Sense Beauty via Face, Dressing, and/or Voice

Tam V. Nguyen
National University of Singapore
tamnguyen@nus.edu.sg

Si Liu
National Laboratory of Pattern Recognition
sliu@nlpr.ia.ac.cn

Bingbing Ni
Advanced Digital Sciences Center
bingbing.ni@adsc.com.sg

Jun Tan
National University of Defense Technology
tanjun.nudt@gmail.com

Yong Rui
Microsoft Research Asia
yongrui@microsoft.com

Shuicheng Yan
National University of Singapore
eleyans@nus.edu.sg

ABSTRACT

Discovering the secret of beauty has been the pursuit of artists and philosophers for centuries. Nowadays, the computational model for beauty estimation has been actively explored in computer science community, yet with the focus mainly on facial features. In this work, we perform a comprehensive study of female attractiveness conveyed by single/multiple modalities of cues, i.e., face, dressing and/or voice, and aim to uncover how different modalities individually and collectively affect the human sense of beauty. To this end, we collect the first Multi-Modality Beauty (M^2B) dataset in the world for female attractiveness study, which is thoroughly annotated with attractiveness levels converted from manual k-wise ratings and semantic attributes of different modalities. A novel Dual-supervised Feature-Attribute-Task (DFAT) network is proposed to jointly learn the beauty estimation models of single/multiple modalities as well as the attribute estimation models. The DFAT network differentiates itself by its supervision in both attribute and task layers. Several interesting beauty-sense observations over single/multiple modalities are reported, and the extensive experimental evaluations on the collected M^2B dataset well demonstrate the effectiveness of the proposed DFAT network for female attractiveness estimation.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Algorithms, Experimentation, Human Factors

Keywords

{Face, Dressing, Voice} attractiveness, Attributes, Dual-supervised Feature-Attribute-Task network

MM'12, October 29–November 2, 2012, Nara, Japan.

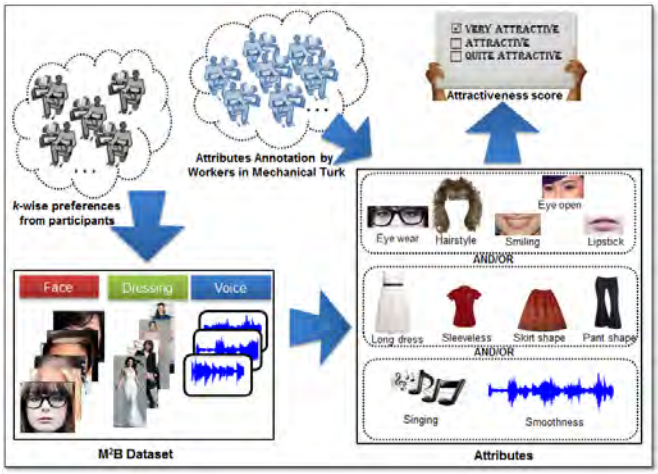


Figure 1: The proposed comprehensive study of sensing human beauty via single/multi-modality cues. The collected dataset contains three modalities, *i.e.*, face, dressing and voice. The attractiveness scores are given by k-wise preferences from participants. All modalities are labeled with extensive attributes by Amazon Mechanical Turk due to its large amount. Both visual and vocal features as well as labeled attributes are collectively utilized to build computational models for estimating female beauty score.

1. INTRODUCTION

Discovering the secret of beauty has been the pursuit of artists and philosophers for centuries [15, 20, 8]. The study on what are the essential elements and how their combinatorial mechanism affects the attractiveness of a person is valuable for many potential applications. For example, when we know the underlying rules how the female's dress and face jointly influence her attractiveness, one system can be developed to recommend how a female can become more attractive by choosing a specific type of lipstick or other make-ups according to her face shape and dress. This research benefits other areas such as fashion, cosmetic, and targeted advertisement. The problem has also attracted many interests from computer science researchers recently. There exist softwares for both automatic human-like facial beauty assessment [19, 25] as well as face beautification [34, 21]. There exist some works from multimedia and social science communities on

attractiveness study based on faces [7, 16, 25], bodies [18, 29], and voices [23].

In essence, most of these studies attempt to answer one question: "what elements constitute beauty or attractiveness for human". However, how these individual elements collaborate with each other and jointly affect the human sense of beauty has received little attention. We believe that different modalities can complement and affect each other and there do exist certain underlying interacting mechanism that makes a lady attractive, which is even more important than the elements themselves. There exist obvious examples to support this argument. In real life, a female may not have very attractive face, but she may have a good taste of how to select dresses and makeups to match her face shape, which then makes her also very attractive entirely. Therefore, in this paper, we study how different modalities, *i.e.*, face, dress and voice individually and collectively affect the human sense of beauty (or attractiveness).

To facilitate the human attractiveness study, we first collect the largest multi-culture (Eastern and Western females), Multi-Modality (face, dressing, and voice) Beauty (M^2B) dataset to date. Then, the participants are invited to annotate the k-wise preference for each k randomly selected examples (with modalities of face, and/or dressing, and/or voice). Afterwards, the k-wise preferences are converted into the global attractiveness scores for all the samples. In addition, a set of carefully designed attributes are annotated for each modality of samples by using Mechanical Turk [1], and used as the bridge for boosting attractiveness estimation performance. Finally, we present a novel tri-layer learning framework, called Dual-supervised Feature-Attribute-Task (DFAT) network, to unify the attribute prediction and multi-task attractiveness prediction within a unified formulation. Eventually, the extensive experiments on the collected M^2B dataset demonstrate several interesting cross-modality observations as well as the effectiveness of our proposed DFAT framework.

Figure 1 illustrates the proposed framework for sensing beauty via multi-modality cues. The main contributions of this work can be summarized as follows.

1. To the best of our knowledge, we conduct the first comprehensive study on how multiple interacting modalities (*i.e.*, face, dress and voice) individually and collectively affect the sense of female attractiveness.
2. We propose a user friendly k-wise ranking tool for reliable large-scale attractiveness annotation.
3. We propose a novel dual-supervised framework where attribute models and attractiveness models are learned simultaneously, which is superior over conventional two-stage framework, namely first learning the attribute models followed by learning the models from attributes to attractiveness score.
4. Last but not least, using our computational models, we study the commonalities and differences between the Eastern and Western on how they sense beauty.

The rest of the paper is organized as follows. Section 2 discusses the related work. Then, we describe the process of dataset collection and annotation, and propose the DFAT framework for attractiveness estimation in Section 3 and 4, respectively. Experiments and discussions are presented in Section 5. Section 6 concludes this work.

2. RELATED WORK

In literature, most computer science researches have focused on identifying attractive facial characteristics. Most approaches to this problem can be considered as geometric or landmark feature based methods. Aarabi et al. built a classification system based on 8 landmark ratios and evaluated the method on a dataset of 80 images rated on a scale of $1-4$ [7]. Eisenthal et al. used an ensemble of features that include landmark distances and ratios, an indicator of facial symmetry, skin smoothness, hair color, and the coefficients of an eigenface decomposition [16]. Their method was evaluated on two datasets of 92 images each with ratings $1-7$. Kagian et al. later improved upon their method using an improved feature selection method [25]. Recently, Guo et al. have explored the related problem of automatic makeup/beautification application, which uses an example to transfer a style of makeup to a new face [21]. Gray et al. presented a method of both quantifying and predicting female facial beauty using a hierarchical feed-forward model without the landmarks [19].

The attractiveness of bodies has also been investigated. Glassenberg et al. discussed the attractive women have a high-degree of facial symmetry, a relatively narrow waist, and V-shaped torso [18]. Studies on attractiveness based on clothing are more centralized in the area of sociology. For example, Lennon's study [29] investigated whether people perceive others differentially as a function of the attractiveness of their clothing with a set of experiments. To the best of our knowledge, there is no existing work specially studying the attractiveness of clothing, which shows the advantage and uniqueness of our work.

Apart from visual attractiveness, Zuckerman et al. investigated the voice attractiveness [40]. They found that attractive voices were louder and more resonant. In addition to this, they found some gender differences, including low-pitch-male voices are perceived as more attractive, while the attractiveness of female voices could not be captured by spectrographic analysis. Susan et al. investigated the relationship between ratings of voice attractiveness and sexually dimorphic differences in shoulder-to-hip ratios and waist-to-hip ratios, as well as different features of sexual behavior [23].

In general, the contemporary datasets used by the previous studies are small-scale and usually restricted to a very small and subset of the population (e.g. uniform ethnicity, age), with less-changed expression, pose and lighting condition. The images are generally studio-quality photos taken by professional photographers. Furthermore, there is no multi-modality datasets for our proposed study on how multiple interacting modalities affect the human sense of beauty. Thus, we construct the new dataset which will be described in the next section.

3. DATASET CONSTRUCTION

3.1 Data Collection

There exist several datasets [19, 7, 25] for attractiveness study but none of them is suitable as they usually contain only one modality of features. Therefore, in order to make a study on our proposed problem, we require a large dataset of faces, dressings and voices along with ground-truth attractiveness scores. However, no such the datasets

Table 1: Exemplar keywords used for downloading online videos from YouTube and their corresponding numbers of high quality video clips downloaded to construct M^{2}B dataset.

Query	#Clip	Query	#Clip
SuperGirl	35	X Factor Auditions	60
Happy Girl	20	Got Talent Auditions	70
Guess-Guess	40	Eurovision	10
Korea Got Talent	5	American Idol	10
Chinese New Year Event	3	Next Top Model	5
Others	22	Total	280

Figure 2: The exemplar user interface of our attractiveness ranking tool on one batch of instances for Western labelling task. The corresponding information of selected top 2 instances disappeared; and the user then proceeds to click "Like" for the next most attractive instance.

are currently publicly available. Thus, we construct Multi-Modality Beauty (M^{2}B) dataset with face, dressing images and voices. The attractiveness scores are annotated by human subjects. To study how people from different cultures sense beauty, the constructed dataset includes two ethnic groups: Western and Eastern.

The data are collected mainly from the popular video sharing website YouTube [2] similar to [12]. To diversify the dataset, we selected images from videos of various TV reality shows, talk shows, looking-for-idol-like programs with contestants from both Western and Eastern countries. Some of the exemplar programs are SuperGirl, Happy Girl, Guess-Guess, Chinese New Year Event, American Dancing with the stars, Eurovision, Britain Next Top Model, American Idol, X Factor, and Britain Got Talent show with its franchises [1]. Besides, we also collected data from the online academic talks, poems, songs from TED Talk [3], VideoLectures [4], etc. Unlike the previous datasets [31] the face images of which are cropped from low-quality photos taken by cell-phone cameras, we only select high quality video clips, e.g. at least 360 pixels wide. The durations of the clips are varied from 28 seconds to 2 hours 48 minutes (averagely 21 minutes per clip). Note that for each clip, there may exist multiple females. The details of actual videos utilized in M^{2}B are reported in Table 1.

For each female instance, we extract several frames from the video. We run Viola-Jones face detector [38] on these frames to extract frontal faces. All the faces are resized to 128×128 pixels. A well-trained human detector [39] is applied on all images, and only the high-confidence detection outputs are kept. Note that for the dressing image, the face size is small, and generally cannot be used for sensing beauty. We extract 5 seconds duration for voice information of each instance. Eventually, we select only one face photo, one full body photo and one voice snippet for one female instance. Totally, our dataset consists of equal 620 vs. 620 instances for Westerners and Easterners, respectively. This database shall be released for the public usage on the research of female beauty.

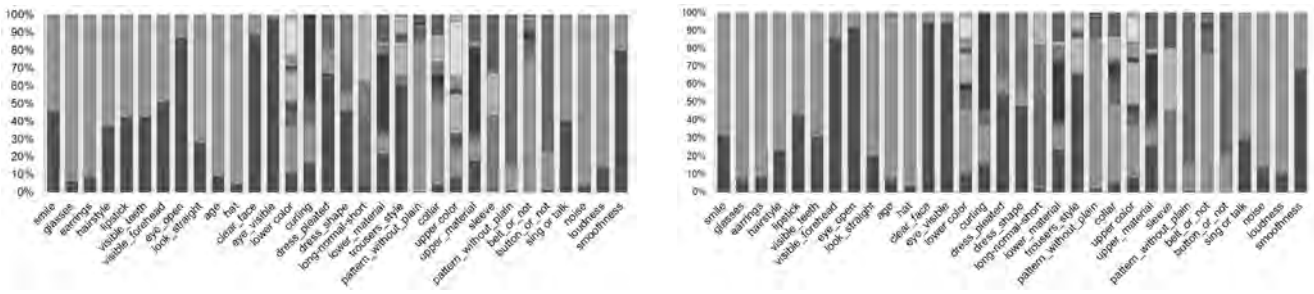
3.2 Ground-truth Attractiveness Score

3.2.1 *Absolute value vs. pair-wise vs. k-wise ratings*

There are several kinds of ratings that can be used for annotation for this task. The most popular ones are absolute ratings where a user is presented with a single image and asked to give a score, typically between 1 and 10. Most previous works have used some versions of absolute value ratings, which are usually presented in the form of a Likert scale [30]. This form of rating requires each image to be rated by many users such that a distribution of ratings can be gathered and averaged to estimate the true score. This method is obviously not ideal because different users with different backgrounds have different priors in rating images. Another method used in [33] is to ask a user to sort a collection of images according to some criteria. This method is likely to give reliable ratings, but it is impractical for users to sort a large dataset. The most recent method is to present a user with a pair of images and ask which one is more attractive. This method presents a user with a binary decision, which they have found can be made more quickly than an absolute rating. Greg et al. applied pair-wise comparison for attractiveness study [19]. However, it is usually non-trivial to convert these pair-wise ratings into global scores, which is important for subsequent tasks.

In this work, we try to avoid the disadvantages of all above methods and propose a k-wise comparison (with k set as 10). The number of pair-wise preferences obtained from one k-wise rating is $\binom{k}{2}$. For example, when k is 10, the number of pair-wise preferences is 45. Totally, 40 participants (17 females and 23 males who are students and staff members of a university) ranged from 19 to 40 years old (μ=26.4, σ=4.1) participated in the data ranking task. There exists cross-race effect or other-race bias, *i.e.*, the tendency for people of one race to have difficulty in recognizing and processing faces and facial expressions of members of a race or ethnic group other than their own [37, 9]. Therefore, the participants have been split into two groups based on their ethnicities. Westerners labeled for Western group while Easterners labeled for Eastern group. This is the main reason we did not ask workers from Mechanical Turk since the current system cannot well control the ethnicity of the workers. Each participant performs some of the six following tasks: 1) faces (F), 2) voices (V), 3) dressings (D), 4) faces and dressing (FD), 5) faces and voices (FV), and 6) faces, dressings, and voices (FDV). We exclude DV (dressings and voices) task since it is an unnatural and impractical scenario. Each k-wise rating shows 10 random instances to the participant, who ranks each instance from the most attractive to the least attractive. Figure 2 shows the user interface of the k-

[1]Got Talent franchises are at America, Australia, Albania, Bulgaria, China, Denmark, France, Holland, Korea, Romania, and Serbia.



(a) Eastern (b) Western

Figure 3: The distributions of attributes annotated by Mechanical Turk workers. For the two-option attributes, the bottom part corresponds to 'yes', the top part corresponds to 'no'. For the multiple-option attributes, please refer to Figure 4 for their options in order. (Please view in high 400% resolution).

wise rating tool for one k-wise rating. When the user clicks "Like" button, the information of the corresponding instance disappears and the user proceeds to the remaining instances. The rating process continues until every instance is ranked. Each instance in each task has been ranked by at least 15 different participants.

3.2.2 *k-wise ratings to global attractiveness score*

In our study, we assume that in a large sense people agree on a consistent opinion on facial attractiveness, which is also the assumption of the previous studies [19]. Each individual's opinion can be varied due to factors like culture, race, and education. As aforementioned, k-wise ratings are fast to collect, but in order to use them for subsequent learning tasks, we need to convert the ratings into the global attractiveness scores. To obtain the scores from k-wise, we minimize a cost function defined such that as many of the pairwise preferences as possible are preserved and the scores lie within a specified range, where the pairwise preferences are converted from the k-wise ratings and $\binom{k}{2}$ pair-wise preferences can be obtained from each k-wise rating. Denote Ω as the set of pairwise preferences for k-wise ratings, we formulate the conversion problem as,

$$\begin{aligned} &\min_{s} J(s) = ss^T + \tau \sum_{(p,q)\in\Omega} \xi_{pq}, \\ &\text{s.t.} \quad \begin{cases} \xi_{pq} \geq 0, \quad \forall\ (p,q) \in \Omega, \\ s_p - s_q \geq 1 - \xi_{pq}, \ \forall\ (p,q) \in \Omega, \end{cases} \end{aligned} \tag{1}$$

where $s = [s_1, s_2, ..., s_n]$ is the global attractiveness score row vector for all the n instances of one task, and the constraints correspond to the pairwise preferences. The problem of (1) is right the popular Ranking SVM [24]. Finally, all the scores are re-scaled to be within [1, 10] for each of six tasks.

3.3 Attributes Annotation

Recently, methods that exploit the semantic attributes of objects have attracted significant attention in the computer vision community. The usefulness of attributes has been demonstrated in several different application areas [10, 17, 26, 27]. Visual attributes are important for understanding object appearance and for describing objects to other people. Automatic learning and recognition of attributes can complement category-level recognition and improve the degree for machines to perceive visual objects. Therefore, we also want to explore the usage of attributes in the attractiveness study.

In this context, the attributes defined are not limited to visual attributes. Attributes associated with different modalities such as *eye wear*, *dressing collar*, or *voice smoothness* are also used. In this work, we manually define different

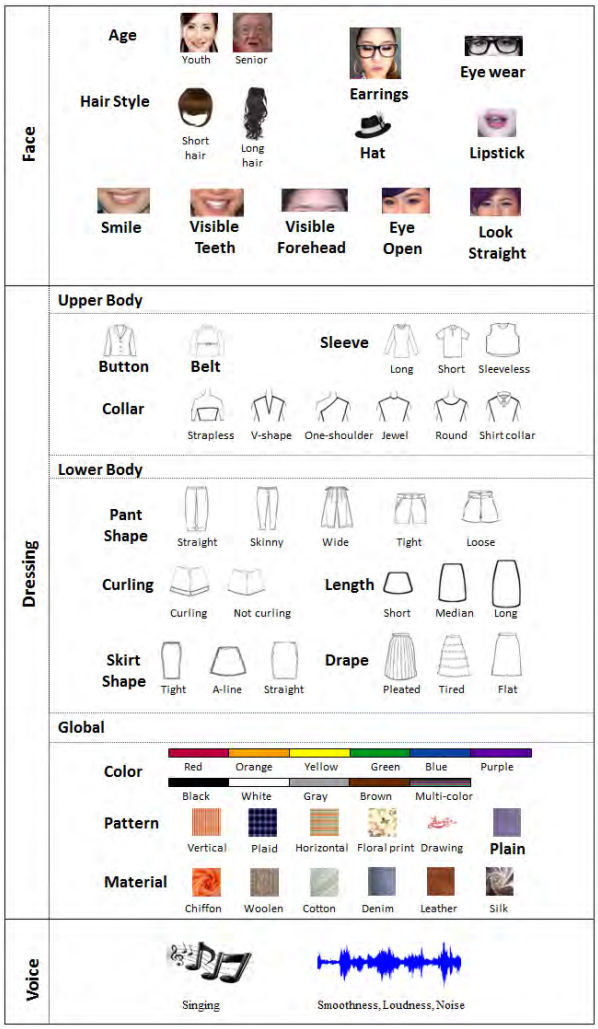


Figure 4: The attributes of different modalities. An example or line drawing is shown to illustrate each attribute value.

types of attributes. The selection of the attributes is determined by the discussions founded on previous related research papers [10, 26] and also related Internet contents, which mean stylists have already implicitly contributed. All the attributes labeled from the dataset are listed in Figure 4. The defined attributes can be summarized into three classes, i.e., face, dressing and voice attributes. Mechanical Turk workers are responsible for labeling the attributes of the newly built dataset [1]. Due to the difficulty in distinguishing the attribute values, different numbers of annotators are assigned to each labeling task. A label was considered as a ground truth if at least more than half of the annotators agreed on the value of the label. To the best of our knowledge, this dataset has the most complete attribute annotations among all contemporary datasets.

4. THE PROPOSED FRAMEWORK

In this section, we first explain the feature extraction applied on the extracted face, body and voice of the collected M^2B dataset. A novel framework, which learns attributes and attractiveness simultaneously, is later introduced.

4.1 Features

4.1.1 Facial features

We extract the following popular features, local binary patterns (LBP) [32], Gabor filter responses [14], Color moment for the frontal faces.

LBP is basically a finescale descriptor that captures small texture details. We adopt the same notation $\mathrm{LBP}_{\mathrm{P,R}}$ as in [32], where R is the radius of the circle to be sampled, and P is the number of sampling points. Denote the ring feature for image pixel (x, y) as $B(x, y) =< b_{P-1}, \ldots, b_1, b_0 >$, where $b_i \in \{0, 1\}$. It is common to transform $B(x, y)$ into decimal code via binomial weighting:

$$\mathrm{LBP}_{\mathrm{P,R}}(x, y) = \sum_{i=0}^{P-1} b_i 2^i,$$

which characterizes image textures over the neighborhood of (x, y).

Gabor filter is another appropriate feature for texture representation. Gabor filters encode facial shape and appearance information over a range of spatial scales. The Gabor functions applied for location (x, y) are used as the following form.

$$\mathrm{G}(x, y) = \exp\left(\frac{X^2 + \gamma^2 Y^2}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} X\right),$$

where $X = x\cos\theta + y\sin\theta$ and $Y = -x\sin\theta + y\cos\theta$ are the orientations of the Gabor filters with angle θ which varies between 0 and π. The other parameters, aspect ratio γ, effective width σ, wavelength λ are set as in [35]. In the implementation, we apply 5 scales and 8 orientations to obtain Gabor responses.

Color moment is a low-level color measurement and consists of the first order (mean of color values) and the second order moments (variance of color values) of the input image block. In this work, we divide the input image into 4×4 blocks and then compute the overall color moment. Color information is highly correlated to some attributes such as lipstick or visible forehead.

Since the concatenated features are high-dimensional, we use PCA to reduce the facial dimensionality to 250.

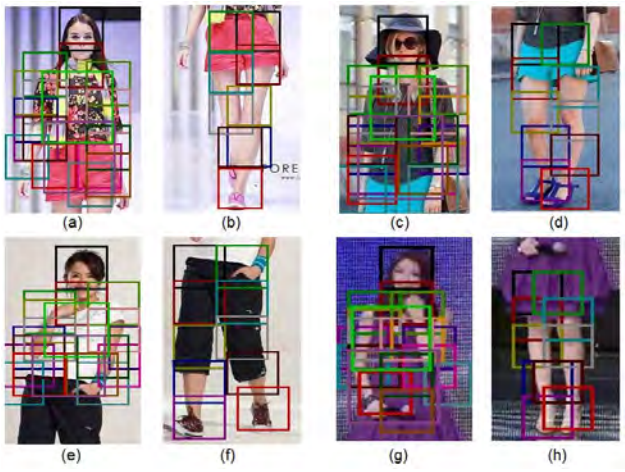


Figure 5: The dressing bounding boxes for dressing feature extraction for Western (a-d) and Eastern (e-h). Each region roughly corresponds to functional parts of the dressing.

4.1.2 Dressing features

We consider dressing as the combination of two main parts, upper and lower. Each part consists of the mixture of mini parts, similar to the human body. Following [11, 36], we extract 5 kinds of features from the 20 upper-body parts and 10 lower-body parts. Figure 5 shows the exemplar boxes of dressings. The features used include HOG [13], LBP, Color moment, Color histogram and Skin descriptor. HOG and LBP features are related to dressing texture attributes such as *collar* or *curling*. Meanwhile, Color moment, Color histogram and skin descriptor are useful for depicting color-relevant dressing attributes such as *shirt color* or *pattern*. More specifically, each human part is first partitioned into 16 smaller, spatially evenly distributed regular blocks. 5 features are extracted from each block and features from all blocks are finally concatenated to represent a human part. The block based features can roughly preserve relative spatial information inside each human part. The dimensionality of dressing feature after PCA is 300.

4.1.3 Vocal features

We apply the audio feature extraction for the audio snippets in M^2B dataset. The vocal features are extracted by using MIRToolbox [28]. Each voice feature is related to one of the audio dimensions traditionally defined in audio theory. The audio sequence is decomposed into successive frames, which are then converted into the spectral domain, frequency domain and pitch domain. Accordingly, the audio features related to pitch, to spectrum (zerocross, low energy, rolloff, entropy, irregularity, brightness, skewness, flatness, roughness), to tonality (chromagram, key strength and key self-organising map) and to dynamics (root mean square energy) are extracted. Another set of features inherited from automatic speech recognition is used, which is the set of mel-frequency spectral coefficients. Additionally, some features related to rhythm, namely tempo, pulse clarity and fluctuation, are also used. Eventually, these audio features are concatenated and reduced to 50-D by PCA.

4.2 Dual-supervised Feature-Attribute-Task (DFAT) Network

Most previous studies [7, 19, 16] utilize features directly in order to predict the attractiveness score. As earlier mentioned, in this work, we explore the usage of attributes serving as the intermediate layer in order to perform the tasks. The conventional approach to integrate attributes is to perform the following two steps separately: 1) learn the regression model from raw features to attributes and 2) learn another regression model from the output attributes of training data to attractiveness scores. The drawback of this approach is to introduce the unexpected errors into the second regression stage, and it cannot guarantee the outputs from the first model are optimal for the second model. Therefore, we propose to fuse these two steps together and simultaneously optimize them in the sense that two steps mutually affect each other.

We propose the novel Dual-supervised Feature-Attribute-Task (DFAT) network, which jointly learns the beauty estimation models of single/multiple modalities, where the semantic attributes are shared by different tasks, namely the beauty estimation models of different types of features and their combinations. The model contains three layers, i.e., feature, attribute and task layers. The proposed DFAT Network is illustrated in Figure 6. DFAT learns two types of regression models simultaneouly by minimizing two types of prediction errors, one is feature-to-attribute error, and the other is attribute-to-attractiveness error. Note that the main difference between conventional Neural Network [22] and our proposed method is the supervision existing in both attribute and task layers of DFAT. Formally, let us denote X^m as the training data matrix for modality m, where each column is a feature vector and $m \in \{1, 2, 3\}$, A^m as the regression matrix from raw features to attributes, $Attr^m$ as the groundtruth attributes of modality m, X_t^m as training data for modality m in task t where $t \in \{1, 2, 3, 4, 5, 6\}$, s_t as the groundtruth attractiveness score row vector for task t, and ω_t^m as the row regression vector for converting attributes of modality m to task t. The learning problem for DFAT network is then formulated as:

Algorithm 1 Procedure to solve Problem (2)

Input: matrices X^m, X_t^m, $Attr^m$, s_t , parameter λ_1, λ_2.
Initialize: A^m by solving $\min \|A^m X^m - Att^m\|^2$, $e_1 = \infty, e_2 = 0$.
while not converged **do**

1. Fix the others and update ω_t^m by:

$$\omega_t^m = \big(s_t - \sum_{p \in \{1,2,3\} \setminus m} \omega_t^p A^p X_t^p\big)(A^m X_t^m)^T (A^m X_t^m (A^m X_t^m)^T + \frac{\lambda_2}{\lambda_1} I)^{-1},$$

2. Fix the others and update A^m by gradient descent.

$$A^m = A^m - \gamma \nabla F(A^m),$$

where γ is the step size and $\nabla F(A^m)$ is defined as

$$\nabla F(A^m) = A^m X^m X^{mT} - Att^m X^{mT} + \lambda_2 A^m + \lambda_1 \sum_{t=1}^{6} \omega_t^{mT} \big(\sum_p \omega_t^p A^p X_t^p - s_t \big) X_t^{mT}.$$

3. Compute $e_2 = \|F\|_F$.
4. Check the convergence condition: $\|e_1 - e_2\| < \varepsilon$.
5. Update e_1: $e_1 = e_2$.
end while
Output: The optimal solution $\{A^{m*}\}, \{\omega_t^{m*}\}$.

$$\min_{\{A^m\},\{\omega_t^m\}} F = \frac{1}{2} \sum_{m=1}^{3} \|A^m X^m - Attr^m\|^2 + \frac{\lambda_1}{2} \sum_{t=1}^{6} \| \sum_{m=1}^{3} \omega_t^m A^m X_t^m - s_t\|^2 + \frac{\lambda_2}{2} \sum_{m=1}^{3} \big(\|A^m\|^2 + \sum_{t=1}^{6} \|\omega_t^m\|^2\big). \quad (2)$$

The first term is the regression from features to attributes, the second term is the regression from attributes to attrac-

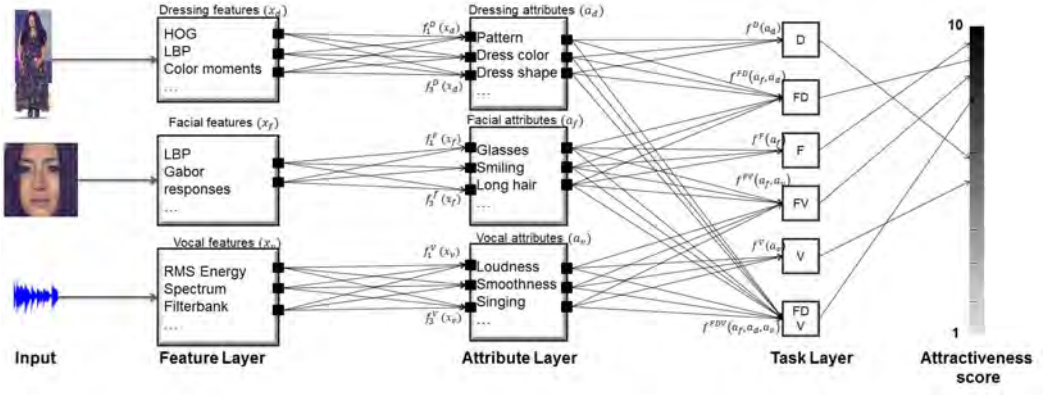


Figure 6: Dual-supervised Feature-Attribute-Task Learning Framework. First, for each modality, i.e, face, dress and voice, different kinds of features are extracted. Then the beauty estimation models of single/multiple modalities are jointly learned. During the learning process, the semantic attributes are shared by different tasks. Different with traditional Neural Network, the proposed DFAT network can seamlessly combine the training attribute labels in the learning process.

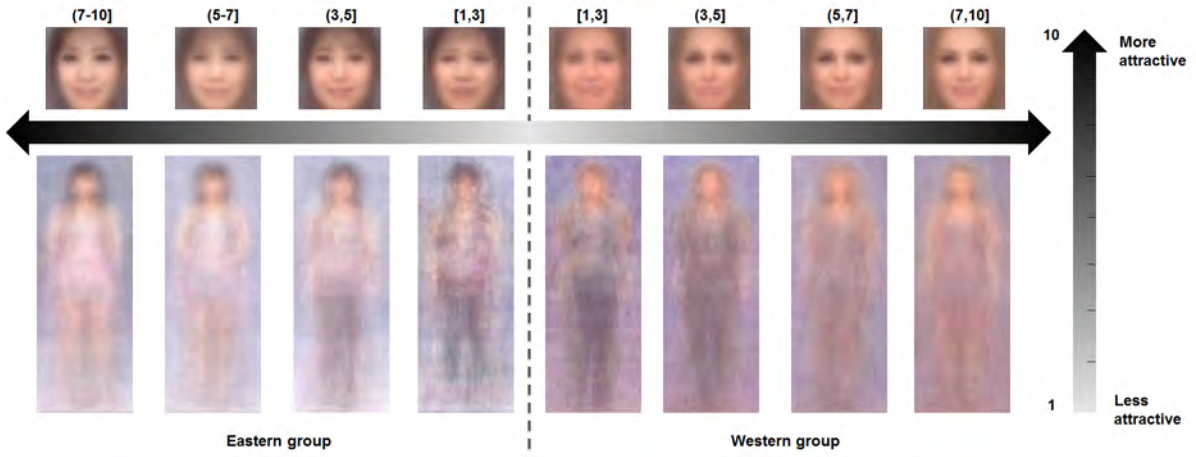


Figure 7: Average faces and dressings of Eastern and Western groups at different attractiveness scores. For better viewing, please see original color pdf file (greatly encouraged for this figure).

tiveness scores, and the last term is the regularization term. Note that for one task t, if the modality m does not exist, we set the corresponding feature matrix X_t^m be all-zero matrix for ease of formulation. The above optimization problem can be solved by any gradient based method and the iterative optimization procedure is listed in Algorithm 1.

5. EXPERIMENTS

In this section, we describe the extensive experiments conducted on the collected M^{2}B dataset for the better understanding of beauty sensing.

5.1 The Attractiveness Scores Distribution

The score of each instance in each task is achieved by solving Equation (1). In order to investigate how the scores distribute in the dataset, we compute the histogram of attractiveness scores across six tasks of both Western and Eastern participants. Recall that Western participants labeled for Western group while Eastern counterparts labeled for Eastern group. We can notice the distributions different between two groups. In Figure 8, the scores of Eastern participants are higher in the range from 4 to 6, while the scores from Western participants dominate the rest of score ranges. This reflects the Golden Mean in traditional Eastern culture [5], where people are more conservative to give the conclusion.

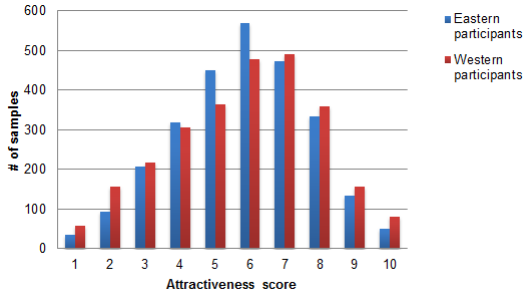


Figure 8: The histograms of attractiveness scores across six tasks of Western and Eastern participants.

5.2 Average Faces and Average Body Shapes

The average faces and dressings show the first glance how people sense the attractiveness. The average faces from the dataset, presented in the top row of Figure 7, have a score within [1, 10]. The average faces present interesting patterns in attractiveness study. One of the early observations in the study of facial beauty was that averaged faces are attractive [8]. There also exist computational tools to generate the average face [6]. Additionally, women with averaged faces have been shown to be considered more attractive. This is possibly due to average features being smoother and, therefore, more comfortable. Average faces are attractive, but not all of them. It is crucial which faces are used to compute an average face. Average face computed from unattractive faces may remain rather unattractive and other ones from attractive faces shall remain attractive. The average faces of *higher scores* look younger and smoother. In contrast, the average faces of *lower scores* look older, and less smooth. Another interesting observation is that Western faces have blonde hair which may blend into the background, while Eastern faces have black hair which has good contrast from the face color.

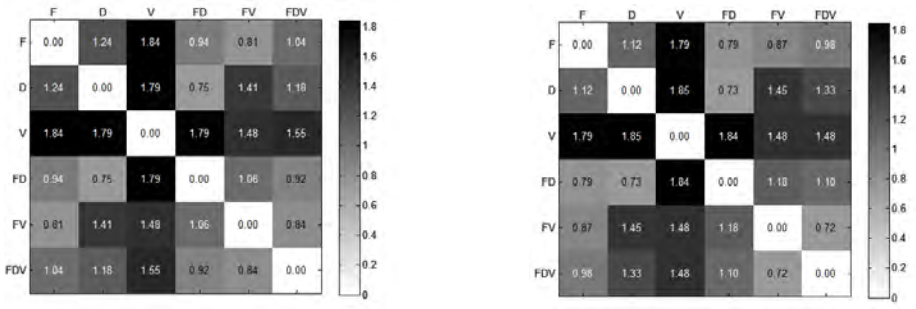
Similarly, the average dressings are shown in the bottom row of Figure 7 with scores within [1,10]. The less attractive dressings are trouser-like while the more attractive dressings are skirt-like. The background is the same for both groups. The Western group prefers dark color dressing, while the Eastern group favors bright color.

5.3 Cross-modalities Beauty Sense Discrepancy

We compute the distance matrices of both Eastern and Western groups. The distance matrices provide the dissimilarities of attractiveness scores in different tasks. Each element of the matrix represents the distance score between two tasks. This allows more detailed analysis on the differences among tasks. Figure 9 shows the distance matrices that represent the distance scores between the tasks, for Easterns and Westerns, respectively. The distance of task i and task j is computed as follows.

$$d_{i,j} = \frac{1}{n}\sum_{l=1}^{n} |s_l^i - s_l^j|, \tag{3}$$

where s_l^i is the attractiveness score of instance l in task i, s_l^j is the corresponding score of the same instance l in task j, and n is the number of instances. As can be seen from Figure 9, the results from the Voice-only task are far away from



(a) Eastern (b) Western

Figure 9: The distance matrices of Eastern and Western groups. The distances are calculated based on the groudtruth attractiveness scores of different tasks. The small values indicate that two tasks are similar with each other.

all of other tasks' results. In other words, the attractiveness score of one instance in the Voice-only task is greatly different to her score in another task. There also exist the differences between Western and Eastern. Face vs. Voice has the largest dissimilarity of Eastern group. Meanwhile, Dressing vs. Voice has the largest distance in the Western group. Face vs. Face-Dressing has the smallest dissimilarity in Eastern group, while the smallest dissimilarity in Western group is Face-Dressing-Voice vs. Face-Voice. Generally, the scores given in the Face-only task have been changed when other modalities are added.

5.4 Within-culture Attractiveness Prediction

In this subsection, we investigate the attractiveness prediction problem within cultures. For each experiment, we perform a standard 2-fold cross validation test to evaluate the accuracy of our algorithms on the M^{2}B dataset. In 2-fold cross-validation, the original dataset is randomly evenly partitioned into 2 subsets. The cross-validation process is then repeated 10 times, with each of the 2 subsets used as the testing data and the other subset as training data. The 10 results from the folds are averaged to report the final results. We use the Mean Absolute Error (MAE) to evaluate the accuracy of the attractiveness prediction. The MAE is defined as the average of the absolute errors between the predicted attractiveness score and the ground truth.

$$MAE = \sum_{i=1}^{n} |\hat{s}_i - s_i|/n, \tag{4}$$

where s_i is the ground truth attractiveness score for the test instance i, $\hat{s}_i$ is the estimated score, and n is the total number of test instances for one task.

We then compare the performance of DFAT network with four baselines.

1. 1-NN: 1-NN classifier is applied to find the nearest neighbor, and assign the score of the neighbor to the query instance.

2. Ridge Regression: We apply the Ridge Regression to obtain the predicted attractiveness score from the raw features directly.

3. Neural Network: We apply feed-forward neural network to retrieve the attractiveness score from the raw features directly. Note that the difference between NN and DFAT network is the hidden layers. The DFAT network differentiates itself by its supervision in both attribute and task layers.

4. F-A-T: We learn the linear regression between the features and attributes, and then train the second linear regression between the output attributes of training data and the attractiveness scores.

Note that for the first 3 baselines, the attributes are not used. Regarding DFAT, we implement the Algorithm 1 with $\lambda_1 = 0.01$, $\lambda_2 = 10^{-3}$, $\gamma = 10^{-3}$, and $\varepsilon = 10^{-4}$ to learn the transfer matrices. In our experiment, 10 to 20 iterations are required for convergence. As can be seen in Table 2, MAEs of 1-NN are worst in all cases. F-A-T achieves better performance than two baselines, Ridge Regression and Neural Network. The better performance of F-A-T shows the advantage of using attributes in attractiveness study. Meanwhile, our proposed DFAT outperforms all of compared algorithms. Face-only task gets the highest MAE in both two cultures. In the opposite side, Face-Voice task achieves the lowest MAE among tasks across Eastern and Western. In addition, the task of Face-Dressing-Voice also reaches the similar MAE to Face-Voice task. For all baselines, the MAEs tend to have the large value when more modalities are added. In contrast, for DFAT, the results show that more modalities combining with face generally reduce the error when predicting the attractiveness level. In other words, multiple modalities boost the performance of attractiveness prediction.

5.5 Cross-culture Attractiveness Prediction

People from different cultures are often attracted to the same type of faces. This agreement among individuals of different ages and from different cultures suggests attractiveness judgements are not arbitrary but have a "biological basis". Thus, we are interested in exploring the cross-culture attractiveness prediction between Eastern and Western. For this experiment, we train the data on one group and test on the different group by using DFAT.

Table 3 shows the MAEs of the cross-culture experiment on M^{2}B dataset. The MAEs of all tasks increase dramatically compared to the results of training and testing on the same ethnic group. The highest error lies on dressing tasks. The difference can be explained by the significant difference in the average dressings. Recall that Westerners prefer darker color, while Easterners favor brighter color. Also, the MAEs of Face and Face-Dressing task are also high due to the significant difference of faces. Meanwhile, the MAEs

Table 2: MAEs of different algorithms on M^2B dataset (the training/testing data within the same culture).

	Eastern						Western					
Algorithm	**F**	**D**	**V**	**FD**	**FV**	**FDV**	**F**	**D**	**V**	**FD**	**FV**	**FDV**
1-NN	2.11	1.50	1.39	1.74	2.16	1.96	1.92	2.02	1.78	1.98	2.25	2.23
Ridge Regression	1.95	1.39	1.15	1.52	1.93	1.79	1.87	1.76	1.37	1.66	2.09	2.13
Neural Network	1.82	1.37	1.12	1.47	1.79	1.82	1.76	1.62	1.38	1.53	1.85	1.87
F-A-T	1.80	1.33	1.12	1.45	1.79	1.67	1.69	1.54	1.34	1.54	1.91	1.93
DFAT	1.77	1.33	1.09	1.42	0.98	1.04	1.66	1.50	1.29	1.50	1.01	1.12

of Face-Voice task achieve is the lowest. This result agrees with the previous finding in [40] that attractive voices have the same effect as attractive faces, meaning that vocal attractiveness parallels visual attractiveness.

Table 3: MAEs of cross-culture attractiveness prediction experiment on M^2B dataset.

Task	Train on Eastern - Test on Western	Train on Western - Test on Eastern
F	1.91	2.22
D	2.55	2.71
V	1.55	1.62
FD	2.39	2.42
FV	1.15	1.20
FDV	1.57	1.52

5.6 Task-specific Important Attributes

Firstly we conduct the experiment to measure the accuracy [2] of attribute prediction whose results are shown in Figure 10. Generally, the prediction results are acceptable, except for some attributes such as *lipstick* and *hairstyle*. Then we investigate attributes' importance in different tasks. It is curious to know what the model really learns. We use the absolute values of coefficients obtained from DFAT to represent the importance of attributes to the tasks. All of the values are rescaled within [0, 1] for each task.

Figure 11 depicts all the attributes' responses in both Eastern and Western group. For the face modality, the first impression is that a bright smile is attractive for both groups. Besides, the results show that the ageing has the large impact to the female attractiveness. At a closer look, the attribute *age* is extremely sensitive in Eastern compared with Western. Meanwhile, *glasses* is well-responded in Western. For the dressing modality, the responses of attributes are different in two ethnic groups. The skirt or pant 'length' (i.e. long, normal, short) is very important to determine the attractiveness in Eastern group, but not for Western group. For Western group, *sleeve* is important compared with *dressing patterns*. Surprisingly, *color* has the small impact on the dressing attractiveness. For the voice modality, *smoothness* is the most important attribute to Eastern people. The reason can be explained as the language of Easterners is the isolating language. In the meantime, the loudness of voice plays the main factor to Western people to decide the voice attractiveness. Additionally, there is an 'overridden' phenomena which means one attribute is important but shows less important after new modality is added. For example, in Eastern group, the age's importance decreases when Face is combined with Voice. Another example is that in Western group, *loudness* is important in Voice modality, but its importance lowers when Face, Dressing and Voice are combined together.

[2]Accuracy is $(TruePositive + TrueNegative)/Total$.

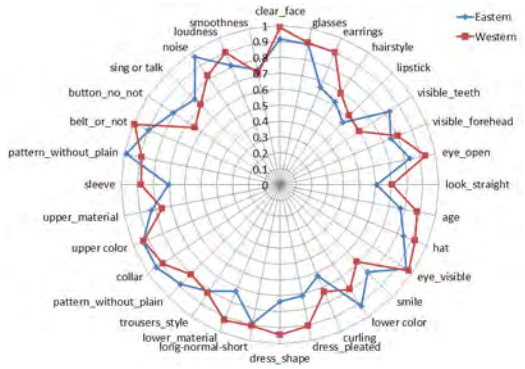


Figure 10: The accuracy of attribute prediction.

6. CONCLUSION AND FUTURE WORK

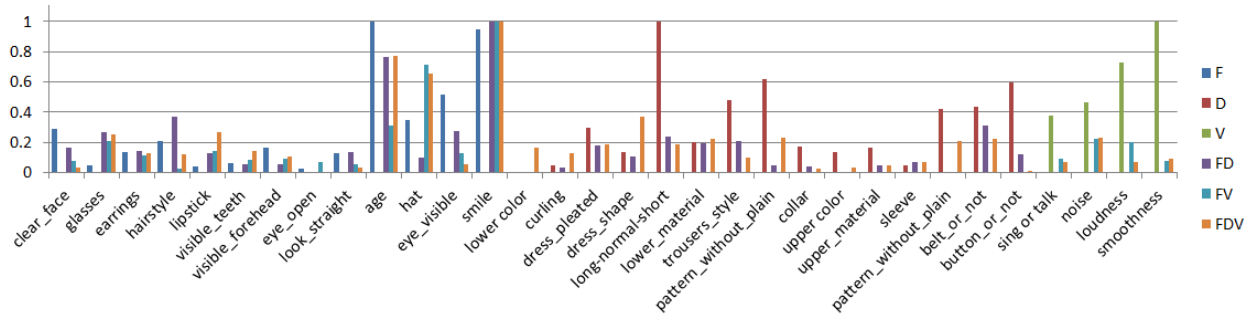
We have investigated the human sense of beauty via multimodality cues. To the best of our knowledge, we are the first to build female attractiveness dataset of multiple modalities, facial, dressing and voice. Its multi-cultural property may also be helpful for the further researches on cultures. We also proposed a tri-layer learning framework, namely DFAT, to learn attributes and attractiveness simultaneously. Extensive experimental evaluations on the M^2B dataset well demonstrate the effectiveness of the proposed DFAT framework for female attractiveness prediction. We believe that the work may help artificial intelligent reaches a new step in order to sense the beauty as human. One of our future directions is to use the latent variable approach for attribute mining. We also plan to investigate more suitable features of different modalities for attractiveness prediction, and build workable real system for practical applications.

7. ACKNOWLEDGEMENT

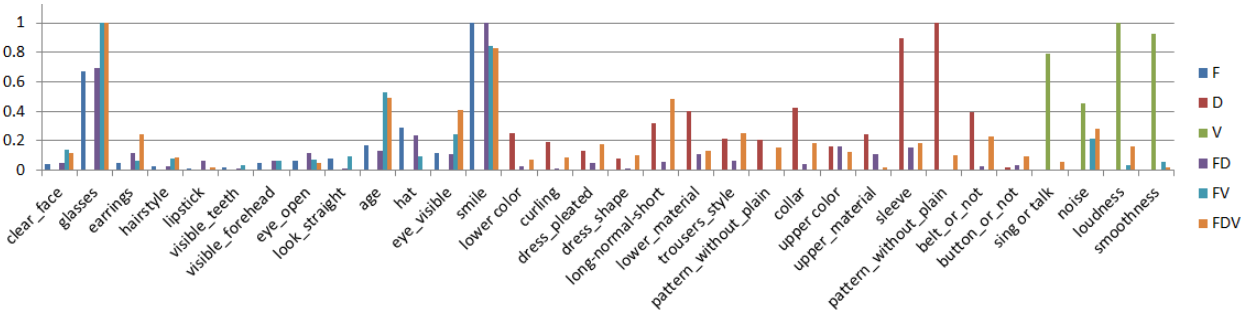
This work was supported by NExT Research Center funded under the research grant WBS. R-252-300-001-490 by MDA, Singapore. Bingbing Ni is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR).

8. REFERENCES

[1] Amazon Mechanical Turk. http://www.mturk.com.
[2] YouTube. http://www.youtube.com.
[3] TED Talk. http://www.ted.com.
[4] Video Lectures. http://videolectures.net.



(a) Eastern



(b) Western

Figure 11: The importance of different attributes with respect to different tasks for (a) Eastern and (b) Western. For better viewing, please see original color pdf file.

[5] Doctrine of the Mean. `http://en.wikipedia.org/wiki/Doctrine_of_the_Mean`.

[6] Average Face. `http://www.faceresearch.org/demos/average`.

[7] P. Aarabi, D. Hughes, K. Mohajer, and E. M. The automatic measurement of facial beauty. *ICSMC*, 2001.

[8] T. Alley and M. Cunningham. Averaged faces are attractive, but very attractive faces are not average. *Psychological Science*, 1991.

[9] M. Beaupre. An ingroup advantage for confidence in emotion recognition judgments: The moderating effect of familiarity with the expressions of outgroup members. In *Personality and Social Psychology Bulletin*, 2006.

[10] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[11] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet- based approach to attribute classification. *ICCV*, 2011.

[12] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *ACM Multimedia*, 2010.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[14] J. Daugman. Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. In *Journal of the Optical Society of America*, 1985.

[15] K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. In *Journal of Applied Social Psychology*, 1972.

[16] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 2006.

[17] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[18] A. Glassenberg, D. Feinberg, B. Jones, A. Little, and L. DeBruine. Sex-dimorphic face shape preference in heterosexual and homosexual men and women. *ASB*, 2009.

[19] D. Gray, K. Yu, W. Xu, and H. Gong. Predicting facial beauty without landmarks. *ECCV*, 2010.

[20] C. Green. All that glitters: A review of psychological research on the aesthetics of the golden section. In *Perception*, 1995.

[21] D. Guo and T. Sim. Digital face makeup by example. *CVPR*, 2009.

[22] S. Haykin. Neural networks. In *Prentice Hall*, 1999.

[23] S. Hughes, F. Dispenza, and G. Gallup. Ratings of voice attractiveness predict sexual behaviour and body configuration. *Evolution and Human Behavior*, 2004.

[24] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

[25] A. Kagian, G. Dror, T. Leyvand, D. Cohen-Or, and E. Ruppin. A humanlike predictor of facial attractiveness. *NIPS*, 2005.

[26] N. Kumar, P. N. Belhumeur, and S. K. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008.

[27] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[28] O. Lartillot and P. Toiviainen. MIR in Matlab: A toolbox for musical feature extraction from audio. In *ISMIR*, 2007.

[29] S. Lennon. Effects of clothing attractiveness on perceptions. In *Home Economics Research Journal*, 1990.

[30] R. Likert. Technique for the measurement of attitudes. *Arch. Psychol*, 1932.

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

[32] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *T-PAMI*, 2002.

[33] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[34] P. Pallett, S. Link, and K. Lee. New golden ratios for facial beauty. In *Vision Research*, 2009.

[35] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. In *Nature Neuroscience*, 1999.

[36] Z. Song, M. Wang, X. Hua, and S. Yan. Predicting occupation via human clothing and contexts. *ICCV*, 2011.

[37] J. Tanaka, M. Kiefer, and C. Bukach. A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study. In *Cognition*, 2004.

[38] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.

[39] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[40] M. Zuckerman and K. Miyake. The attractive voice: What makes it so? In *Journal of Nonverbal Behavior*, 1993.