}^{y_l} < 1,\\
&\quad \sum_l d_{\boldsymbol{x}_n}^{y_l} = 1.
\end{aligned}$$

where $v(\boldsymbol{d}_n)$ computes the variance of all the labels assigned by the raters to the n-th image, and λ controls the tradeoff between the main target function and the regularizing term. This algorithm is named as *Beauty distribution Transformation* (BDT) algorithm, and solved by BFGS[Malouf, 2002]. Because BFGS avoids explicit calculation of the Hessian matrix and performs much more efficiently than standard gradient descending methods.

3 Structural Label Distribution Learning

The main purpose of our work is to train a model that maps from the image to the label distribution. In most previous work[Geng, 2016; Geng and Xia, 2014; Geng *et al.*, 2013], the description degree is regarded as a form of conditional probability, i.e., $d_{\boldsymbol{x}}^{y} = P(y|\boldsymbol{x})$, and a parametric model $p(y|\boldsymbol{x}; \boldsymbol{\theta})$ is learned. However, there is some correlation among the labels in this work, e.g., when one rater gives a certain rating, it is highly possible that the other raters will give the ratings similar to the former rating; take the absolute rating into consideration, one rater may rate slightly different because of the order of showing pictures or some other influence factors that mentioned before. Thus, we take $\boldsymbol{d} = \{d_{\boldsymbol{x}}^{y_1}, d_{\boldsymbol{x}}^{y_2}, \cdots, d_{\boldsymbol{x}}^{y_c}\}$ as a special structure, which can be learned simultaneously in order to make use of the correlation. Then the problem of structural label distribution learning can be formulated as follow,

> Let $X = \mathbf{R}^b$ denotes the input space and $Y = \{y_1, y_2, \cdots, y_c\}$ denotes the complete set of labels. $d_{\boldsymbol{x}}^{y}$ is the description degree of y to $\boldsymbol{x}$. $\boldsymbol{d}_n = \{d_{\boldsymbol{x}_n}^{y_1}, d_{\boldsymbol{x}_n}^{y_2}, \cdots, d_{\boldsymbol{x}_n}^{y_c}\}$ is the label distribution of $\boldsymbol{x}_n$, where $d_{\boldsymbol{x}}^{y_l} \in [0, 1]$ and $\sum_{l=1}^{c} d_{\boldsymbol{x}}^{y_l} = 1$. Suppose all possible $\boldsymbol{d}$'s span a special space V called structural label distribution space. Given a training set $D = \{(\boldsymbol{x}_1, \boldsymbol{d}_1), (\boldsymbol{x}_2, \boldsymbol{d}_2,), \cdots, (\boldsymbol{x}_N, \boldsymbol{d}_N)\}$, the goal of structural label distribution learning is to learn a structural function $f\colon X \longrightarrow V$ from D.

Suppose a discriminant function $F\colon X \times V \longrightarrow \mathbf{R}$ over the input and output pairs. Ideally, it achieves the maximum when $\boldsymbol{d}$ is the true label distribution of the input $\boldsymbol{x}$. Assuming F is linear in some combined feature representation of $\boldsymbol{x}$ and $\boldsymbol{d}$, represented as $\psi(\boldsymbol{x}, \boldsymbol{d})$,

$$F(\boldsymbol{x}, \boldsymbol{d}; \boldsymbol{w}) = \langle \boldsymbol{w}, \psi(\boldsymbol{x}, \boldsymbol{d}) \rangle. \qquad (4)$$

The specific form of $\psi(\boldsymbol{x}, \boldsymbol{d})$ is based on the nature of the task and will be introduced later. Then the hypotheses f is formulated as,

$$f(\boldsymbol{x}; \boldsymbol{w}) = \arg\max_{\boldsymbol{d}\in V} F(\boldsymbol{x}, \boldsymbol{d}; \boldsymbol{w}), \qquad (5)$$

where $\boldsymbol{w}$ denotes a parameter vector. The goal of structural label distribution learning is to find the $\boldsymbol{w}$ that can maximize F as the label distribution is close to the $\boldsymbol{d}_i$, given the instance $\boldsymbol{x}_i$. So, the purpose of our algorithm is to minimize the risk,

$$R_h^{\triangle}(f) = \int_{X\times V} \triangle(\boldsymbol{d}, f(\boldsymbol{x})) dh(\boldsymbol{x}, \boldsymbol{d}). \qquad (6)$$

where $\triangle(\boldsymbol{d}, \overline{\boldsymbol{d}})$ quantifies the loss associated with a prediction $\overline{\boldsymbol{d}}$, if the true label distribution is $\boldsymbol{d}$. $h(\boldsymbol{x}, \boldsymbol{y})$ denotes the data generating distribution. We assume that h is unknown, but the pairs of input $\boldsymbol{x}$ and the corresponding output $\boldsymbol{d}$ are generated from the training set D according to h is given. We solve this problem by a structural Support Vector Machine [Joachims *et al.*, 2009] with 1-slake formulation:

$$\begin{aligned}
&\min_{\boldsymbol{w}, \xi \geqslant 0} \quad \tfrac{1}{2}||\boldsymbol{w}||^2 + C\xi \qquad (7)\\
&\text{s.t.} \quad \forall (\overline{\boldsymbol{d}}_i, \cdots, \overline{\boldsymbol{d}}_N) \in V^n :\\
&\quad \tfrac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{w}, \delta\psi_i(\overline{\boldsymbol{d}}_i) \rangle \geqslant \tfrac{1}{N} \sum_{i=1}^{N} \triangle(\boldsymbol{d}_i, \overline{\boldsymbol{d}}_i) - \xi.
\end{aligned}$$

where $\delta\psi_i(\overline{\boldsymbol{d}}_i) = \psi(\boldsymbol{x}_i, \boldsymbol{d}_i) - \psi(\boldsymbol{x}_i, \overline{\boldsymbol{d}}_i)$, ξ is a slack variable, and C controls slackness.

As $\psi(\boldsymbol{x}, \boldsymbol{d})$ is a combined feature representation of $\boldsymbol{x}$ and $\boldsymbol{d}$, we divide it into two parts, following the work in [Wu *et al.*, 2011]. The first part reflects the correlation between the input features and the output label distributions,

$$\psi_1(\boldsymbol{x}, \boldsymbol{d}) = \boldsymbol{x} \otimes \boldsymbol{d} = [x_1 d_{\boldsymbol{x}}^{y_1}, \cdots, x_b d_{\boldsymbol{x}}^{y_c}], \tag{8}$$

where $\otimes$ is the tensor product, which computes all the possible combinations of x and $d_{\boldsymbol{x}}^{y}$ in each dimensional. The second part describes the interaction among basic rating distributions,

$$\begin{aligned}\psi_2(\boldsymbol{d}) = [&d_{\boldsymbol{x}}^{y_1} d_{\boldsymbol{x}}^{y_2}, \cdots, d_{\boldsymbol{x}}^{y_1} d_{\boldsymbol{x}}^{y_c}, d_{\boldsymbol{x}}^{y_2} d_{\boldsymbol{x}}^{y_3}, \\ &\cdots, d_{\boldsymbol{x}}^{y_2} d_{\boldsymbol{x}}^{y_c}, \cdots, d_{\boldsymbol{x}}^{y_{c-1}} d_{\boldsymbol{x}}^{y_c}].\end{aligned} \tag{9}$$

Just like the form of tensor product, it computes all the possible combinations between two $d_{\boldsymbol{x}}^{y}$ in different dimensional without repeat. Thus the definition of $\psi(\boldsymbol{x}, \boldsymbol{d})$ is

$$\psi(\boldsymbol{x}, \boldsymbol{d}) = [\psi_1(\boldsymbol{x}, \boldsymbol{d}), \psi_2(\boldsymbol{d})]. \tag{10}$$

There are various distance measures that are applicable to compare two distribution [Cha, 2007]. In this task, we choose the widely used measure, the Euclidean distance as the loss function $\triangle(\boldsymbol{d}, \overline{\boldsymbol{d}})$

$$\triangle(\boldsymbol{d}, \overline{\boldsymbol{d}}) = dis_{Euc} = \sqrt{\sum_{l=1}^{c} |d_{\boldsymbol{x}}^{y_l} - \overline{d}_{\boldsymbol{x}}^{y_l}|^2}. \tag{11}$$

Of course, other distance measures can also be applied depending on the dataset and the task.

The key challenge in solving Eq. (7) is the infinite number of constraints. As V is defined as a space spanned by all possible $\boldsymbol{d}$'s, there are infinite conditions that can match the constrains. Following [Joachims *et al.*, 2009], we apply a cutting-plane method to solve this problem. The cutting-plane method is implemented as a variable selection approach, which helps us build a polynomially-sized subset Q of the infinite constrains with a precision of at least ϵ. Then, Eq. (7) can be rewritten as,

$$\begin{aligned}\min_{\boldsymbol{w}, \xi \geqslant 0} \quad & \tfrac{1}{2}||\boldsymbol{w}||^2 + C\xi \\ \text{s.t.} \quad & \forall(\overline{\boldsymbol{d}}_1, \cdots, \overline{\boldsymbol{d}}_N) \in Q : \\ & \tfrac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{w}, \delta\psi_i(\overline{\boldsymbol{d}}_i)\rangle \geqslant \tfrac{1}{N}\sum_{i=1}^{N} \triangle(\boldsymbol{d}_i, \overline{\boldsymbol{d}}_i) - \xi.\end{aligned} \tag{12}$$

The pseudocode of the algorithm is depicted in Algorithm 1. The algorithm proceeds by finding the most "violated constrain", which is used to update the working set Q. In each iteration, Eq. (12) is solved with the current working set Q (line 4). After $(\boldsymbol{w}, \xi)$ is updated, we find the "most violated" constraint (line 6). If it is violated by more than ϵ (line 8), we add it into the working set Q(line 9). The iteration continues until no constraint has been added into the working set Q(line 11).

Algorithm 1 Structural Label Distribution Learning(SLDL)

1: Input: $D = \{(\boldsymbol{x}_1, \boldsymbol{d}_1), \cdots, (\boldsymbol{x}_N, \boldsymbol{d}_N)\}, C, \epsilon$
2: $Q \longleftarrow \varnothing$
3: **repeat**
4: compute $(\boldsymbol{w}, \xi)$ in Eq.(12)
5: **for** $i = 1, \cdots, N$ **do**
6: $\hat{\boldsymbol{d}}_i \longleftarrow \arg\max_{\hat{\boldsymbol{d}}_i \in V} \triangle(\boldsymbol{d}_i, \hat{\boldsymbol{d}}_i) + \langle \boldsymbol{w}, \psi(\boldsymbol{x}_i, \hat{\boldsymbol{d}})\rangle$
7: **end for**
8: **if** $\frac{1}{N}\sum_{i=1}^{N} \triangle(\boldsymbol{d}_i, \hat{\boldsymbol{d}}_i) - \frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{w}, \delta\psi_i(\hat{\boldsymbol{d}}_i)\rangle > \xi + \epsilon$ **then**
9: $Q \longleftarrow Q \bigcup (\hat{\boldsymbol{d}}_1, \cdots, \hat{\boldsymbol{d}}_N)$
10: **end if**
11: **until** Q has no change
12: **return** $(\boldsymbol{w}, \xi)$

Finding the "most violated" constraint in line 6 can be seen as a primary quadratic programming problem as

$$\begin{aligned}\arg\max_{\hat{\boldsymbol{d}}} \quad & \triangle(\boldsymbol{d}, \hat{\boldsymbol{d}}) + \langle \boldsymbol{w}, \psi(\boldsymbol{x}, \hat{\boldsymbol{d}})\rangle \\ \text{s.t.} \quad & \forall l = 1, \cdots, c, \quad 0 < d_{\boldsymbol{x}}^{\hat{y}_l} < 1 \\ & \sum_{l=1}^{c} d_{\boldsymbol{x}}^{\hat{y}_l} = 1,\end{aligned} \tag{13}$$

which can be solved by conventional optimization techniques. Here we choose BFGS as the advantages mentioned before.

4 Experiments

4.1 Datasets

There are a few publicly available datasets for facial beauty assessment. But most of them only contain the average score of the images, thus are not suitable to our task. In the experiments, we use the SCUT-FBP dataset[Xie *et al.*, 2015] and the Multi-Modality Beauty (M^2B) dataset [Nguyen *et al.*, 2012], which are just fit to the two conditions mentioned before, the former contains the whole ratings of each image, and the latter contains the k-wise comparisons.

SCUT-FBP dataset contains 1500 instances with the original facial images in the size of 350×350 pixels and the ratings given by 75 raters. The ratings were collected by randomly showing one image at a time, and asking the raters to rate the attractiveness within five degrees. With the ratings given by the raters, we generate the label distributions of each instance by Eq. (1). The features of the images are extracted by three popular descriptors, i.e., LPB [Ojala *et al.*, 2002] with a cell size of 64×64 pixels; HOG [Dalal and Triggs, 2005] with a cell size of 32×32 pixels; Gabor filter [Jain and Farrokhnia, 1991] with 2 scales and 4 orientations. Since the features we extracted are high-dimensional, we use PCA to reduce the dimensionality to 300.

M^2B dataset contains 1240 instances that includes the information of their faces, dressings and voices. In this task, we only use the part of the faces, which contains 1240 facial images in the size of 128×128 pixels. A k-wise comparison method (k = 10) is used in this data to collect ratings. The

Algorithm	RA
Ranking SVM	68.48%
BDT	**74.98%**

Table 1: The experimental results of the transformation task on the M^{2}B dataset

Algorithm	MAE	RMSE
sDFAT	0.4065	0.5647
SVR	0.3549	0.4643
k-NN	0.3920	0.5116
SLDL	**0.3015**	**0.4076**

Table 2: The experimental results of the predicting task on SCUT-FBP dataset

raters are showed with randomly selected ten images of faces at the same time, and asked to sort them from the most attractive one to the least attractive one. As suggested in [Nguyen *et al.*, 2012], we extracted features by LBP, Gabor filter and Color moment with the same settings, and use PCA to reduce the dimensionality to 250.

4.2 The Label Distribution Transformation

For the dataset collecting ratings by the absolute methods, we generate the label distributions by Eq. (1), which can be seen as the ground truth in some aspect.

As for the comparison methods, we convert the comparison results to the label distributions by BDT algorithm. We do the experiment on the M^{2}B dataset, and set λ to 20. To prove the accuracy, BDT algorithm is compared with the conventional pairwise ranking method Ranking SVM following the settings in [Nguyen *et al.*, 2012]. We measure the performance by comparing the results with the pairwise preferences obtained from k-wise comparisons. We name the measurement as *Ranking Accuracy* and define it as follow,

$$RA = \frac{\sum\limits_{(i,j)\in S} \mathbb{I}(\overline{\boldsymbol{x}}_i \succ \overline{\boldsymbol{x}}_j)}{M} \tag{14}$$

where M is the total number of pairs in set S. (i,j) means $\boldsymbol{x}_i$ is preferred to $\boldsymbol{x}_j$, and $\overline{\boldsymbol{x}}$ represents the predicted results. $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if $\overline{\boldsymbol{x}}_i \succ \overline{\boldsymbol{x}}_j$ and return 0 otherwise.

As the description degree can be regarded as a conditional probability, we define the expectation in Eq. (15) to rank the images, i.e., higher expectation value means more attractive. The expectation is computed as

$$E(\boldsymbol{x}) = \sum_{l=1}^{c} y_l d_{\boldsymbol{x}}^{y_l}; \tag{15}$$

For Ranking SVM, the higher rank means more attractive.

Table 1 shows the performance of Ranking SVM and BDT, where BDT achieves a better result. It proves that representing human sense by label distributions is much closer to the true nature.

Algorithm	MAE	RMSE	RA
sDFAT	0.6272	0.7969	63.37%
SVR	0.6014	0.7752	62.41%
k-NN	0.8077	1.0050	54.18%
SLDL	**0.5927**	**0.7723**	**65.14%**

Table 3: The experimental results of the predicting task on M^{2}B dataset

4.3 The Human Sense Prediction

The main purpose of our work is to learn the human sense. So, in this subsection, we first compare SLDL with the existing facial beauty assessment methods DFAT[Nguyen *et al.*, 2012] and some regression methods that are used to solve the facial attractiveness assessment problem in the early work[Kagian *et al.*, 2006; Eisenthal *et al.*, 2006], i.e., SVR and k-NN.

On the SCUT-FBP dataset, for the previous methods, i.e, DFAT, SVR and k-NN, we take the mean value of the ratings given by the whole raters to each image as the input, just as the previous work does; As for the SLDL, we use the label distributions that transformed by Eq. (1). DFAT [Nguyen *et al.*, 2012] is proposed as a multi-modality model, which considers face, dressing and voice. It builds a network from features to attributes, then to attractiveness scores. However, as the attributes are publicly unavailable and need manually annotated, we cut off the attribute layer, as our task is to learn a model that maps from the face to the label distribution straightly. The regularization parameter in simplified DFAT network is set to 0.005. SVR is set with a linear kernel. The k in k-NN is set to 1, and the distance is computed by the Euclidean distance. SLDL is set with a linear kernel and C is 400. Two traditional criteria are adopted to measure the performance of the compared methods, i.e., mean absolute error(MAE) and root mean squared error(RMSE). However, the output of SLDL is the label distributions, we compute the expectation by Eq. (15) for the measurement. Ten-fold cross validation is conducted.

On the M^{2}B dataset, we use the results of BDT to train the model, because of the higher Ranking Accuracy. For the previous methods, we take the expectation computed by Eq. (15) as the input. As for SLDL, we use the label distributions converted by BDT as the input. The regularization parameter in simplified DFAT network is set to 0.001. SVR is set with a linear kernel. The k in k-NN is set to 1, with the Euclidean distance. SLDL is set with a linear kernel and C is 400. Apart from MAE and RMSE, RA is adopted to measure the performance, as well. The expectation of the label distributions predicted by SLDL is used for measurement. Ten-fold cross validation is conducted, where we divide images into ten folds, and the pairs only containing the images in the test fold are selected for RA measurement.

Table 2 reports the comparative results between SLDL and previous methods, i.e., sDFAT, SVR, k-NN on the SCUT-FBP dataset, and Table 3 reports the results on the M^{2}B dataset. On both dataset, the best performance is highlighted by boldface. As shown in the tables, SLDL performs the best, which proves that the label distribution can well model the human

Algorithm	Evaluation						Average rank
	Chebyshev↓	Clark↓	Sϕrensen ↓	Topsϕe ↓	Cosine ↑	Intersection ↑	
PT-Bayes	0.2895(4)	1.4373(5)	0.3538(3)	0.2531(3)	0.6462(3)	0.8002(2)	3.33
PT-SVM	0.4324(7)	1.5787(7)	0.4933(6)	0.4342(7)	0.5067(6)	0.5729(7)	6.67
AA-kNN	0.2262(2)	1.2002(2)	0.2692(2)	0.1650(2)	0.7308(2)	**0.8542(1)**	1.83
AA-BP	0.2859(3)	1.3684(4)	0.3621(4)	0.2545(4)	0.6379(4)	0.7806(4)	3.83
SA-IIS	0.3939(6)	1.3385(3)	0.4957(7)	0.4091(5)	0.5043(7)	0.6527(6)	5.67
SA-BFGS	0.3643(5)	1.5729(6)	0.4832(5)	0.4172(6)	0.5374(5)	0.7454(5)	5.33
SLDL	**0.2062(1)**	**1.0039(1)**	**0.2196(1)**	**0.1273(1)**	**0.8984(1)**	0.7810(3)	**1.33**

Table 4: The experimental results of the predicting task on SCUT-FBP dataset compared with previous label distribution methods

sense toward beauty. Though, in Table 3, SLDL is slightly better than SVR in MAE and RMSE, but it is quite better in RA, which suggests the closer to the true nature. As the label distributions can contain more original information, it is helpful to learn the human sense.

Also, we compare the SLDL with the previous LDL methods, i.e., PT-Bayes, PT-SVM with the linear kernel, AA-kNN with k set to 4, AA-BP with 80 units in hidden layer, SA-IIS, and BFGS following the same settings in [Geng, 2016]. We conduct the ten-fold cross validation on the SCUT-FBP dataset. Six measures in [Cha, 2007] are selected to evaluate the performances, i.e., Chebyshev distance, Clark distance, Sϕrensen distance, Topsϕe distance, Cosine similarity, and Intersection similarity, which are from different families summarized in [Cha, 2007]. We do not use the same measures as [Geng, 2016] suggests. Because there are some label distributions valuing zero in the two datasets, Kullback-Leibler distance and Canberra distance are unsuitable to measure the performance. The former will cause a situation that the molecule is divided by zero, the latter magnifies the errors where zero appears. Thus we select two other measures from the same family instead.

Table 4 reports the experimental results for prediction task on the SCUT-FBP dataset compared with previous label distribution methods. For each measures, "↓" indicates "the smaller the better", while "↑" indicates "the larger the better". On each measure, the algorithms are ranked in decreasing order of their performance, and the best performance is highlighted by boldface. The ranks are given in the parentheses right after the measure values and the average ranks are given in the last column. From the results, we can see, SLDL ranks the 1*st* on five over six measures, and gets the lowest average rank, which indicates a best performance. However, it does not perform best in the "Intersection" evaluation, because the "Intersection is computed by $\sum_{i=1}^{d} min(P_i, Q_i)$, and SLDL considers the correlation among the labels, thus some results are smoother than the ground truth, which may lead to slightly worse results. However, the better performance of SLDL proves that considering the correlation among the labels can improve the prediction precision.

4.4 Parameter Analysis

Figure 2 shows the result of BDT trained with different λ in $\{0.1, 10, 20, 30\}$. The Ranking Accuracy increases when λ increases within 20, and goes down when λ continues increasing. It reveals that too small or too large value of λ can both lead to performance deterioration, which proves that the cross-cultural agreement does existed in some extent, but accompanied with the individual differences. This further supports our idea of using label distribution in the learning of human sense of facial beauty.

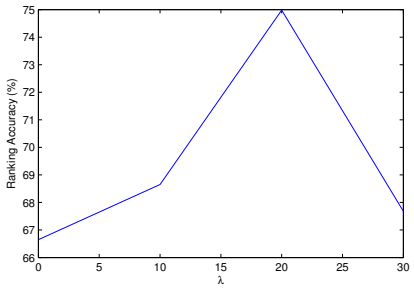

Figure 2: The Ranking Accuracy of BDT with different value of λ on the M^{2}B dataset.

5 Conclusion

In this paper, we propose to use the label distribution to represent human sense toward facial beauty, instead of a scalar value that is commonly used in many previous work. The reason is that, the human sense is quite subjective, so that a scalar cannot sufficiently represent it. Toward this purpose, we propose a new method BDT to transfer the rating comparisons to the label distributions, and propose an SLDL algorithm based on structural SVM, which considers the correlation among labels and has a better performance on the predicting task than previous label distribution learning algorithms on the facial beauty dataset.

Acknowledgments

This research was supported by National Science Foundation of China (61622203, 61232007), Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), Collaborative Innovation Center of Novel Software Technology and Industrialization, and Collaborative Innovation Center ofWireless Communications Technology.

References

[Aarabi *et al.*, 2001] Parham Aarabi, Dominic Hughes, Keyvan Mohajer, and Majid Emami. The automatic measurement of facial beauty. In *2001 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2644–2647, Tucson, Arizona USA, 2001.

[Alley and Cunningham, 1991] Thomas R Alley and Michael R Cunningham. Averaged faces are attractive, but very attractive faces are not average. *Psychological science*, 2(2):123–125, 1991.

[Cha, 2007] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[Cross and Cross, 1971] John F Cross and Jane Cross. Age, sex, race, and the perception of facial beauty. *Developmental Psychology*, 5(3):433, 1971.

[Cunningham *et al.*, 1995] Michael R Cunningham, Alan R Roberts, Anita P Barbee, Perri B Druen, and Cheng-Huan Wu. " their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, 68(2):261, 1995.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893, San Diego, CA, USA, 2005.

[Dion *et al.*, 1972] Karen Dion, Ellen Berscheid, and Elaine Walster. What is beautiful is good. *Journal of personality and social psychology*, 24(3):285, 1972.

[Eisenthal *et al.*, 2006] Yael Eisenthal, Gideon Dror, and Eytan Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2006.

[Geng and Xia, 2014] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1837–1842, Columbus, OH, USA, 2014.

[Geng *et al.*, 2013] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013.

[Geng, 2016] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[Gray *et al.*, 2010] Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong. Predicting facial beauty without landmarks. In *European Conference on Computer Vision*, pages 434–447, Crete, Greece, 2010.

[Jain and Farrokhnia, 1991] Anil K Jain and Farshid Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186, 1991.

[Joachims *et al.*, 2009] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[Jones *et al.*, 2001] Benedict C Jones, Anthony C Little, Ian S Penton-Voak, Bernard P Tiddeman, D Michael Burt, and David I Perrett. Facial symmetry and judgements of apparent health: support for a good genes explanation of the attractiveness–symmetry relationship. *Evolution and human behavior*, 22(6):417–429, 2001.

[Kagian *et al.*, 2006] Amit Kagian, Gideon Dror, Tommer Leyvand, Daniel Cohen-Or, and Eytan Ruppin. A human-like predictor of facial attractiveness. In *Advances in Neural Information Processing Systems*, pages 649–656, Hyatt Regency Vancouver, in Vancouver, B.C., Canada, 2006.

[Kalayeh *et al.*, 2015] Mahdi M Kalayeh, Misrak Seifu, Wesna LaLanne, and Mubarak Shah. How to take a good selfie? In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 923–926, Brisbane, Australia, 2015.

[Malouf, 2002] Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7, Taipei, Taiwan, 2002.

[Nguyen *et al.*, 2012] Tam V Nguyen, Si Liu, Bingbing Ni, Jun Tan, Yong Rui, and Shuicheng Yan. Sense beauty via face, dressing, and/or voice. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 239–248, Nara,Japan, 2012.

[Ojala *et al.*, 2002] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[Pallett *et al.*, 2010] Pamela M Pallett, Stephen Link, and Kang Lee. New golden ratios for facial beauty. *Vision research*, 50(2):149–154, 2010.

[Redi *et al.*, 2015] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. The beauty of capturing faces: Rating the quality of digital portraits. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8, Ljubljana, Slovenia, 2015.

[Wu *et al.*, 2011] Ou Wu, Weiming Hu, and Jun Gao. Learning to predict the perceived visual quality of photos. In *2011 International Conference on Computer Vision*, pages 225–232, Barcelona, Spain, 2011.

[Xie *et al.*, 2015] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. Scut-fbp: A benchmark dataset for facial beauty perception. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 1821–1826, Hong Kong, Chian, 2015.