Expert Systems With Applications 142 (2020) 112990

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Toward graph-based semi-supervised face beauty prediction

Fadi Dornaika[a,b,*], Kunwei Wang[a,c], Ignacio Arganda-Carreras[a,b], Anne Elorza[a], Abdelmalik Moujahid[a]

[a] University of the Basque Country (UPV/EHU), Spain
[b] IKERBASQUE, Basque Foundation for Science, Spain
[c] Northwestern Polytechnic University, Xian, China

ARTICLE INFO

Article history:
Received 18 February 2019
Revised 25 September 2019
Accepted 27 September 2019
Available online 5 October 2019

Keywords:
Image-based face beauty analysis
Graph-based semi-supervised learning
Graph-based label propagation
Deep face features,

ABSTRACT

Assessing beauty using facial images analysis is an emerging computer vision problem. To the best of our knowledge, all existing methods for automatic facial beauty scoring rely on fully supervised schemes. In this paper, we introduce the use of semi-supervised learning schemes for solving the problem of face beauty scoring when the image descriptor is holistic and the score is given by a real number. The paper has two main contributions. Firstly, we introduce the use of graph-based semi-supervised learning for face beauty scoring. The proposed method is based on texture and utilizes continuous scores in a full range. Secondly, we adapt and kernelize an existing linear Flexible Manifold Embedding scheme (that works with discrete classes) to the case of real scores propagation. The resulting model can be used for transductive and inductive settings. The proposed semi-supervised schemes were evaluated on three recent public datasets for face beauty analysis: SCUT-FBP, M^2B, and SCUT-FBP5500. The obtained experimental results, as well as many comparisons with fully supervised methods, demonstrate that the non-linear semi-supervised scheme compares favorably with many supervised schemes. The proposed semi-supervised scoring framework paves the way to virtually all applications to adopt continuous scores instead of the usual discrete labels.

1. Introduction

It has been widely reported that the concept of facial beauty is universal, and may be defined throughout different races, eras and cultures (Chen, German, & Zaidel, 1997; Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Larrabee, 1997). Attempts at quantifying facial beauty have been addressed by investigators from different areas such as psychology, arts, plastic surgeons, and more recently image analysis (Fan et al., 2018; Gunes & Piccardi, 2006; Langlois & Roggman, 1990; Liu, Fan, Guo, Samal, & Ali, 2017).

Despite some progress, measuring facial beauty remains a challenging task. The existing approaches either lack some human or expert intervention, or require the computation of several manual measurements, or both. Therefore, automatic assessment of face beauty can be valuable for applications such as plastic surgery, computer assisted personalized search of partners (Whitehill & Movellan, 2008), animation, advertising, and computer games. Indeed, with the rapid development of pattern analysis and machine learning techniques, computer aided approaches for beauty analysis are emerging in recent years (Eisenthal, Dror, & Ruppin, 2006; Xie, Liang, Jin, Xu, & Li, 2015).

In general, the task of assessing face beauty by a machine can be carried out using three categories of schemes. The first category concerns classification whenever beauty is defined by levels or classes. The second category seeks face image ordering where the task is to order the face images by their attractiveness. The third category deals with automatic scoring of face beauty. Schemes belonging to this category are the most generic ones since the solution related to the two first categories is implicitly solved.

Early efforts to introduce automatic facial beauty system (Aarabi, Hughes, Mohajer, & Emami, 2001), aimed at classifying beauty in four different levels. Each image was described by eight ratios (e.g., the ratio of the horizontal distance between the eyes to the average vertical distance between the eyes and the mouth), which were supposed to capture the essence of beauty according to previous studies made by psychologists and biologists. This means there were only eight geometric features. The performance of the classifier strongly depended on the accuracy of localization of the eyes and the mouth. The dataset used consisted of 80 photos, each of them rated by 12 individuals among the four

* Corresponding author.
E-mail addresses: fadi.dornaika@ehu.es (F. Dornaika), wkwkee@mail.nwpu.edu.cn (K. Wang), ignacio.arganda@ehu.eus (I. Arganda-Carreras), aelorza026@ikasle.ehu.eus (A. Elorza), abdelmalik.moujahid@ehu.eus (A. Moujahid).

https://doi.org/10.1016/j.eswa.2019.112990
0957-4174/

classes, and the ground truth label was set to the median of all raters' ratings. The classifier was a variant of *k*-nearest neighbors.

Since then, many researches have been developed (Whitehill & Movellan, 2008; Yan, Duan, Deng, Zhu, & Wu, 2016; Zhang, Chen, & Xu, 2016). The first accurate facial attractiveness predictor (Kagian, Dror, Leyvand, Cohen-Or, & Ruppin, 2007; Kagian et al., 2008) used a high number of features: 98 image features: 90 geometric features and eight related to face symmetry, hair and skin color and skin smoothness. Their study, along with others like Eisenthal et al. (2006), reinforced the idea that beauty perception is something a machine can learn. However, together with the limited size of the dataset (around 90 images), this study still had another major flaw: the process of feature extraction was not automatic. In fact, the geometric features were based on landmarks. It was shown that an automatic engine could detect the eyes, the nose, the lips, the eyebrows, and the head contour; yet some of them have failed to be correctly identified and had to be adjusted manually.

One of the principal challenges of the research in this field is to build accurate facial representations, which can be either feature-based, holistic or hybrid. In the first category one can find geometric, color, texture or other local representations. Geometric representations are based on the use of facial landmark points and consider their positions, distances between them or ratios of these distances (Zhang, Zhao, & Chen, 2011). Some studies have also focused on the relationship between the golden ratio and beauty (Schmid, Marx, & Samal, 2008). On the other hand, holistic approaches use the global information of the face, instead of local features, e.g., using eigenfaces (Eisenthal et al., 2006).

In Gan, Wang, and Xu (2015), for instance, the authors extract face features by applying multi-scale and multi-patch K-means in order to extract multi-scale apparent features. They demonstrate the efficacy of their method on their own dataset that consists of 5000 labeled female images and 5000 labeled male images. They consider five levels of face attractiveness. In Nguyen et al. (2013), the authors propose the Dual-supervised Feature-Attribute-Task (DFAT) learning framework. First, for each modality, i.e., face, dress and voice, different kinds of features are extracted. Then the beauty estimation models of the single/multiple modalities are jointly learned. During the learning process, the semantic attributes are shared by different tasks. DFAT learns two types of regression models simultaneously by minimizing two types of prediction errors, one is feature-to-attribute error, and the other is attribute-to-attractiveness score error. Other approaches were based on a family of convolutional neural networks whose input was the raw image and the output is the beauty score (Gray, Yu, Xu, & Gong, 2010).

Deep learning based schemes have been also adopted for holistic approaches (Gan, Li, Zhai, & Liu, 2014; Wang, Shao, & Fu, 2014). Indeed, during the last decade, the studies in this area have partly focused on constructing larger and reliable benchmarks (Mu, 2013; Nguyen et al., 2013; Xie et al., 2015), where many works make use of various neural networks such as Gan et al. (2014), Nguyen et al. (2013), Xie et al. (2015) and Xu, Jin, Liang, Feng, and Xie (2015). For a recent review on face beauty analysis we refer the reader to Laurentini and Bottino (2014) and Liu, Fan, Samal, and Guo (2016).

Most of these works were based on supervised/unsupervised learning, however, little has been done to address automatic face beauty scoring using semi-supervised learning despite the difficulties to get labeled training points. In fact, semi-supervised learning has proven to be useful in the situation where relatively few labeled training samples are available, but a large number of unlabeled samples are given. Although, semi-supervised learning has been widely explored in both classification and clustering problems (Kamnitsas et al., 2018; Smieja, Myronov, & Tabor, 2018), to the best of our knowledge, the use of semi-supervised learning to solve the automatic assessment of face beauty has not been sufficiently addressed. The first attempts in that direction have some drawbacks. In Zhang, Zhang, Sun, and Chen (2017), for example, one may raise three major limitations. First, face features are pure geometric since they are represented by the 2D location of 68 facial points. Second, many training images are synthetically generated by deforming the shape of the face and warping the face image. Third, the beauty scores are related to only two classes: attractive faces having a score equal to +1 and unattractive faces having a score equal to -1. In Gan et al. (2014), the authors propose a learning scheme in which a convolutional deep belief network (CDBN) is learned in two phases. In the first phase, unlabeled natural images are used for pre-training. In the second phase, the deep net is trained using a supervised scheme on the labeled face images. At test stage, high level features are extracted from the net for each face image. Assessing face beauty is then performed using a classic supervised scheme adopting regression techniques.

In this work, we report two main novelties. **(1)** Introducing the use of graph-based semi-supervised learning for face beauty scoring using real scores and non-geometric face features. The proposed method is based on texture and utilizes continuous scores in a full range. **(2)** Adapting and kernelizing an existing linear Flexible Manifold Embedding scheme (that works with discrete classes) to the case of real scores propagation. The importance of the second contribution is twofold. Firstly, the proposed non-linear prediction model can better tackle the beauty prediction problem by adopting this nonlinear model. Secondly, since the proposed semi-supervised model is inductive, it is applicable and scalable in real-world applications especially when the dataset sizes are large.

The paper is organized as follows. Section 2 presents the motivation of the work. Section 3 describes two adapted graph-based semi-supervised learning methods for beauty score propagation. Section 4 describes the proposed non-linear Flexible Manifold Embedding for score propagation. Section 5 describes the experimental setup used for evaluating the performance of the proposed scheme. Section 6 provides the results on three public datasets: SCUT-FBP, M^2B and SCUT-FBP5500. Section 7 concludes the paper.

2. Background

2.1. Motivation and contribution

For face beauty analysis and its applications, pattern analysis and machine learning based techniques, such as neural networks, support vector machines, K-nearest neighbor, linear regression, and non-linear regression, have proven to be promising supervised approaches (Bottino & Laurentini, 2010; Mu, 2013; Sutic, Rreskovic, Huic, & Jukic, 2010). However, little work (if any) has been done for the semi-supervised case.

In fact, one of the main problems in face beauty analysis is the scarcity of labeled data. It is well known that obtaining reliable and large databases requires a great deal of human labor. In order to obtain meaningful labels, many raters are needed for each face photo and, usually, the ground truth label of an image is considered to be the average score of all the ratings given by different raters. As a concrete example showing the difference between traditional classification problems and face beauty assessment, let us consider the public ImageNet2012 classification dataset (Russakovsky et al., 2018) that consists of 1,000 classes. It contains 1.28 million training images and is used to train supervised models. In comparison, in 2017, the largest public dataset for Face Beauty Prediction had 1,240 labeled face images, given by the Multi Modality-Beauty (M^2B) dataset (Nguyen et al., 2013). It is obvious that semi-supervised learning techniques appear to be particularly appropriate since they can exploit unlabeled photos.

Moreover, it is expected that semi-supervised techniques can enrich the model by exploiting labeled and unlabeled images.

Unlike supervised learning, which just makes use of labeled data, semi-supervised learning utilizes the information underlying the unlabeled data as well (Dornaika & Traboulsi, 2017; Traboulsi & Dornaika, 2018). The generic semi-supervised learning methods using graph-based label propagation attracted much attention in the last decade. All of them impose that samples with high similarity should share similar labels. They differ by the regularization term as well as by the loss function used for fitting label information associated with the labeled samples. All of these methods use the graph similarity matrix and the initial labels of some samples. Some recent label propagation algorithms (they can also be called classifiers (Sousa, Rezende, & Batista, 2013)) are: Gaussian Fields and Harmonic Functions **(GFHF)** (Zhu, Ghahramani, Lafferty, & others, 2003), Local and Global Consistency **(LGC)** (Zhou, Bousquet, Lal, Weston, & Schölkopf, 2004), Laplacian Regularized Least Square **(LapRLS)** (Belkin, Niyogi, & Sindhwani, 2006), Robust Multi-class Graph Transduction **(RMGT)** (Liu & Chang, 2009), Flexible Manifold Embedding (**FME**) (Nie, Xu, Tsang, & Zhang, 2010). These techniques can be either transductive (defined for training samples only) or inductive (defined for both training and unseen samples). The method proposed in Nie, Tian, Wang, and Li (2019), learns a unified graph via a structural regularization term. Instead of weight regularization which is adopted by previous works, the work presented in Nie et al. (2019) learns a unified graph and weights from a priori individual graphs.

Thinking about beauty, it seems quite reasonable to assume that when two faces resemble each other they should have similar attractiveness scores. This abstract idea can be materialized by constructing a weighted graph, in which nodes are images (or their descriptors) and the weights between each pair of nodes represent their similarities. Therefore, in our case, we exploit manifold structure of face images (both labeled and unlabeled) via graphs. In this assumption, similar images should share similar beauty scores.

In this paper, we introduce the semi-supervised paradigm to face beauty prediction field. We explore some manifold based semi-supervised algorithms to the specific problem of automatic facial beauty assessment. Moreover, we propose a non-linear Flexible Manifold Embedding for carrying out the score propagation. This proposed method can achieve state-of-the-art results.

2.2. Notations and preliminaries

In the sequel, capital bold letters denote matrices and bold letters denote vectors. Assume that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ are l labeled face images (or their descriptors) and that $\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \ldots, \mathbf{x}_N$ are the u unlabeled face images. Here, the vector $\mathbf{x}_i$ refers to the ith face image.

The data matrix $\mathbf{X}$ is defined by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$. The total number of training images is $N = l + u$. The semi-supervised algorithms we are using were originally developed for classification tasks, where the ground truth labels $\mathbf{Y}$ and the predicted labels $\mathbf{F}$ are matrices in $\mathbb{R}^{N \times C}$, where $Y_{ij} = 1$ if sample i belongs to class j and $Y_{ij} = 0$ otherwise. C denotes the number of classes.

Since our problem is essentially a regression problem, the labels are real numbers that can be represented as a column vector $\mathbf{y} \in \mathbb{R}^N$. The first l rows of $\mathbf{y}$ will contain the scores of the l labeled images, while the last u rows will generally be 0, since they correspond to the u unlabeled images. In addition to the initial label (or score) vector $\mathbf{y}$, we consider the unknown label vector $\mathbf{f} \in \mathbb{R}^N$ that should be estimated.

As already stated, the way of exploiting the information contained in unlabeled data is to consider a similarity graph that encodes the pairwise similarity between images. To this end, we have to introduce a similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ (which will be symmetric, so our graph has to be undirected). Each element S_{ij} of $\mathbf{S}$ is the similarity between samples i and j (i.e., face i and face j). This graph is assumed to capture much information about the data manifold. In our work, without loss of generality, the affinity matrix $\mathbf{S}$ is set to the KNN graph similarity matrix as it offers a simple and very efficient method for graph construction. It proceeds as follows. First, the adjacency matrix is constructed (the edges are set). Second, the weights of the edges are estimated.

For adjacency matrix construction, K-Nearest Neighbor can be used in order to find the neighbors of a datum. There is a function that defines the distance (similarity) of one input with respect to the others.

In the second phase, a weight should be assigned to each constructed edge. In general, this weight should quantify the similarity between two connected nodes. Let $sim(\mathbf{x}_i, \mathbf{x}_j)$ be the similarity score between neighbors $\mathbf{x}_i$ and $\mathbf{x}_j$, then the elements of the graph weight matrix $\mathbf{S}$ are given by Eq. (1).

$$S_{ij} = \begin{cases} sim(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

There are several choices for $sim(\mathbf{x}_i, \mathbf{x}_j)$. For instance, in Belkin and Niyogi (2003) the authors use the heat kernel $sim(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}}$ where t can be set to the average of squared distances in the training set. We adopt the above similarity and set the neighborhood size of the KNN graph to 10 as in many studies (Dornaika & El Traboulsi, 2016)). We emphasize that more sophisticated graph construction methods can be used in order to estimate the similarity matrix $\mathbf{S}$ (e.g., Cheng, Yang, Yan, Fu, & Huang, 2010; Dornaika & Bosaghzadeh, 2015; Dornaika, Dhabi, Bosaghzadeh, & Ruichek, 2016; Dornaika, Traboulsi, & Assoum, 2013; He, Zheng, Hu, & Kong, 2011; Nie, Wang, Jordan, & Huang, 2016).

The Laplacian matrix of $\mathbf{S}$ is given by $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is the diagonal matrix whose elements are the row sums of $\mathbf{S}$. The normalized Laplacian matrix is given by $\hat{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is the identity matrix of size N. Finally, $\mathbf{1}, \mathbf{0} \in \mathbb{R}^N$ denote vectors with all elements as 1 and 0 respectively. The norm $\|\cdot\|$ denotes the Euclidean norm.

2.3. Problem statement

The input data are given by a set of face images or their descriptors $\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \ldots, \mathbf{x}_N$. l and $u = N - l$ represent the numbers of labeled and unlabeled face images, respectively.

The labels are given by real scores $y_1, y_2, \ldots, y_l$. The goal is to infer the scores of the unlabeled face images. For illustration, Fig. 1 shows a toy example that demonstrates the principle of graph-based score propagation. In this example, we have seven face images, from which only 4 images are labeled with face beauty score. The remaining images are unlabeled. The objective is to recover the face beauty score of all unlabeled images by performing score propagation over the graph. The similarity matrix of the graph, $\mathbf{S}$, is constructed using all training images. Unless stated otherwise, the paper targets learning from single-valued scores. Thus, learning from discrete classes or from label distribution is beyond the scope of the paper.

3. Graph-based score propagation schemes

In this section, we will describe some existing label propagation methods that were developed for discrete classification. We also show their adaptation for the score propagation, where the score is a continuous variable. All existing graph-based label propagation scheme use either a non normalized graph or a normalized graph. We emphasize that the three semi-supervised schemes that are

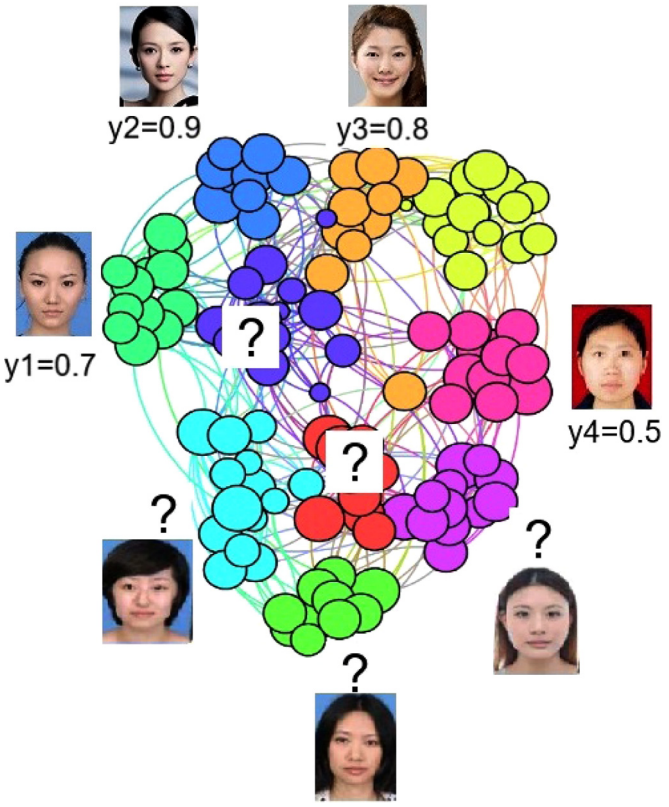

Fig. 1. Principle of graph-based beauty score propagation.

presented and developed in Sections 3 and 4 use a normalized graph. In other words, the objective functional of each method uses a normalized Laplacian matrix.

3.1. Local and Global Consistency

The Local and Global Consistency (LGC) method was introduced in Zhou et al. (2004). It aims at predicting the discrete labels of all labeled and unlabeled instances, $\mathbf{F}$, by minimizing the following function:

$$q(\mathbf{F}) = \sum_{i,j=1}^{N} S_{ij} \left\| \frac{\mathbf{f}_i}{\sqrt{D_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{D_{jj}}} \right\|^2 + \mu \sum_{i=1}^{N} \|\mathbf{f}_i - \mathbf{y}_i\|^2, \tag{2}$$

where $\mathbf{f}_i$ is the ith row of $\mathbf{F}$ and D_{ii} is the sum of the ith row of $\mathbf{S}$ or, in other words, the sum of the similarities of sample i with all the other images. The first term is the smoothness constraint. The second term is the fitting constraint and μ is the parameter which controls the trade-off between them. The optimal solution to this problem can be found analytically by vanishing the first derivatives w.r.t. the unknown. It is given by $\mathbf{F} = (\mathbf{I} + \hat{\mathbf{L}}/\mu)^{-1}\mathbf{Y}$, where $\mathbf{I}$ is the identity matrix of N dimensions. Once the matrix $\mathbf{F}$ is estimated, the predicted class of an instance i will be the maximum index j of the ith row of $\mathbf{F}$.

Our revisited LGC should predict the scores of all images, namely the vector $\mathbf{f}$. It minimizes the following:

$$q(\mathbf{f}) = \sum_{i,j=1}^{N} \hat{S}_{ij}(f_i - f_j)^2 + \mu \|\mathbf{f} - \mathbf{y}\|^2, \tag{3}$$

where the matrix $\hat{\mathbf{S}}$ is given by $\hat{\mathbf{S}} = \mathbf{D}^{-1}\mathbf{S}$. As it can be seen, the criterion is similar to the one presented in (2). However, there are two significant differences. First, the distance in scores between any pair of nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ is not any more depending on the degree of these two nodes. Indeed, empirically, we found that keeping the degrees in the pairwise distance lead to worse results. This is due to the fact that our problem is score estimation that should fit some fixed values provided by the labeled images. Therefore, whenever the degrees are different the vectors $\frac{\mathbf{f}_i}{\sqrt{D_{ii}}}$ and $\frac{\mathbf{f}_j}{\sqrt{D_{jj}}}$ will not be equally significant in the distance depicted in the first term of Eq. (2). By recalling the definition of D_{ii} (it is the sum of the ith row in $\mathbf{S}$), a simple analysis of this term $\frac{\mathbf{f}_i}{\sqrt{D_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{D_{jj}}}$ will show that the most influential samples are the ones being the least similar to the rest.

Second, unlike discrete label propagation where it is safe to set the label distribution $\mathbf{Y}_{ij}$ associated with the unlabeled images to zero vectors, in our case these values (i.e., $y_i, i = 1, \ldots, u$) are set to the average scores that can be easily known from the labeled images. The solution for $\mathbf{f}$ is again similar to the one provided by the LGC discrete label propagation in which the Laplacian matrix is now associated with the graph $\frac{\hat{\mathbf{S}}+\hat{\mathbf{S}}^T}{2}$:

$$\mathbf{f} = (\mathbf{I} + \hat{\mathbf{L}}/\mu)^{-1}\mathbf{y}$$

3.2. Flexible Manifold Embedding

Similarly to the LGC method, the Flexible Manifold Embedding (FME) method (Nie et al., 2010) estimates the labels $\mathbf{F}$ by minimizing the following cost function:

$$\begin{aligned} g(\mathbf{F}, \mathbf{W}, \mathbf{b}) = {} & \mathrm{tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \beta\, \mathrm{tr}[(\mathbf{F} - \mathbf{Y})^T\mathbf{U}(\mathbf{F} - \mathbf{Y})] \\ & + \mu(||\mathbf{W}||^2 + \gamma||\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||^2), \end{aligned} \tag{4}$$

where $\mathbf{U}$ is an indicator matrix, that is a diagonal matrix, with its first l diagonal elements, corresponding to labeled instances, equal to 1, while the last u diagonal elements, corresponding to unlabeled instances, are equal to 0. This assumes that the l labeled images are the first l samples in the data matrix $\mathbf{X}$ and in the similarity matrix $\mathbf{S}$. $\mathbf{W}$ and $\mathbf{b}$ denote the unknown linear regressor which maps the original samples to the label space.

As with the LGC method, a closed-form solution can be found by setting the derivatives of g with respect to $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{F}$ as 0 (Nie et al., 2010). The solution is given by:

$$\mathbf{b} = \frac{1}{l+u}\left(\mathbf{F}^T\mathbf{1} - \mathbf{W}^T\mathbf{X}\mathbf{1}\right) \tag{5}$$

$$\mathbf{W} = \gamma(\gamma\mathbf{X}\mathbf{H}_c\mathbf{X}^T + \mathbf{I})^{-1}\mathbf{X}\mathbf{H}_c\mathbf{F} \tag{6}$$

$$\mathbf{F} = \beta(\beta\mathbf{U} + \mathbf{L} + \mu\gamma\mathbf{H}_c - \mu\gamma^2\mathbf{Q})^{-1}\mathbf{U}\mathbf{Y}, \tag{7}$$

where $\mathbf{Q} = \mathbf{X}_c^T\mathbf{X}_c(\gamma\mathbf{X}_c^T\mathbf{X}_c + \mathbf{I})^{-1}$, $\mathbf{X}_c = \mathbf{X}\mathbf{H}_c$ and $\mathbf{H}_c = \mathbf{I} - (1/(l+u))\mathbf{1}\mathbf{1}^T$.

Flexible Manifold Embedding and regression: As with the LGC method, this algorithm was originally designed for classification tasks. Nonetheless, it can easily be adapted to work as a regressor. If $\mathbf{f}$ and $\mathbf{y}$ are real-valued vectors representing the predicted labels and the ground-truth scores, the cost function becomes

$$\begin{aligned} g(\mathbf{f}, \mathbf{w}, b) = {} & \mathbf{f}^T\mathbf{L}\mathbf{f} + \beta(\mathbf{f} - \mathbf{y})^T\mathbf{U}(\mathbf{f} - \mathbf{y}) \\ & + \mu\,(||\mathbf{w}||^2 + \gamma||\mathbf{X}^T\mathbf{w} + b\,\mathbf{1} - \mathbf{f}||^2). \end{aligned} \tag{8}$$

The first term controls the label smoothness, the second one the label fitness and the last term fits a linear regression between features and labels, where $||\mathbf{w}||^2$ is a regularization term controlling the complexity of the model (thus, avoiding over-fitting). β, μ and γ are the parameters controlling the trade-off between all the terms. The solution is given by:

$$b = \frac{1}{N}\left(\mathbf{f}^T\mathbf{1} - \mathbf{w}^T\mathbf{X}\mathbf{1}\right) \tag{9}$$

$$\mathbf{w} = \gamma(\gamma\mathbf{X}\mathbf{H}_c\mathbf{X}^T + \mathbf{I})^{-1}\mathbf{X}\mathbf{H}_c\mathbf{f} \tag{10}$$

$$\mathbf{f} = \beta(\beta\mathbf{U} + \mathbf{L} + \mu\gamma\mathbf{H}_c - \mu\gamma^2\mathbf{Q})^{-1}\mathbf{U}\mathbf{y}. \quad (11)$$

FME and unseen data: One advantage of the FME method, which distinguishes it from many proposed label propagation methods, e.g., from the LGC method, is its capacity of dealing with unseen data. Apart from predicting the labels of the unlabeled data in vector $\mathbf{f}$, it can predict unseen data using the regression model that is already learned in (8). Given the features of the unseen data $\mathbf{X}_{uns}$, its scores would be:

$$\mathbf{f}_{uns} = \mathbf{X}_{uns}^T\mathbf{w} + b\mathbf{1}.$$

4. Proposed approach: Non-linear Flexible Manifold Embedding

The FME model works directly on the data samples. In many recent machine learning works, it was shown that working with a non-linear representation of the data can improve the final performance of the learner. In our work, we propose to use the column generation trick in order to get the non-linear representation of the original data matrix $\mathbf{X}$.

Column generation replaces each sample $\mathbf{x}_i$ by a vector of similarities of that sample with the samples contained in a fixed set of samples (Klare & Jain, 2013). Very often, the latter set is given by the training samples or a subset of them (Zhang, Liu, Shen, Shen, & Shao, 2018).

In our work, we use all training samples as reference samples. The data matrix $\mathbf{X}$ is thus replaced by the matrix $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N]$, where each vector $\mathbf{g}_i$ is formed by the similarities, i.e. $\mathbf{g}_i = sim(\mathbf{x}_i, \mathbf{x}_1), \ldots, sim(\mathbf{x}_i, \mathbf{x}_N) \in \mathbb{R}^{N\times N}$. In our case, we use the Gaussian similarity. This means that

$$G_{ij} = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2^{t_0}\sigma^2}},$$

where σ^2 is a measure of the variability of the data. Concretely, it is set to the mean of the squares of the distances between all pairs of samples. t_0 is a kernel parameter that can control the similarity function. Therefore, the Non-linear Flexible Manifold Embedding (NFME) can be formulated as follows:

$$\begin{aligned} g(\mathbf{f}, \mathbf{w}, b) &= \mathbf{f}^T\mathbf{L}\mathbf{f} + \beta(\mathbf{f} - \mathbf{y})^T\mathbf{U}(\mathbf{f} - \mathbf{y}) \\ &\quad + \mu\,(||\mathbf{w}||^2 + \gamma\,||\mathbf{G}^T\mathbf{w} + b\mathbf{1} - \mathbf{f}||^2). \end{aligned} \quad (12)$$

The NFME method can cope with non-linear data when a linear regression may have a poor performance in the FME. A closed-form solution can be found again by setting the derivatives of g with respect to $\mathbf{f}$, $\mathbf{w}$, and b as 0. The solution is given by:

$$b = \frac{1}{N}\left(\mathbf{f}^T\mathbf{1} - \mathbf{w}^T\mathbf{G}\,\mathbf{1}\right) \quad (13)$$

$$\mathbf{w} = \gamma\,(\gamma\,\mathbf{G}\,\mathbf{H}_c\mathbf{G}^T + \mathbf{I})^{-1}\mathbf{G}\,\mathbf{H}_c\mathbf{f} \quad (14)$$

$$\mathbf{f} = \beta\,(\beta\,\mathbf{U} + \mathbf{L} + \mu\,\gamma\,\mathbf{H}_c - \mu\,\gamma^2\mathbf{Q})^{-1}\mathbf{U}\mathbf{y} \quad (15)$$

where $\mathbf{Q} = \mathbf{G}_c^T\mathbf{G}_c(\gamma\,\mathbf{G}_c^T\mathbf{G}_c + \mathbf{I})^{-1}$, $\mathbf{G}_c = \mathbf{G}\,\mathbf{H}_c$, and $\mathbf{H}_c = \mathbf{I} - (1/(l + u))\mathbf{1}\mathbf{1}^T$.

NFME and unseen data: Similarly to the FME method, in the NFME method one can easily handle unseen data using the regression term in (12). To do so, given a set of unseen samples $\mathbf{x}_i^{uns}|_{i=1}^m$, one has to build the similarity matrix of the unseen samples $\mathbf{G}_{uns} \in \mathbb{R}^{m\times N}$, where the element (i, j) is the similarity function of the unseen sample i, $\mathbf{x}_i^{uns}$, and the training sample j, $\mathbf{x}_j$. Then, the predicted scores, $\mathbf{f}_{uns} \in \mathbb{R}^m$, would be

$$\mathbf{f}_{uns} = \mathbf{G}_{uns}^T\mathbf{w} + b\mathbf{1}. \quad (16)$$

5. Experimental setup

5.1. Datasets

Three datasets are used in this work: the SCUT-FBP dataset (Xie et al., 2015), the Multi Modality-Beauty (M^2B) dataset (Nguyen et al., 2013), and the SCUT-FBP5500 dataset (Liang, Lin, Jin, Xie, & Li, 2018). The first was specifically designed for automatic facial beauty perception and contains high resolution front-on face portraits of Asian females. Moreover, the second was developed to evaluate beauty via a face, dressing and/or voice on both Eastern and Western females and each instance in the dataset contains information about the three modalities. However, we are only focusing on the facial images, which unlike the ones in the SCUT-FBP dataset, show very different poses and expressions. This complicates, in consequence, the beauty assessment, which could be found difficult even by a human rater.

SCUT-FBP dataset: The SCUT-FBP dataset contains 500 high resolution front-on face portraits of Asian females with neutral expressions, simple background and minimal occlusion, as can be seen in Fig. 2. These characteristics prevent from taking into account irrelevant factors in the beauty classification task. The beauty rankings (scores) lie in the interval (1, 5) and are the result of averaging various ratings. The ratings were collected among 75 individuals using a web-based tool with an average number of 70 raters per image. The scores approximately follow a normal distribution (Fig. 3) with a small peak around 4.5.

Raters' consistency and self-consistency are checked in different ways by the authors of the paper. For instance, low standard deviations in the ratings of each image indicate rater's agreement in the perception of beauty.

M^2B dataset: The Multi-Modality Beauty dataset has been developed to study beauty perception in three different modalities, in dressing, in the face and in the voice, as well as the global beauty perceived when any of these three aspects are combined. Therefore, the dataset contains one face photo, one full body photo and one voice snippet of 1240 females belonging to two ethnic groups: westerners and easterners (620 individuals in each group). In addition, each of the females of the dataset is rated, in the different modalities and their combinations, with various scores in the interval [1, 10].

The ratings were collected among 40 participants, which were split into two groups depending on their ethnicity, so that each of the participants rated females of their own ethnic group. The web tool used for this purpose can be seen in Fig. 4. The ratings

Fig. 2. Examples of face portraits of SCUT-FBP dataset from Xie et al. (2015).

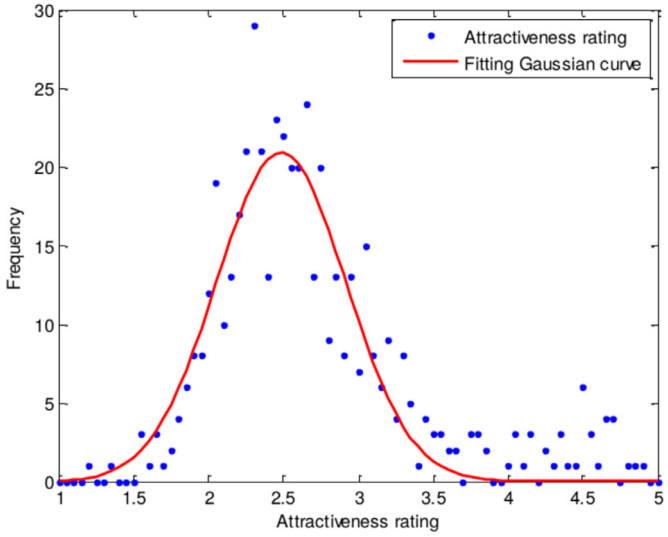

Fig. 3. Histogram of the rating distribution from Xie et al. (2015).

Fig. 4. User interface of the attractiveness ranking tool from Nguyen et al. (2013).

were obtained using k-wise comparison, which means that the raters are asked to sort k females according to their beauty, and then these k-wise ratings were converted into global ratings in the interval [1, 10] by solving an optimization problem to preserve as many pairwise preferences as possible. The drawback of this method of collecting the labels is that, unlike the SCUT-FBP dataset, where we had the ratings of various raters per image, here we have a unique rating. Thus, we cannot really measure the uncertainty of each of the labels, even if it seems to be important, since beauty is not an absolute concept.

SCUT-FBP5500 dataset: In order to overcome the small size of the SCUT-FBP database, the authors of Liang et al. (2018) introduced the SCUT-FBP5500 dataset. This dataset contains 5,500 frontal, unoccluded faces aged from 15 to 60 with neutral expression. It can be divided into four subsets with different races and gender, including 2,000 Asian females, 2,000 Asian males, 750 Caucasian females and 750 Caucasian males. Most of the images of the SCUT-FBP5500 dataset were collected from the Internet, where some portions of Asian faces were from the DataTang (DataTang, 2019) and some Caucasian faces were from the 10k US Adult database (Bainbridge, Isola, & Oliva, 2013). The SCUT-FBP5500 dataset allows different computational models with different prediction paradigms, such as appearance-based/shape-based facial beauty classification/regression model for male/female of Asian/Caucasian. Some samples are shown in Fig. 5.

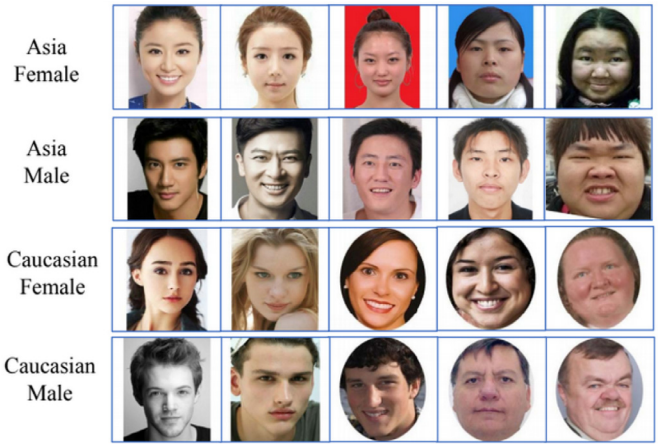

Fig. 5. Examples of face portraits of SCUT-FBP55500 dataset from Liang et al. (2018).

5.2. Face features

Extracting face features from images can be done in many ways. For instance, one can use hand-crafted features such as Local Binary Patterns, Gabor features, Histograms of Oriented Gradients, etc. However, recent studies have proved that extracting features using a pre-trained Convolutional Neural Network (CNN) can provide more than satisfactory results in computer vision (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014), often improving methods using other types of features, even if they are used in a task which is not exactly the one the network was trained to accomplish. Therefore, we adopt this transfer learning approach for our study and use the VGG-face network (Parkhi, Vedaldi, Zisserman et al., 2015) as fixed feature extractor. The VGG-face network was trained for face identification. This trained network is available for Matlab.[1] Fig. 6 shows its architecture.

The face features are extracted from layer 7 (fc7), right before the last fully-connected layer (fc8 in the diagram). Moreover, some experiments used the output of the layer (fc6). Before propagating an image through VGG-face, it must be resized and the average image associated with the network has to be subtracted. After extracting the features using the CNN, they are organized in a matrix, where each column represents the 4096 features of an image. The data matrix is normalized using L_2 normalization (each column vector is divided by its Euclidean norm). The dimensionality of all samples is then reduced from 4096 to 200 features using Principal Component Analysis (PCA).

5.3. Evaluation protocol

As mentioned above, three datasets are used in our study: SCUT-FBP, M^2B, and SCUT-FBP5500. Labels (scores) in the SCUT-FBP and SCUT-FBP5500 datasets lie in the interval [1, 5], whereas in the M^2B dataset they lie in [1, 10]. Labels in the SCUT-FBP and SCUT-FBP5500 datasets are divided by 5 so they lie in [0.2, 1]. Labels in the M^2B dataset are divided by 10 so they lie in [0.1, 1].

The providers of the M^2B and SCUT-FBP5500 datasets have also provided the evaluation protocol. For the M^2B dataset, the providers used the 2-fold cross validation scheme. For the SCUT-FBP5500 dataset, the authors adopted the five-fold cross validation scheme (Liang et al., 2018). We note that the five folds associated with the SCUT-FBP5500 dataset were already prepared by the

[1] The network can be downloaded from http://www.vlfeat.org/matconvnet/pretrained/.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Fig. 6. Architecture of VGG-face from Parkhi et al. (2015).

Table 1
Transformation of the labels from real numbers to discrete classes for the purpose of data stratification. The objective of this stratification is to roughly preserve score distribution in the training and testing parts.

Class	Interval in SCUT-FBP	Interval in M^{2}B
1	[1,2)	[1,3)
2	[2, 3)	[3, 5)
3	[3, 4)	[5, 7)
4	[4, 4.5)	[7, 8)
5	[4.5, 5)	[8, 10)

providers of this dataset. For the SCUT-FBP, there is no specific evaluation protocol.

In the sequel, we adopt the same protocols for the M^{2}B and SCUT-FBP5500.

For the SCUT-FBP we adopt random splits with different labeled data sizes, which can be particularly interesting considering that we are working in a semi-supervised setting. In most of the experiments on the SCUT-FBP dataset, the labeled data constitute the 50%, the 70% or the 90% of all the data. In order to have a meaningful evaluation of performances of the models, for a given training percentage (50%, 70% or 90%), each model is evaluated 10 times with 10 different random splits of the data. The final performance is set to the average performance over the 10 splits.

We also conducted experiments in order to study the effect of testing in unseen data. In these cases, the dataset has three parts: labeled images, unlabeled images, and unseen images.

Since the scores are continuous numbers, five virtual classes are created to carry out the stratification easily, as shown in Table 1. The objective of this stratification is to preserve the score distribution in the training part and in the testing part. It is worth noting that these classes are pure virtual classes. Indeed, other number of classes and intervals can also be adopted without significantly altering the final model which is a semi-supervised regressor. Each model is evaluated using the following performance metrics: the mean absolute error (MAE), the root mean square error (RMSE), the Pearson correlation coefficient (PC) and the ϵ-error. Let $y_1, y_2, \ldots, y_n$ be the ground truth labels (scores) and $f_1, f_2, \ldots, f_n$ the estimated labels (scores). The four metrics are defined as follows:

Mean absolute error: The mean absolute error (MAE) of the prediction is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f_i|. \tag{17}$$

Root mean square error: The root mean square error of the prediction is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}. \tag{18}$$

Pearson correlation coefficient: The Pearson's correlation (PC) coefficient measures the linear correlation between the ground truth and the estimated labels and is given by:

$$PC = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(f_i - \bar{f})^2}}, \tag{19}$$

where $\bar{f}$ and $\bar{y}$ are the means of the estimated labels and the ground truth labels respectively.

The PC is a number in the interval $[-1, 1]$. It measures to what extent the points (y_i, f_i), with $i = 1, \cdots, n$, fit in their regression line. A PC around 1 or -1 means that there is a high linear correlation. On the other hand, a PC around 0 means that there exists no linear correlation.

In our particular case, a PC as close as possible to 1 is desirable, since a perfect prediction would be one in which $f_i = y_i$, in other words, the pairs (y_i, f_i) lie on the line $y = x$. ***ϵ-error*** It is possible to compute the ϵ-error whenever the ground-truth data contain both the scores and their uncertainty, that is given by the standard deviation. The ϵ-error of the prediction is given by:

$$\epsilon\text{-error} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - e^{-\frac{(y_i - f_i)^2}{2\sigma_i^2}}\right), \tag{20}$$

where σ_i is the standard deviation of the scores of all the raters of image i. This error lies in the interval [0, 1]. Unlike the MAE and the RMSE, the ϵ- error takes into account the uncertainty of the rate given by the standard deviation. In this formula, the square difference between the real score and the estimated score $(y_i - f_i)^2$ is divided by σ_i^2. The ϵ-error function is based on $1 - Gaussian(x)$ where $x = |y_i - f_i|$. This function gives zero for a zero error. However, for other non-zero increasing values of x, the function grows less rapidly than the absolute difference and the square functions that are respectively adopted by the MAE and the RMSE. Moreover, the less growth is due to the division by the standard deviation. In brief, this function can down-weigh the contribution of large errors related to rates having large uncertainties.

Several of the algorithms used in this study have a number of parameters to be fixed. In all of these cases a grid search is conducted, that is, given a finite set of possible values for each parameter, all of the combinations are tried and the one giving the smallest prediction error is selected. All the Matlab code used in

Table 2
Average performances obtained with three graph-based score propagation schemes: LGC, FME, and NFME. The dataset used is SCUT-FBP.

Data partition	Method	MAE ↓	RMSE ↓	PC % ↑	ϵ-error ↓
50% labeled– 50% unlabeled (FC7)	LGC	0.0928	0.1217	59.52	0.2375
	FME	0.0626	0.0814	81.62	0.1421
	NFME	**0.0603**	**0.0784**	**82.36**	**0.1326**
70% labeled – 30% unlabeled (FC7)	LGC	0.0861	0.1132	69.05	0.2169
	FME	0.0563	0.0746	83.96	0.1219
	NFME	**0.0551**	**0.0727**	**84.35**	**0.1133**

these experiments is replicable downloading it from the following link: https://github.com/CVPD/aesthetic-analysis.

6. Experimental results

6.1. SCUT-FBP dataset

6.1.1. Graph-based score propagation

In all the experiments related to LGC, the MSE has been optimized adjusting the parameter μ for values ranging in $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$.

Regarding the FME and NFME methods, which are dependent on the parameters β, μ and γ, we optimized the PC for values of these parameters in the range $\{10^{-9}, 10^{-6}, 10^{-3}, 1, 10^3, 10^6, 10^9\}$ in order to optimize the PC.The parameter t_0 associated with the NFME method is selected in {1/8, 1/4, 1/2, 1, 2, 4, 8}.

Table 2 shows the results of applying the three different graph-based label propagation on the SCUT-FBP dataset with different training/test sizes. The best performances are shown in bold. Ideally, MAE, RMSE, and ϵ-error should be zero, and PC should be one. As it can be seen, the NFME method gave the best performance indicating that the non-linear regression is preferred over linear regression. In this table, we also depict the PC obtained with a supervised technique using a Deep CNN (Xie et al., 2015). The authors used 80% of images for training and 20% for testing and averaged the performance on five random trials. As it can be seen, the PC obtained with our NFME (0.847) is higher than the one obtained with the deep CNN (0.820). The same table depicted the performances obtained with the fc6 layer output (see lower part of the table). As it can be seen for the SCUT-FBP dataset, the use of the fc6 layer can improve the performance.

6.1.2. NFME: An experiment on unseen data

Any learner should be able to handle unseen images (Raducanu & Dornaika, 2014). The NFME method is inductive in the sense it can handle unseen images. In the following experiment, we attempt to address the unseen data case that is handled by the NFME scheme. We consider the following partition of the data (sampled with stratification): 200 images will form the labeled data, 200 images will be the unlabeled data and the remaining 100 images will be the unseen test data. The model of NFME is learned on the 400 labeled and unlabeled instances (the 80% of the data) and it will be tested on the remaining 100 images (20% of the data), as explained in Section 4 Eq. (16), using the linear regression on the kernel matrix associated to unseen samples.

Table 3 shows the results of the learned NFME model when applied to the unseen test images. In this case, only one data split has been considered. The obtained PC is now 0.79 which is lower than the one obtained with the 50% labeled– 50% unlabeled case (0.82). This is not surprising given the fact that the size of the training data is now smaller (200 labeled images and 200 unlabeled images) than the NFME model depicted in the first row of Table 2.

Table 3
Performance of NFME on unseen faces.

	MAE ↓	RMSE ↓	PC % ↑	ϵ-error ↓
NFME	0.075	0.0926	79.88	0.1714

Table 4
Summary of the performances of the supervised and semi-supervised methods on SCUT-FBP with a 90%–10% data partition.

Method	MAE ↓	RMSE ↓	PC % ↑	ϵ-error ↓
1-NN	0.0891	0.1139	63.77	0.2208
Ridge Regression	0.0645	0.0827	77.72	0.1429
Gaussian ϵ-SVR	0.0561	0.0711	84.34	0.1151
LGC	0.0830	0.1076	73.90	0.2090
FME	0.0567	0.0723	84.32	0.1196
NFME	**0.0554**	**0.0704**	**84.64**	**0.1119**

6.1.3. Comparisons with supervised algorithms

Due to the lack of semi-supervised approaches for face beauty scoring in the literature, we will compare the performances with that of the supervised techniques. Keeping in mind that supervised schemes have their own context and testing protocols, our comparison is intended to see to which extent the semi-supervised schemes can be useful and successful.

In this section, we will report the performances of three supervised methods, namely, **1-Nearest Neighbor** (1-NN), **Ridge Regression** (RR) and **ϵ-insensitive Support Vector Regression** (ϵ-insensitive SVR), in order to compare them with the semi-supervised schemes. For all compared methods, the percentage of labeled images is kept to 90% and the percentage of unlabeled images is kept to 10%. For the supervised methods, the labeled images are used for training and the unlabeled images are used for testing. The Nearest Neighbor classifier sets the score of the test image to that of its nearest neighbor in the training set. Three distances over the image features were used: Euclidean, Manhattan, and cosine. Empirical results show that the NN classifier provided the best performance when the cosine similarity is used. With regard to the Ridge Regression, the objective is to compute a linear mapping between raw image features (the deep features in PCA subspace) and the scores using a regularized least square problem. The balance parameter of ridge regression has been chosen among {0.0001, 0.001, 0.01, 0.1, 1, 10, 50, 100, 250, 500, 1000, 5000, 10000}. The ϵ-insensitive SVR used a Gaussian kernel. In this case, two parameters have to be adjusted, C and ϵ, which have been selected in the sets {0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 10, 20, 30} and {0, 0.001, 0.0025, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.15}, respectively.

Table 4 summarizes the best results obtained by each method with a fixed data partition of 90%/10%.

6.1.4. Effect of few labeled images

In order to study the effect of using a very few labeled images on the prediction performance, we use again the SCUT-FBP images in which the size of labeled images is set to 10% (50 images) and the remaining 90% images are used for testing.

Table 5
Summary of the performances of the supervised and semi-supervised methods on SCUT-FBP with a 10%–90% data partition.

Method	MAE ↓	RMSE ↓	PC % ↑	ϵ-error ↓
1-NN	0.0956	0.1248	51.59	0.2423
Ridge Regression	0.0946	0.1310	23.89	0.2328
Gaussian ϵ-SVR	0.0735	0.0955	71.51	0.1761
LGC	0.0983	0.1336	17.91	0.2470
FME	**0.0707**	**0.0923**	**72.91**	**0.1682**
NFME	0.0724	0.0946	72.90	0.1711

Table 6
Average performances obtained with three graph-based score propagation schemes: LGC, FME, and NFME. The dataset used is M^2B.

Set	Method	MAE ↓	RMSE ↓	PC % ↑
Eastern	LGC	0.1502	0.1827	20.51
	FME	**0.1352**	**0.1668**	**44.82**
	NFME	0.1358	0.1671	44.55
Western	LGC	0.1438	0.1742	34.99
	FME	0.1141	**0.1422**	**63.38**
	NFME	**0.1132**	0.1424	63.22
Both	LGC	0.1484	0.1801	22.90
	FME	0.1346	0.1665	43.58
	NFME	**0.1303**	**0.1624**	**48.05**

Table 5 summarizes the average results obtained over ten random splits.

As it can be seen, the performance of all methods dropped since only 10% of the images (50 images) were labeled. We can also observe that the supervised NN and linear regression methods as well as the semi-supervised LGC method gave poor results. By comparing the PC of the ϵ-SVR, FME and NFME methods in Tables 4 and 5, we can conclude that the superiority of the semi-supervised FME and NFME methods with respect to the supervised ϵ-SVR becomes more significant. This shows the benefits of using the graph-based semi-supervised frameworks for the face beauty assessment whenever labeled photos are scarce.

6.2. M^2B dataset

In this section, we present the results obtained on M^2B dataset, with a configuration of 50% of the samples as labeled data and the other 50% of the samples as the unlabeled/test data. All experiments are carried out doing 10 stratified splits of the data.

Table 6 shows the results of applying the three different graph-based label propagation schemes on M^2B dataset. In this experiment, we consider three sets: the first set contains 620 images of Eastern subjects, the second set contains 620 images of Western subjects. The third one contains the whole M^2B dataset. The best performances are shown in bold. As it can be seen, the FME and NFME methods achieved the best performances. We can also observe that when the two types of faces Eastern and Western are mixed, the performance of all schemes dropped. This suggests that for a computational model the features about beauty are not the same for every ethnicity. Recall that, in M^2B dataset, the eastern faces were rated by eastern subjects, and western faces were rated by western subjects. Since machine learning tries to imitate human expertise, using face images belonging to mixed ethnicities and rated by more than one ethnicity will be more difficult than using face images belonging to one ethnicity and rated by that ethnicity.

We can also observe that the NFME method gave the best performances for the mixed case. Compared to the SCUT results, the performances obtained on M^2B dataset are worse than those obtained with SCUT-FBP dataset. This is due to two main reasons: (i) the images in M^2B dataset are more challenging (some faces correspond to mannequin faces), (ii) the rating process is based on ordering ten images at a time using 40 raters. On the other hand, in SCUT-FBP dataset every face image got the opinion of 70 raters on average.

Table 7
MAEs obtained with supervised schemes and the proposed three graph-based score propagation schemes: LGC, FME, and NFME. The dataset used is M^2B.

	Method	Eastern	Western
State-of-the art	1-NN	2.11	1.92
	Ridge Regression	1.95	1.87
	Neural Network	1.80	1.76
	F-A-T	1.80	1.69
	DFAT	1.77	1.66
Our schemes	LGC	1.50	1.42
	FME	1.35	1.14
	NFME	**1.35**	**1.13**

Table 7 depicts the comparison between the best MAEs of some supervised methods (Nguyen et al., 2013), which were carried out with 2-fold cross-validation, and ours, which has the same data proportions, i.e., train/test is 50%/50%. The FAT method is a cascaded estimation of the regression defined the DFAT method. Our MAEs are multiplied by 10, because we used the normalized labels (dividing the original score by 10). As it can be seen, the NFME method has provided the best MAE.

6.3. SCUT-FBP5500 dataset

In this section, we present the results obtained on the SCUT-FBP5500 dataset. Table 8 shows the results of applying six different methods on the SCUT-FBP5500 dataset. This table depicts the performance obtained with five different experiments. The first four experiments correspond to learning from a given gender and ethnicity. These experiments correspond to Asian Female data (2,000 images), Asian Male data (750 images), Caucasian Female data (2,000 images), and Caucasian Male data (750 images), respectively. The fifth experiment corresponds to the use of the whole dataset (5,500 images). All experiments were conducted using the five-fold cross validation scheme in which 80% of images are used for training and the remaining 20% are used for testing. The features are given by the layer fc6 of the VGG-Face net.

From the results depicted in Table 8, we can observe that the best supervised method was the non-linear ϵ-SVR method, and the best semi-supervised method was the NFME method. The best PCs were obtained for Asian Female and Caucasian Female models. This can be explained by the fact that these two datasets have 2000 images each, making the size of labeled images equal to 1600 images. On the other hand, the size of the labeled images for the Asian Male and Caucasian Male cases is 600 images. It is interesting to note that the model learned on a mixture of genders and ethnicities have provided a performance that can be slightly worse than that obtained in the other cases. Nevertheless, this performance has not been significantly deteriorated. This can be plausible since ethnicity and gender have different representations.

6.4. Score-based propagation versus label distribution based propagation

For some datasets, the face photos may have label distributions instead of a single score. In our case, these label distributions are available for SCUT-FBP and SCUT-FBP5500 in which the distribution is over five levels of beauty. We compare the performance of the NFME method when it propagates single valued scores and when it propagates the label distribution over the same graph. The NFME was chosen since it gave the best results among the semi-supervised methods and since it can perform a basic label

Table 8
Average performances obtained with three graph-based score propagation schemes: LGC, FME, and NFME. The dataset used is SCUT-FBP5500.

Subset	Method	MAE ↓	RMSE ↓	PC % ↑	ϵ-error ↓
Asian Female	1-NN	0.0923	0.1206	63.09	0.2610
	Ridge Regression	0.0580	0.0742	85.41	0.1362
	Gaussian ϵ-SVR	0.0550	0.0714	86.57	0.1264
	LGC	0.1202	0.1427	58.52	0.3554
	FME	0.0564	**0.0707**	86.09	0.1309
	NFME	**0.0550**	0.0713	**86.72**	**0.1256**
Asian Male	1-NN	0.0863	0.1182	59.29	0.2248
	Ridge Regression	0.0569	0.0739	82.88	0.1249
	Gaussian ϵ-SVR	**0.0538**	**0.0701**	84.68	**0.1136**
	LGC	0.1032	0.1316	53.08	0.2768
	FME	0.0557	0.0722	83.72	0.1202
	NFME	0.0540	0.0704	**84.80**	0.1150
Caucasian Female	1-NN	0.0825	0.1080	70.71	0.2133
	Ridge Regression	0.0624	0.0778	83.98	0.1395
	Gaussian ϵ-SVR	**0.0550**	0.0701	87.27	**0.1161**
	LGC	0.1226	0.1424	62.64	0.3553
	FME	0.0554	**0.0701**	87.07	0.1172
	NFME	0.0553	0.0702	**87.52**	0.1181
Caucasian Male	1-NN	0.0752	0.0978	68.71	0.2003
	Ridge Regression	0.0567	0.0733	81.32	0.1316
	Gaussian ϵ-SVR	**0.0505**	0.0650	85.53	**0.1064**
	LGC	0.0983	0.1232	59.50	0.2721
	FME	0.0506	0.0650	85.46	0.1071
	NFME	0.0525	0.0663	**85.67**	0.1173
Whole dataset	1-NN	0.0864	0.1153	64.52	0.2330
	Ridge Regression	0.0570	0.0736	84.50	0.1293
	Gaussian ϵ-SVR	0.0539	0.0693	86.41	0.1167
	LGC	0.1137	0.1376	56.45	0.3239
	FME	0.0570	0.0734	84.60	0.1283
	NFME	**0.0535**	**0.0691**	**86.60**	**0.1151**

Table 9
Performance of two NFME variants on Asian-Female data of the SCUT-FBP5500 dataset.

	MAE ↓	RMSE ↓	PC ↑	ϵ-error ↓
NFME (score)	**0.0550**	**0.0713**	**86.72**	**0.1256**
NFME (label dist.)	0.0750	0.0926	85.35	0.1714

distribution propagation. When the NFME method is used for label distribution propagation, the label distributions of labeled images are given by the label matrix $\mathbf{Y}$, and the unknown soft label distributions are given by the matrix $\mathbf{F}$. In these cases, each of these matrices has five columns. In order to compare the performances of the NFME that outputs single valued predicted scores and the NFME that outputs label distributions, the latter are converted to a single valued score. This is achieved by averaging the levels using the obtained probability distribution.

Table 9 summarizes the performance of the two NFME variants on the Asian-Female data of the SCUT-FBP5500 dataset. These results were obtained using the five-fold cross validation scheme. As it can be seen, the performance of the variant that estimates label distribution is worse than the variant that propagates a single valued score. One plausible explanation is that NFME estimates soft class label distribution that can be more suitable for discrete classification. While label distribution can be useful for uncertainty estimation, it is still not clear what can be gained when a single valued score should be used.

7. Conclusions and discussions

Current methods for face beauty assessment are fully supervised. A limitation of these approaches is the scarcity of labeled face images. The paper has introduced two main contributions. Firstly, semi-supervised paradigms are proposed for the face beauty prediction problem. We exploit graph-based score propagation methods in order to enrich model learning without the need of additional labeled face images. Secondly, a non-linear flexible manifold embedding for solving the score propagation was proposed.

The proposals were tested on three public datasets for face beauty analysis: SCUT-FBP, M^2B, and SCUT-FBP5500. These experiments as well as many comparisons with supervised schemes show that the non-linear semi-supervised scheme compares favorably with the best supervised scheme.

Obviously, the proposed semi-supervised schemes have the limitation that the similarity graph should be computed prior to the estimation of the beauty prediction model, namely the semi-supervised learning model. In other words, if the graph quality is bad (e.g., the graph is very dense), the prediction accuracy might be affected. Therefore, as a future work we envision to overcome this limitation by deploying a joint estimation of the pairwise similarity graph and the unknown prediction model. The other limitation is related to the data themselves. By nature the face beauty prediction problem suffers from imbalanced data. For instance, it is well known that the number of faces with average attractiveness is very high. A remedy to this limitation is to use state-of-the-art data augmentation techniques in order to increase the least and most attractive faces.

The proposed semi-supervised scoring framework opens the door to other image-based applications. It paves the way to virtually all applications to adopt continuous scores instead of the usual discrete labels that are usually used. These applications can be pain level estimation, driver drowsiness detection, subject concentration detection, facial expression recognition, etc. In addition, all applications whose natural output is a number (age estimation, number of present objects, etc.) can directly benefit from the proposed framework.

Future work may envision exploring the following research directions. Firstly, we may exploit image similarities in order to rectify some human generated ground-truth beauty scores.

Secondly, we may target the joint estimation of the pairwise similarity graph and the unknown prediction model. Thirdly, we would use multiple views as inputs (i.e., multiple image descriptors) to the objective functional that estimates the prediction model in order to improve the prediction results over the use of one single descriptor. In this case, the unknown model will be estimated from a fused graph where the fusion is carried out by auto-weighted schemes. Fourthly, we intend to use multi-task estimation for solving the face beauty prediction problem in a more accurate way. In this kind of schemes, several outputs are simultaneously estimated. For instance, the output can be the gender, the ethnicity and the face beauty score.

Declaration of Competing Interest

None.

Credit authorship contribution statement

Fadi Dornaika: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Writing - original draft, Writing - review & editing. **Kunwei Wang:** Data curation, Software, Validation. **Ignacio Arganda-Carreras:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation. **Anne Elorza:** Data curation, Software. **Abdelmalik Moujahid:** Formal analysis, Investigation, Validation, Writing - original draft, Writing - review & editing.

References

Aarabi, P., Hughes, D., Mohajer, K., & Emami, M. (2001). The automatic measurement of facial beauty. In *Systems, man, and cybernetics, 2001 IEEE international conference on: 4* (pp. 2644–2647). IEEE.

Bainbridge, W., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General, 142*(4), 1323–1334.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computing, 15*(6), 1373–1396.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research, 7*, 2399–2434.

Bottino, A., & Laurentini, A. (2010). The analysis of facial beauty: An emerging area of research in pattern analysis. In *ICIAR. Part I, LNCS (6111)* (pp. 425–435).

Chen, A., German, C., & Zaidel, D. (1997). Brain asymmetry and facial attractiveness: Facial beauty is not simply in the eye of the beholder. *Neuropsychologia, 35*(4), 471–476. Cited By 52. doi: 10.1016/S0028-3932(96)00065-6.

Cheng, B., Yang, J., Yan, S., Fu, Y., & Huang, T. (2010). Learning with l1-graph for image analysis. *IEEE Transactions on Image Processing, 19*(4), 858–866.

Cunningham, M., Roberts, A., Barbee, A., Druen, P., & Wu, C.-H. (1995). "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), 261–279. Cited By 433. doi: 10.1037/0022-3514.68.2.261.

DataTang (2019). Retrieved from: https://en.datatang.com/en/index.

Dornaika, F., & Bosaghzadeh, A. (2015). Adaptive graph construction using data self-representativeness for pattern classification. *Information Sciences, 325*, 118–139.

Dornaika, F., Dhabi, R., Bosaghzadeh, A., & Ruichek, Y. (2016). Dynamic adaptive graph construction: Application to graph-based multi-observation classification. *Ieee international conference on pattern recognition*.

Dornaika, F., & El Traboulsi, Y. (2016). Learning flexible graph-based semi-supervised embedding. *IEEE Transactions on Cybernetics, 46*(1), 206–218.

Dornaika, F., & Traboulsi, Y. E. (2017). Matrix exponential based semi-supervised discriminant embedding. *Pattern Recognition, 61*, 92–103.

Dornaika, F., Traboulsi, Y. E., & Assoum, A. (2013). Adaptive two phase sparse representation classifier for face recognition. *LNCS 8192. advanced concepts for intelligent vision systems*.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation, 18*(1), 119–142.

Fan, Y., Liu, S., Li, B., Guo, Z., Samal, A., Wan, J., & Li, S. Z. (2018). Label distribution-based facial attractiveness computation by deep residual learning. *IEEE Transactions on Multimedia, 20*(8), 2196–2208. doi:10.1109/TMM.2017.2780762.

Gan, J., Li, L., Zhai, Y., & Liu, Y. (2014). Deep self-taught learning for facial beauty prediction. *Neurocomputing, 144*, 295–303.

Gan, J., Wang, B., & Xu, Y. (2015). In Y.-J. Zhang (Ed.), *A novel method for predicting facial beauty under unconstrained condition* (pp. 350–360)). Cham: Springer International Publishing.

Gray, D., Yu, K., Xu, W., & Gong, Y. (2010). Predicting facial beauty without landmarks. In *Computer Vision–ECCV 2010* (pp. 434–447).

Gunes, H., & Piccardi, M. (2006). Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies, 64*(12), 1184–1199. doi:10.1016/j.ijhcs.2006.07.004.

He, R., Zheng, W.-S., Hu, B.-G., & Kong, X.-W. (2011). Nonnegative sparse coding for discriminative semi-supervised learning. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on* (pp. 2849–2856).

Kagian, A., Dror, G., Leyvand, T., Cohen-Or, D., & Ruppin, E. (2007). A humanlike predictor of facial attractiveness. In *Advances in neural information processing systems* (pp. 649–656).

Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., & Ruppin, E. (2008). A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research, 48*(2), 235–243.

Kamnitsas, K., Castro, D. C., Folgoc, L. L., Walker, I., Tanno, R., Rueckert, D., Glocker, B., Criminisi, A., & Nori, A. V. (2018). Semi-supervised learning via compact latent space clustering. ArXiv:abs/1806.02679.

Klare, B., & Jain, A. (2013). Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*(6), 1410–1422.

Langlois, J., & Roggman, L. (1990). Attractive faces are only average. *Psychological Science, 1*(2), 115–121. Cited By 753. doi: 10.1111/j.1467-9280.1990.tb00079.x

Larrabee, W. F. J. (1997). Facial beauty:myth or reality? *Archives of Otolaryngology–Head and Neck Surgery, 123*(6), 571–572. Cited By 15. doi: 10.1001/archotol.1997.01900060013001.

Laurentini, A., & Bottino, A. (2014). Computer analysis of face beauty: A survey. *Computer Vision and Image Understanding, 125*(Supplement C), 184–199.

Liang, L., Lin, L., Jin, L., Xie, D., & Li, M. (2018). SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. arXiv:1801.06345v1 [cs.CV] 19 Jan 2018.

Liu, S., Fan, Y.-Y., Guo, Z., Samal, A., & Ali, A. (2017). A landmark-based data-driven approach on 2.5d facial attractiveness computation. *Neurocomputing, 238*, 168–178. doi:10.1016/j.neucom.2017.01.050.

Liu, S., Fan, Y.-Y., Samal, A., & Guo, Z. (2016). Advances in computational facial attractiveness methods. *Multimedia Tools and Applications, 75*(23), 16633–16663.

Liu, W., & Chang, S.-F. (2009). Robust multi-class transductive learning with graphs. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on* (pp. 381–388). IEEE.

Mu, Y. (2013). Computational facial attractiveness prediction by aesthetics-aware features. *Neurocomputing, 99*, 59–64.

Nguyen, T. V., Liu, S., Ni, B., Tan, J., Rui, Y., & Yan, S. (2013). Towards decrypting attractiveness via multi-modality cues. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 9*(4), 28.

Nie, F., Tian, L., Wang, R., & Li, X. (2019). Multiview semi-supervised learning model for image classification. *IEEE Transactions on Knowledge and Data Engineering*.

Nie, F., Wang, X., Jordan, M. I., & Huang, H. (2016). The constrained laplacian rank algorithm for graph-based clustering. *AAAI conference on artificial intelligence*.

Nie, F., Xu, D., Tsang, I. W.-H., & Zhang, C. (2010). Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing, 19*(7), 1921–1932.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition.. In *BMVC: 1* (p. 6).

Raducanu, B., & Dornaika, F. (2014). Embedding new observations via sparse-coding for non-linear manifold learning. *Pattern Recognition, 47*(1), 480–492.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & M. Bernstein, e. a. (2018). Imagenet large scale visual recognition challenge. arXiv:1409.0575, 2014.

Schmid, K., Marx, D., & Samal, A. (2008). Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition, 41*(8), 2710–2717.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 806–813).

Smieja, M., Myronov, O., & Tabor, J. (2018). Semi-supervised discriminative clustering with graph regularization. *Knowledge-Based Systems, 151*, 24–36. doi:10.1016/j.knosys.2018.03.019.

Sousa, C., Rezende, S., & Batista, G. (2013). Influence of graph construction on semi–supervised learning. In *European conferene on machine learning* (pp. 160–175).

Sutic, D., Rreskovic, I., Huic, R., & Jukic, I. (2010). Automatic evaluation of facial attractiveness. In *Proceedings of the 33rd international convention* (pp. 1339–1342).

Traboulsi, Y. E., & Dornaika, F. (2018). Flexible semi-supervised embedding based on adaptive loss regression: Application to image categorization. *Information Sciences, 444*, 1–9.

Wang, S., Shao, M., & Fu, Y. (2014). Attractive or not?: Beauty prediction with attractiveness-aware encoders and robust late fusion. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 805–808). ACM.

Whitehill, J., & Movellan, J. R. (2008). Personalized facial attractiveness prediction. In *IEEE international conference on automatic face gesture recognition* (pp. 1–7).

Xie, D., Liang, L., Jin, L., Xu, J., & Li, M. (2015). Scut-fbp: A benchmark dataset for facial beauty perception. In *Systems, man, and cybernetics (SMC), 2015 ieee international conference on* (pp. 1821–1826). IEEE.

Xu, J., Jin, L., Liang, L., Feng, Z., & Xie, D. (2015). A new humanlike facial attractiveness predictor with cascaded fine-tuning deep learning model. arXiv:1511.02465v1.

Yan, M., Duan, Y., Deng, S., Zhu, W., & Wu, X. (2016). Facial beauty assessment under unconstrained conditions. In *2016 8th international conference on electronics, computers and artificial intelligence (ECAI)* (pp. 1–6).

Zhang, D., Chen, F., & Xu, Y. (2016). *Computer models for facial beauty analysis*. Springer.
Zhang, D., Zhao, Q., & Chen, F. (2011). Quantitative analysis of human facial beauty using geometric features. *Pattern Recognition, 44*, 940–950.
Zhang, L., Zhang, D., Sun, M., & Chen, F. (2017). Facial beauty analysis based on geometric feature: Toward attractiveness assessment application. *Expert Systems With Applications, 82*, 252–265.
Zhang, Z., Liu, L., Shen, F., Shen, H. T., & Shao, L. (2018). Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in neural information processing systems* (pp. 321–328).
Zhu, X., Ghahramani, Z., Lafferty, J., et al. (2003). *Semi-supervised learning using gaussian fields and harmonic functions* (vol. 3, pp. 912–919).