Towards Decrypting Attractiveness via Multi-Modality Cues

TAM V. NGUYEN and SI LIU, National University of Singapore
BINGBING NI, Advanced Digital Sciences Center
JUN TAN, National University of Defense Technology
YONG RUI, Microsoft Research Asia
SHUICHENG YAN, National University of Singapore

Decrypting the secret of beauty or attractiveness has been the pursuit of artists and philosophers for centuries. To date, the computational model for attractiveness estimation has been actively explored in computer vision and multimedia community, yet with the focus mainly on facial features. In this article, we conduct a comprehensive study on female attractiveness conveyed by single/multiple modalities of cues, that is, face, dressing and/or voice, and aim to discover how different modalities individually and collectively affect the human sense of beauty. To extensively investigate the problem, we collect the Multi-Modality Beauty (M^2B) dataset, which is annotated with attractiveness levels converted from manual k-wise ratings and semantic attributes of different modalities. Inspired by the common consensus that middle-level attribute prediction can assist higher-level computer vision tasks, we manually labeled many attributes for each modality. Next, a tri-layer Dual-supervised Feature-Attribute-Task (DFAT) network is proposed to jointly learn the attribute model and attractiveness model of single/multiple modalities. To remedy possible loss of information caused by incomplete manual attributes, we also propose a novel Latent Dual-supervised Feature-Attribute-Task (LDFAT) network, where latent attributes are combined with manual attributes to contribute to the final attractiveness estimation. The extensive experimental evaluations on the collected M^2B dataset well demonstrate the effectiveness of the proposed DFAT and LDFAT networks for female attractiveness prediction.

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

General Terms: Algorithms, Experimentation, Human Factors

Additional Key Words and Phrases: {Face, dressing, voice} attractiveness, latent attributes

ACM Reference Format:

Nguyen, T. V., Liu, S., Ni, B., Tan, J., Rui, Y., and Yan, S. 2013. Towards decrypting attractiveness via multi-modality cues. ACM Trans. Multimedia Comput. Commun. Appl. 9, 4, Article 28 (August 2013), 20 pages.
DOI: http://dx.doi.org/10.1145/2501643.2501650

This research is supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office. B. Ni is supported by a research grant from the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*Star).

Authors' addresses: T. Nguyen, S. Liu, and S. Yan, National University of Singapore, 4 Engineering Drive 3, 117583 Singapore; email: {tamnguyen, dcslius, eleyans}@nus.edu.sg; B. Ni, Advanced Digital Sciences Center, 1 Fusionopolis Way, 138632 Singapore; email: bingbing.ni@adsc.com.sg; J. Tan, National University of Defense Technology, No. 137, Yanwachi, Changsha, Hunan 410073 China; email: tanjun.nudt@gmail.com; Y. Rui, Microsoft Research Asia, No. 5, Dan Ling Street, Beijing, 100080 China; email: yongrui@microsoft.com.

DOI: http://dx.doi.org/10.1145/2501643.2501650

1. INTRODUCTION

Decrypting the secret of beauty or attractiveness has been the pursuit of artists and philosophers for centuries [Dion et al. 1972; Green 1995; Alley and Cunningham 1991]. 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-up according to her face shape and dress. Hence, this research benefits areas such as fashion, cosmetic, and targeted advertisement. There exists software for both automatic human-like facial beauty assessment [Gray et al. 2010; Kagian et al. 2005] as well as face beautification [Pallett et al. 2009; Guo and Sim 2009]. There exist some works from multimedia and social science communities on attractiveness study based on faces [Aarabi et al. 2001; Eisenthal et al. 2006; Kagian et al. 2005], bodies [Glassenberg et al. 2009; Lennon 1990], and voices [Hughes et al. 2004].

In essence, most of these studies attempt to answer one question: "which elements combine to form beauty or attractiveness for human?". However, how these individual elements are correlated 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 exists a 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 reality, a female may not have a very attractive face, but she may have a good taste of how to select dresses and makeup to match her face shape, which then makes her also very attractive entirely. Therefore, in this article, we study how different modalities, that is, 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 multiculture (Eastern and Western females), Multi-Modality (face, dressing, and voice) Beauty (M^2B) dataset. In this article, Mongoloid females such as the ones with Chinese, Korean and Japanese origin represent the Eastern group, whereas Caucasian females, who are descended from Angles, Celtic, Latin and Germanic people, represent the Western group [Brinton 1890]. We have removed the ambiguous case such as Arabic and Eurasian females. The labellers were invited to annotate the k-wise preference for each k randomly selected examples (with modalities of face, and/or dressing, and/or voice). Totally, 40 participants (17 females and 23 males who are university students and staff) whose ages are ranged from 19 to 40 years old ($\mu = 26.4, \sigma = 4.1$) participated in the data ranking task. Note that the labellers are from both groups. Afterwards, the k-wise preferences were converted into the global attractiveness scores for all the samples. In addition, a set of carefully designed attributes were annotated for each modality 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. Latent Dual-supervised Feature-Attribute-Task (LDFAT), the combination of latent attributes and DFAT is also introduced to remedy any possible loss of information caused by incomplete manual attributes. 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]http://www.mturk.com.

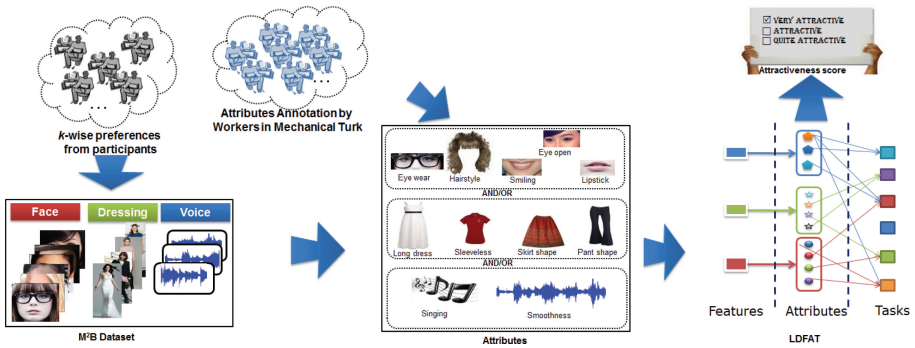


Fig. 1. The proposed framework of sensing human attractiveness via single/multi-modality cues. The collected dataset contains three modalities, that is, face, dressing and voice. The attractiveness scores are given by k-wise preferences from participants. All attributes of each modality are labeled by Amazon Mechanical Turk workers. Both visual and vocal features as well as labeled attributes are collectively utilized to build computational models for estimating female beauty score.

(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 latent 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. Note that we integrate the usage of latent attributes into the framework along with the manually annotated attributes.

(4) Last but not least, using our computational models, we study the commonalities and differences between the Eastern and Western groups on how they sense beauty.

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

2. RELATED WORK

In literature, most computer vision 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 $1-4$ [Aarabi et al. 2001]. 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 [Eisenthal et al. 2006]. 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 [Kagian et al. 2005]. Recently, [Guo and Sim 2009]. have explored the related problem of automatic makeup/beautification application, which uses an example to transfer a style of makeup to a new face. Gray et al. [2010]. presented a method of both quantifying and predicting female facial beauty using a hierarchical feed-forward model without landmarks. The attractiveness of bodies has also been investigated. Glassenberg et al. [2009] found that the attractive women have a high-degree of facial symmetry, a relatively narrow

Table I.
Exemplar categories of downloaded 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

waist, and V-shaped torso. Studies on attractiveness based on clothing are more centralized in the area of Sociology. For example, Lennon's study [Lennon 1990] investigated whether clothing may affect our sense of attraction between humans. Recently, Liu et al. [2012a] introduced an interactive system which recommends the dressings according to the event.

Apart from visual attractiveness, Zuckerman and Miyake [1993] investigated the voice attractiveness. They found that attractive voices were louder and more resonant. In addition to this, they found some gender differences. For example, low-pitch-male voices are perceived as more attractive, while the attractiveness of female voices could not be captured by spectrographic analysis. Hughes et al. [2004] investigated the relationship between ratings of voice attractiveness and sexually dimorphic differences in shoulder-to-hip ratios and waist-to-hip ratios.

An earlier version of this work has been published in ACM Multimedia 2012 [Nguyen et al. 2012]. In this extended work, we integrate the use of latent attributes combining the new facial shape features into the proposed framework.

3. DATASET CONSTRUCTION

3.1 Data Collection

There exist several datasets [Gray et al. 2010; Aarabi et al. 2001; Kagian et al. 2005] for attractiveness study, but none of them is suitable as they usually contain only one modality. 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, such datasets are not currently publicly available. Thus, we constructed Multi-Modality Beauty (M^{2}B) dataset with face, dressing image and voice of each instance. The attractiveness scores are annotated by human subjects. To study how people from different cultures sense beauty, the constructed dataset includes two groups: Western and Eastern.

The data are collected mainly from the popular video sharing website YouTube[2]. 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 Got Talent series.[3]

We then manually select high-quality video clips, for instance, at least 640×360 pixels. The duration of the clips varies from 28 seconds to 2 hours 48 minutes (averagely 21 minutes per clip). The details of actual videos utilized in M^{2}B are reported in Table I. We then cut the longer clips into the smaller snippets of only 5 seconds. The snippets with no female voice have been removed manually. For each

[2]http://www.youtube.com.
[3]Got Talent series are at America, Australia, Albania, Britain, Bulgaria, China, Denmark, France, Holland, Korea, etc.

female instance, we extract several frames from the video. We run Viola-Jones face detector [Viola and Jones 2004] on these frames to extract frontal faces. All the faces are resized to 128×128 pixels. The state-of-the-art human detector [Yang and Ramanan 2011] is applied on all images, and only the high-confidence detection outputs are kept.

The dressing image is the crop of the image around the bounding box of the body including the head. Note that for the dressing image, the face size is small, and generally cannot be used for sensing beauty. We extract 5 seconds duration of voice information for 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. There are 270 voices not matching the face and dressing images. This database is publicly released[4] for the usage on the research of female beauty.

3.2 Ground-Truth Attractiveness Score

3.2.1 *Absolute Value vs. Pairwise 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 [Likert 1932]. This form of rating requires each image to be rated by many users so 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 Oliva and Torralba [2001] 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 can be performed more quickly than an absolute rating. Gray et al. [2010] applied pairwise comparison for attractiveness study. However, it is usually non-trivial to convert these pairwise 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 rating (with k set as 10). The number of pairwise preferences obtained from one k-wise rating is $\binom{k}{2}$. For example, when k is 10, the number of pairwise preferences is 45.

In addition, there exists the cross-race effect or other-race bias, that is, 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 [Tanaka et al. 2004; Beaupre 2006]. 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 scenario. Each k-wise rating shows 10 random instances to the participant, who ranks each instance from the most attractive to the least attractive. Note that when each participant performs annotations on the specific task, he/she will only be shown the corresponding modalities, and the other modalities are hidden. Figure 2 shows the user interface of the k-wise rating tool for one FDV-task 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.

[4]https://sites.google.com/site/vantam/beautysense.

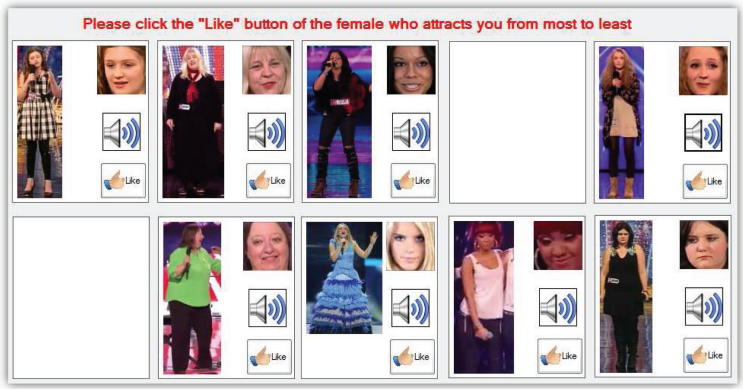


Fig. 2. The user interface of our attractiveness ranking tool on one batch of instances for Western FDV labelling task. The corresponding information of selected top 2 instances disappeared; and the user then proceeds to click "Like" for the next most attractive instance.

3.2.2 *From 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 [Gray et al. 2010]. Individuals' opinions may 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 so that as many pairwise preferences as possible are preserved. The scores lie within a specified range, where the pairwise preferences are converted from the k-wise ratings and $\binom{k}{2}$ pairwise 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,

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

where $s = [s_1, s_2, \ldots, s_n]$ 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) fits well the popular Ranking SVM [Joachims 2002]. Finally, all the scores are rescaled 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 [Parikh and Grauman 2011; Berg et al. 2010; Kumar et al. 2008]. 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 wish to investigate the usage of attributes in the attractiveness study.

In this context, the attributes defined are not limited to facial attributes. Attributes associated with different modalities such as *dressing collar*, or *voice smoothness* are also used. In this work, we manually define different types of attributes. The selection of the attributes is determined by the discussions

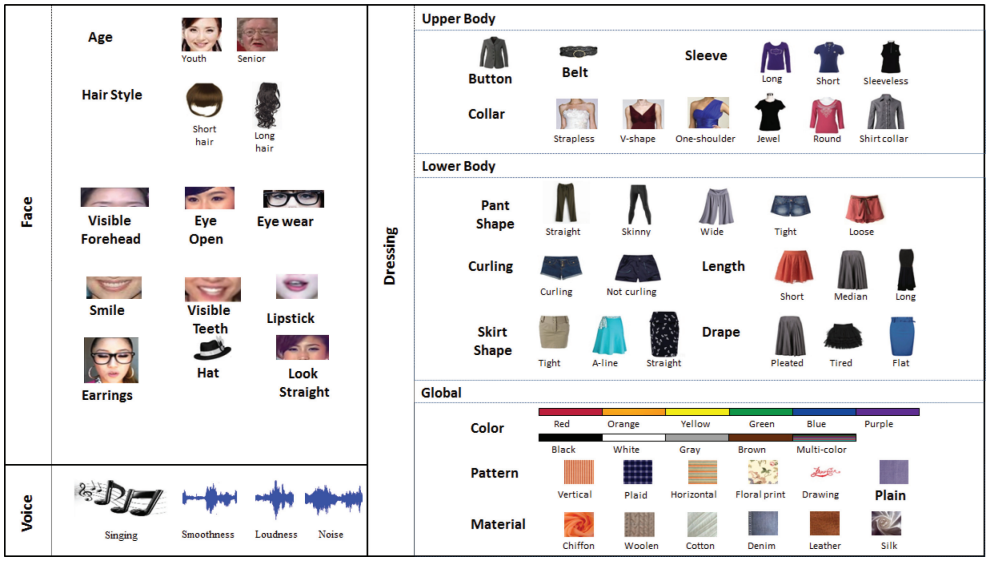


Fig. 3. The attributes of different modalities. An example or line drawing is shown to illustrate each attribute value.

founded on previous related research papers [Berg et al. 2010; Kumar et al. 2008; Liu et al. 2012b] and also Internet-related contents. All the attributes labeled from the dataset are listed in Figure 3. The defined attributes can be summarized into three classes, that is, face, dressing and voice attributes. Mechanical Turk workers are responsible for labeling the attributes of the M^2B dataset. 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. However, by selecting attributes manually, it is clear that this process is subjective and arbitrary, and it does not guarantee that all of the critical features characterizing a task are successfully associated with attribute labels. To address this issue, we propose to integrate manually specified attributes with the latent attributes.

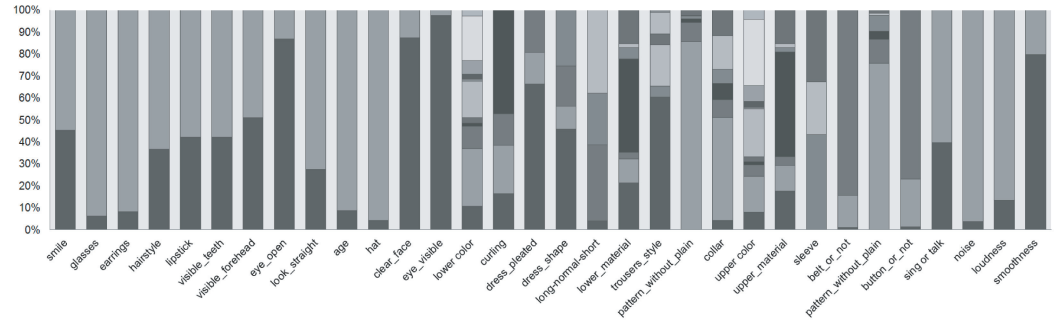
4. THE PROPOSED FRAMEWORK

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.

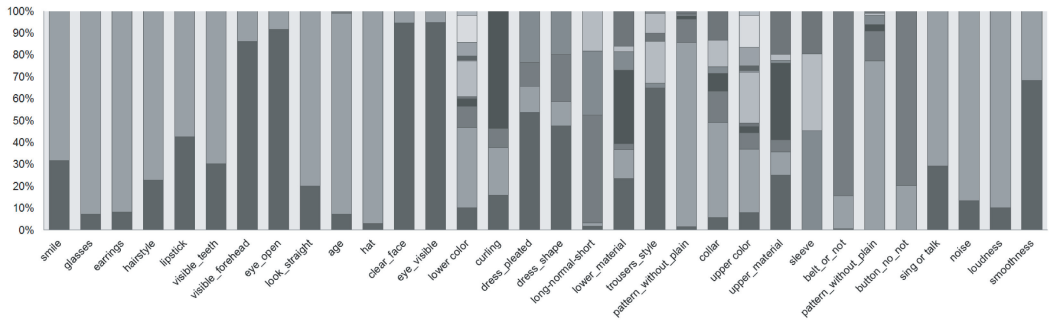
4.1 Features

4.1.1 *Facial Features.* We extract the following popular features, local binary patterns (LBP) [Ojala et al. 2002], Gabor filter responses [Daugman 1985], Color moment, Shape context, and Shape parameters 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 Ojala et al. [2002], 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). In our implementation, the face image is divided into 4×4 grids, for each of which the LBP histogram is created.



(a) Eastern group



(b) Western group

Fig. 4. 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 3. (Please view in high 200% resolution).

Gabor filter is another popular 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, $G(x, y) = \exp(\frac{X^2+\gamma^2Y^2}{2\sigma^2}) \times \cos(\frac{2\pi}{\lambda}X)$, 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 Riesenhuber and Poggio [1999]. 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.

Shape context is utilized as a way of measuring shape similarity [Belongie et al. 2002]. The usage of this feature is based on the assumption that the particular face shape is attractive, that is, oval face shape. We extract 87 landmark points on the contours of the face.[5] For each point p_i on the shape, consider the $n-1$ vectors obtained by connecting p_i to all other points. For the point p_i, the coarse histogram of the relative coordinates of the remaining $n-1$ points, $h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\}$ is defined to be the shape context of p_i. Then we cluster the shape context descriptor space by k-means algorithm identifying a set of k cluster centers and assigning to them a given integer index $I \in [1, k]$.

[5]These points are extracted by commercial software from Omron Corporation.

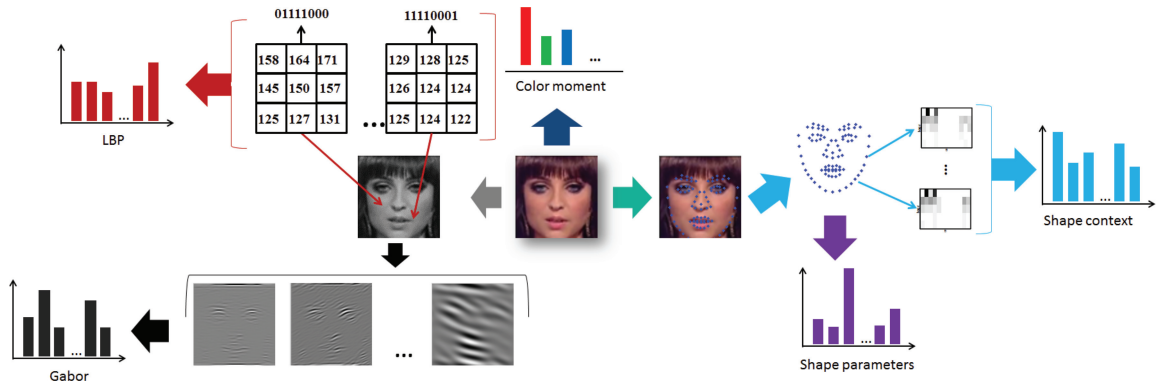


Fig. 5. The representation of the facial features extracted from face. Best viewed in color.

Each shape context descriptor of the points p_i is then projected to the clustered space and can be identified by a single index I_i. The shape context feature can thus be represented by a histogram coding the frequency of appearance of each of the k center indices. k is set as 60 in our implementation.

Shape parameters are used in Active Shape Model (ASM) [Cootes et al. 1995]. We have a set of face images associated with a set of landmark points $s_i \in S$. We consider two-dimensional ASM, and hence there are two scalar values (an x- and a y-coordinate) of each annotation point in the vector s_i. Let us denote $\bar{s}$ as the mean shape from the aligned shapes. We apply PCA to the data as follows, $\hat{s_i} = \bar{s} + P_s b_i^s$, where the b_i^s are vectors of shape parameters. P_s is an $|\bar{s}| \times n$ matrix where columns are orthogonal modes of variation: the first n eigenvectors of the covariance matrix for S, ordered by eigenvalue. n is chosen such that P_s is adequate to represent a certain proportion of the variation in S (95%). Shape parameters b_i^s are used to represent facial shape along with the aforementioned *shape context*.

All of the facial feature representations are illustrated in Figure 5. Note that all of sub-features, that is, LBP, Gabor responses, are normalized with their ℓ_2 norm. Since the concatenated features are high-dimensional, we use PCA to reduce the dimensionality of facial feature to 350.

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 Bourdev et al. [2011] and Song et al. [2011], we extract 5 kinds of features from the 20 upper-body parts and 10 lower-body parts. Figure 6 shows the exemplar parts of dressings. Five features used include HOG [Dalal and Triggs 2005], LBP [Ojala et al. 2002], Color moment, Color histogram and Skin descriptor [Mittal et al. 2011]. Regarding Skin descriptor, we trained a skin classifier, which is a Gaussian Mixture Model with 5 components from the LAB (color space) transformed patches of skin collected from various illuminations. The skin dataset we used is the same as in Mittal et al. [2011]. Based on the GMM classifier, we get a binary mask indicating whether a pixel is skin or not. We split each human part box into 3 different cell structures: 4×4 cells, 3×1 cells and 1×3 cells. Then we calculated the average value of the mask inside each cell to form 22-dim skin features representing 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. Five aforementioned features are extracted from each block and features from 29 blocks (we discard the block including the face) are finally concatenated to form the dressing feature. All of sub-features are normalized with their ℓ_2 norm. The

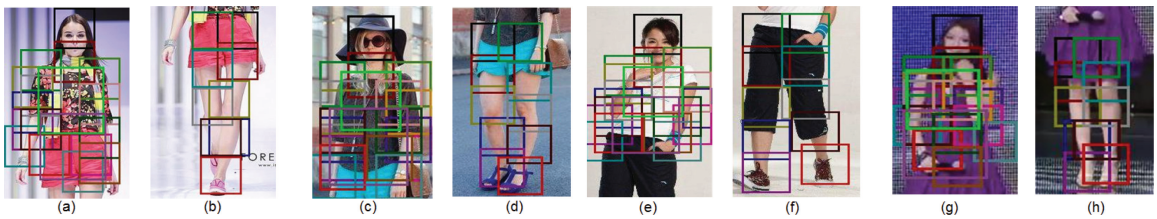


Fig. 6. The dressing bounding boxes for dressing feature extraction for Western (a-d) and Eastern (e-h). Each region roughly corresponds to functional parts of the dressing.

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 [Lartillot and Toiviainen 2007]. 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 [Aarabi et al. 2001; Gray et al. 2010; Eisenthal et al. 2006] 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, that is, feature, attribute and task layers. 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 [Haykin 1999] 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 $m \in \{1, 2, 3\}$ for different modalities, 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 $t \in \{1, 2, 3, 4, 5, 6\}$ for six tasks, s_t as the groundtruth attractiveness score row vector

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:

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

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

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

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 $\{A^{m*}\}$, $\{\omega_t^{m*}\}$.

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:

$$\begin{aligned} \min_{\{A^m\},\{\omega_t^m\}} F = & \frac{1}{2} \sum_{m=1}^{3} \|A^m X^m - Attr^m\|^2 \\ & + \frac{\lambda_1}{2} \sum_{t=1}^{6} \left\| \sum_{m=1}^{3} \omega_t^m A^m X_t^m - s_t \right\|^2 \\ & + \frac{\lambda_2}{2} \sum_{m=1}^{3} \left(\|A^m\|^2 + \sum_{t=1}^{6} \left\|\omega_t^m\right\|^2 \right). \end{aligned} \tag{2}$$

The first term is the regression from features to attributes, the second term is the regression from attributes to attractiveness scores, and the last term is the regularization term. Note that for one task t, if the modality m does not exist, then 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.

4.3 Latent Dual-Supervised Feature-Attribute-Task (LDFAT) Network

As aforementioned, we argue that manually-specified attributes can assist beauty recognition since they provide high-level semantic information that can be used to improve the characterization of attractiveness. However, the manual specification of attributes is subjective, and potentially useful (discriminative) attributes may be ignored. This may significantly affect the performance of classifiers. One way to overcome this weakness is to automatically learn attributes. We call these latent attributes, and argue that they have a complementary role in providing a more complete characterization of human attractiveness. The intuition is that attributes may be characterized by a collection of low-level

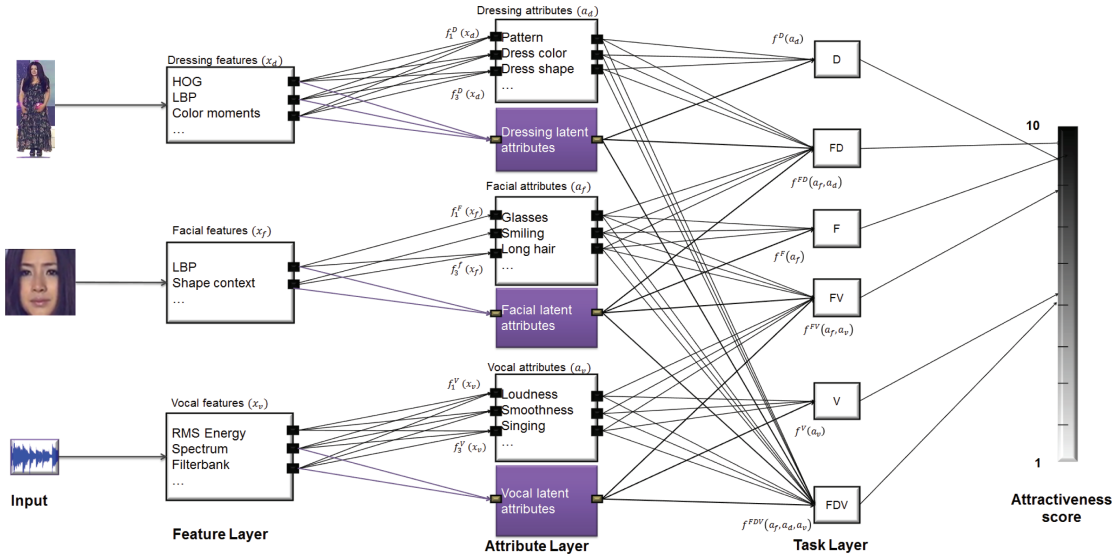


Fig. 7. Latent Dual-supervised Feature-Attribute-Task (LDFAT) Learning Framework. First, for each modality, that is, face, dress, and voice, different kinds of features are extracted. Then the beauty estimation models of single/multiple modalities are jointly learned. During the learning process, the semantic attributes are shared by different tasks. Different with traditional Neural Network, the proposed LDFAT network can seamlessly combine the training attribute labels in the learning process. Note that, without latent attributes, LDFAT falls back to DFAT. For better viewing, please see original color pdf file.

features that tend to cooccur in the training data. We propose to utilize the latent attributes inside the learning framework called Latent Dual-supervised Feature-Attribute-Task (LDFAT). The LDFAT Network is illustrated in Figure 7.

Let us denote $\hat{A}^m$ as the regression matrix from raw features to latent attributes, and $\hat{\omega}_t^m$ as the row regression vector for converting latent attributes of modality m to task t. The learning problem for LDFAT network is then formulated as:

$$\begin{aligned}\min_{\{A^m\},\{\omega_t^m\},\{\hat{A}^m\},\{\hat{\omega}_t^m\}} F = &\frac{1}{2}\sum_{m=1}^{3}\|A^m X^m - Attr^m\|^2 \\ &+ \frac{\lambda_1}{2}\sum_{t=1}^{6}\left\|\sum_{m=1}^{3}\omega_t^m A^m X_t^m + \sum_{m=1}^{3}\hat{\omega}_t^m \hat{A}^m X_t^m - s_t\right\|^2 \\ &+ \frac{\lambda_2}{2}\sum_{m=1}^{3}\left(\|A^m\|^2 + \|\hat{A}^m\|^2 + \sum_{t=1}^{6}\left(\|\omega_t^m\|^2 + \|\hat{\omega}_t^m\|^2\right)\right).\end{aligned} \tag{3}$$

The given optimization problem of LDFAT can be solved by any gradient based method and the iterative optimization procedure is listed in Algorithm 2. When $\hat{A}^m$, and $\hat{\omega}_t^m$ are initialized as zero vectors, it can be seen that Equation (3) falls back to (2). LDFAT is similar to DFAT except for the usage of latent attributes as aforementioned. Thus, we initialize $\hat{\omega}_t^m$ vectors, of all which components are 1, while $\hat{A}^m$ is initiated as zero matrix.

The computational complexities of DFAT and LDFAT lie mostly on matrix computation when updating the learnt parameters ω_t^m, A^m, ω_t^{m*} and A^{m*}. For example, the complexity of computing ω_t^m is $O(n_{att_m} n_{fea_t^m} n_{ins_t^m})$, where n_{att_m} is the number of attributes of modality m, $n_{fea_t^m}$ is the number of feature dimension of modality m and task t, and $n_{ins_t^m}$ is the number of training instance of modality m

ALGORITHM 2: Procedure to solve Problem (3)

Input: matrices X^m, X_t^m, $Attr^m$, s_t , parameter λ_1, λ_2.
Initialize: A^m by solving $\min \|A^m X^m - Att^m\|^2$, $e_1 = \infty, e_2 = 0$, $\hat{A}^m = \vec{0}$, $\hat{\omega}_t^m = \vec{1}$.
while not converged **do**

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

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

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

$$A^m = A^m - \gamma_1 \nabla F(A^m),$$

where γ_1 is the step size and $\nabla F(A^m)$ is defined as

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

3. Fix the others and update $\hat{A}^m$ by gradient descent.

$$\hat{A}^m = \hat{A}^m - \gamma_2 \nabla F(\hat{A}^m),$$

where γ_2 is the step size and $\nabla F(A^m)$ is defined as

$$\nabla F(\hat{A}^m) = \lambda_2 \hat{A}^m + \lambda_1 \sum_{t=1}^{6} \hat{\omega}_t^{mT} \left(\sum_p \omega_t^p A^p X_t^p + \sum_p \hat{\omega}_t^p \hat{A}^p X_t^p - s_t \right) X_t^{mT}.$$

4. Fix the others and update $\hat{\omega}_t^m$ by:

$$\hat{\omega}_t^m = \left(s_t - \sum_{p \in \{1,2,3\} \setminus m} \hat{\omega}_t^p \hat{A}^p X_t^p - \sum_p \omega_t^p A^p X_t^p \right) \left(\hat{A}^m X_t^m\right)^T \left(\hat{A}^m X^m \left(\hat{A}^m X_t^m\right)^T + \frac{\lambda_2}{\lambda_1} I \right)^{-1},$$

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

and task t. 10 to 20 iterations are required for convergence. It takes less than 1 second to learn all parameters on our computer equipped with quad-core 2.67GHz CPU and 8GB RAM.

5. EXPERIMENTS

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

5.1 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 8, have a score within [1, 10]. The average faces, computed by simply averaging the corresponding pixels of all the face images, present interesting patterns in attractiveness study. One of the early observations in the study of facial beauty was that averaged faces are attractive [Alley and Cunningham 1991]. Women faces similar to averaged faces have been shown to be considered more attractive. This is possibly due to average features being

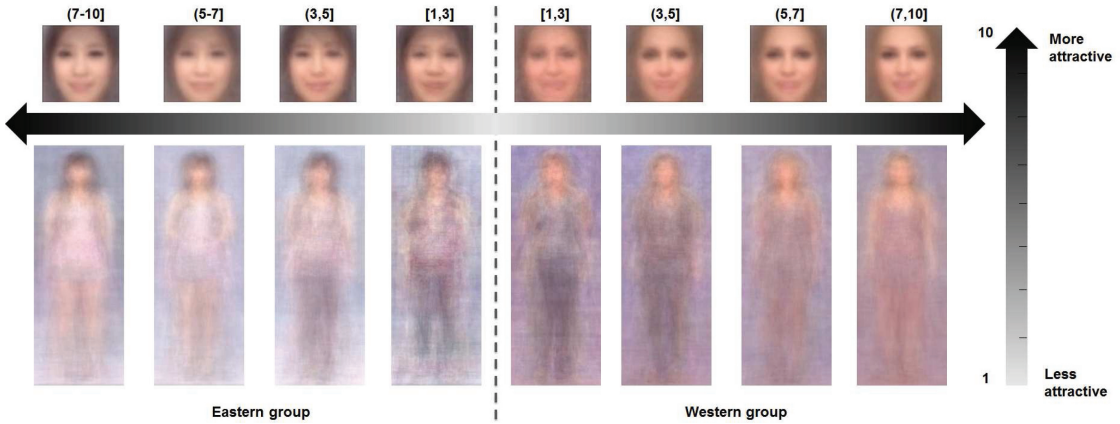
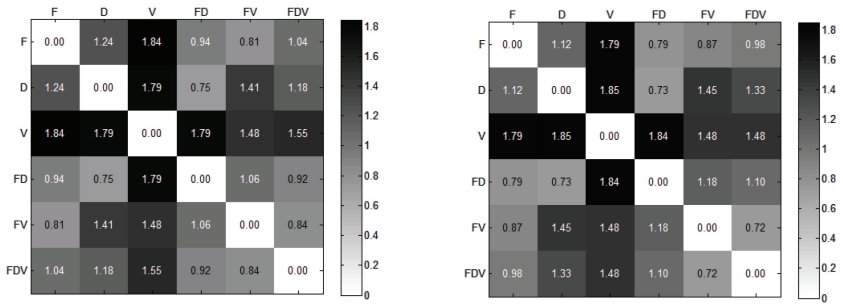


Fig. 8. Average faces and dressings of Eastern and Western groups at different attractiveness scores. For better viewing, please see original color pdf file (greatly encouraged for this figure).



(a) Eastern (b) Western

Fig. 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.

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. As observed from Figure 8, 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 forehead, 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 8 with scores within [1,10]. The less attractive dressings are trouser-like while the more attractive dressings are skirt-like. In addition, the more attractive dressings are brighter than the less attractive ones.

5.2 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

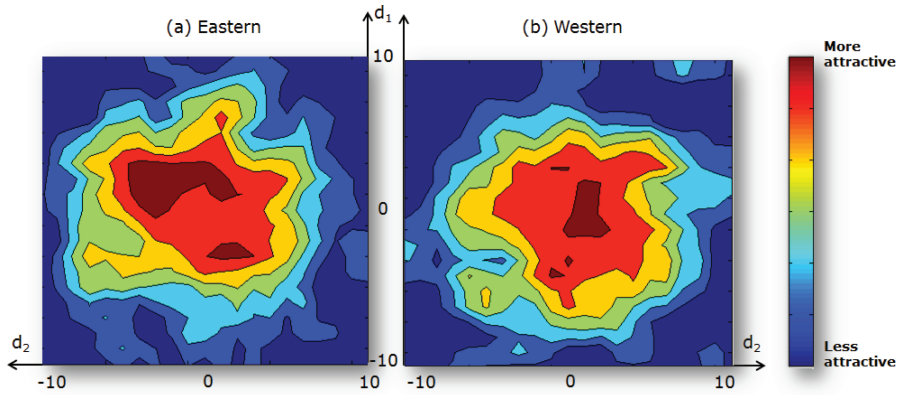


Fig. 10. The contour lines of Eastern and Western groups regarding the shape parameters and attractiveness scores.

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.

$$d_{i,j} = \frac{1}{n}\sum_{l=1}^{n}\left|s_l^i - s_l^j\right|, \tag{4}$$

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 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. Dressing 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.

5.3 Facial Shape Central Bias

In this experiment, we utilize the shape parameters as mentioned in Section 4.1.1. We only use the first two dimensions of this feature for the visualization. The location of each face is the value of its first two dimensions. The value of the face is the attractiveness score. Each dimension value is scaled to $[-10, 10]$. Then the contour fill is applied on the sparse matrix. Figure 10 illustrates the contour lines of both groups, Eastern and Western. The central bias is significantly recognized in the image. The face shape that is close to the mean shape has the higher attractiveness score. Therefore, we would like to investigate the contribution of facial shape into the attractiveness prediction.

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^2B 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

Table II. MAEs of Different Algorithms on M^{2}B 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	2.10	1.50	1.39	1.74	2.16	1.94	1.91	2.02	1.78	1.95	2.25	2.22
Ridge Regression	1.89	1.39	1.15	1.52	1.93	1.79	1.83	1.76	1.37	1.66	2.09	2.13
Neural Network	1.82	1.37	1.12	1.47	1.79	1.82	1.75	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	1.52	1.26	1.01	1.33	0.69	0.71	1.48	1.46	1.24	1.32	0.75	0.76
DFAT (w/o shape features)	1.77	–	–	1.42	0.98	1.04	1.66	–	–	1.50	1.01	1.12
LDFAT	1.46	1.14	0.96	1.18	0.67	0.67	1.46	1.37	1.14	1.28	0.71	0.74
LDFAT (w/o shape features)	1.63	–	–	1.33	0.95	0.94	1.59	–	–	1.45	0.99	1.04

prediction. The MAE is defined as the average of the absolute errors between the predicted attractiveness score and the ground truth. $MAE = \sum_{i=1}^{n} |\hat{s_i} - s_i|/n$, 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 and LDFAT 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 first 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 and LDFAT, we implement the Algorithm 1 and 2 with $\lambda_1 = 0.01$, $\lambda_2 = 10^{-3}$, $\gamma = 10^{-3}$, $\gamma_1 = 10^{-3}$, $\gamma_2 = 10^{-3}$, and $\varepsilon = 10^{-4}$ to learn the transfer matrices. We also conduct the experiment on the proposed DFAT and LDFAT without using shape features, for instance, shape context and shape parameters.

As can be seen in Table II, 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. DFAT achieves the second best results. Meanwhile, our proposed LDFAT 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 LDFAT, 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. Additionally, the usage of latent attributes also improves the performance of beauty sensing. There exist hidden attributes which were not in the predefined list contribute to the attractiveness. It is worth noting that the performance of DFAT and LDFAT decreases when shape features are not utilized. The best performance is achieved when we use the latent attributes along with the facial shape features.

Table III. MAEs of Cross-Culture Attractiveness Prediction Experiment on M^2B Dataset

Task	F	D	V	FD	FV	FDV
Train on Eastern - Test on Western	1.57	1.61	1.44	1.52	1.42	1.48
Train on Western - Test on Eastern	1.43	1.40	1.32	1.45	1.29	1.35

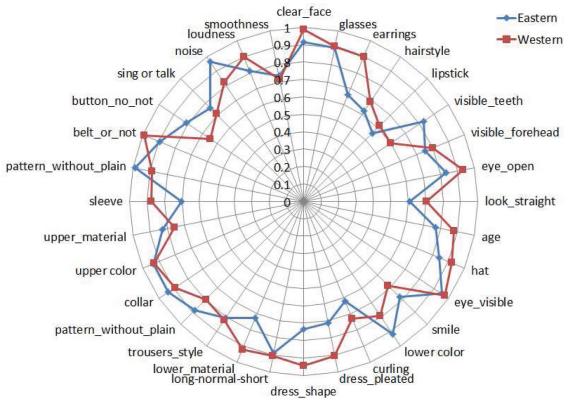


Fig. 11. The accuracy of attribute prediction.

5.5 Cross-Culture Attractiveness Prediction

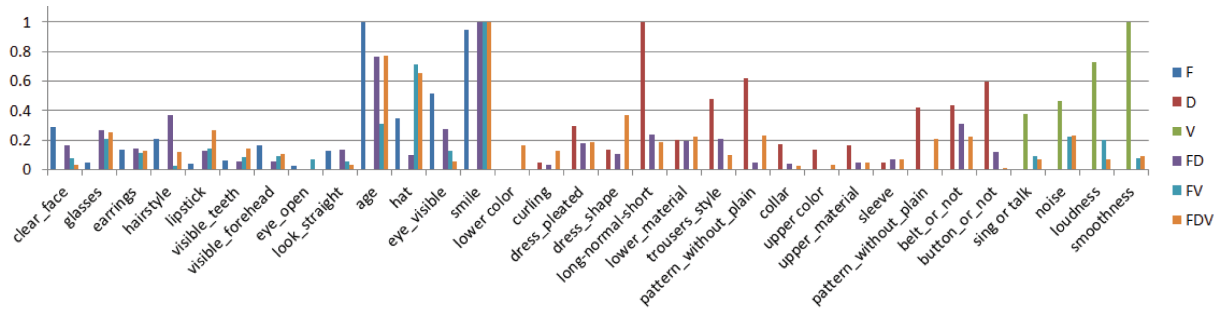
People from different cultures are often attracted by the same type of faces. This agreement among individuals of different ages and from different cultures suggests attractiveness judgements are not arbitrary but have a "biological basis". Thus, we are interested in exploring the cross-culture attractiveness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

Table III shows the MAEs of the cross-culture experiment on M^2B dataset. The MAEs of all tasks increase compared with the results of training and testing on the same ethnic group. The high error lies on dressing-related 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 of Face-Voice task is the lowest. This result agrees with the previous finding in Zuckerman and Miyake [1993] that attractive voices have the same effect as attractive faces, meaning that vocal attractiveness parallels visual attractiveness.

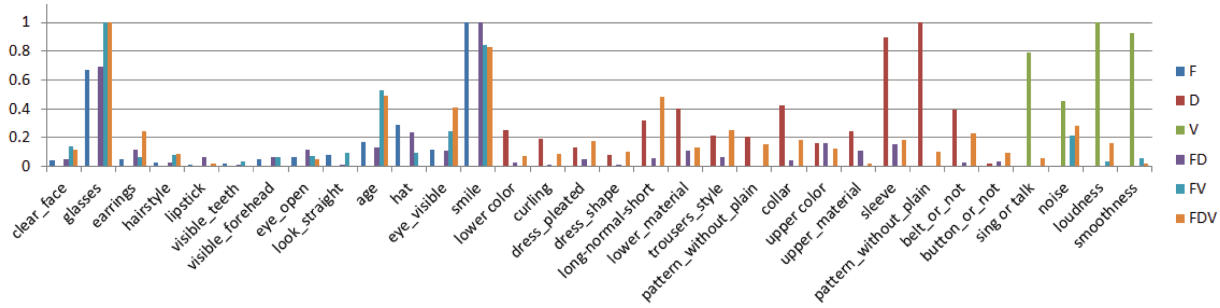
5.6 Task-specific Important Attributes

Firstly we conduct the experiment to measure the accuracy[6] of attribute prediction whose results are shown in Figure 11. Generally, the prediction results are acceptable, except for some attributes such as *lipstick* and *hairstyle*. Then we investigate attributes' importance in different tasks. We are curious to know what the model really learns. We use the absolute values of coefficients obtained from LDFAT to represent the importance of attributes to the tasks. All of the values are rescaled within [0, 1] for each task.

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



(a) Eastern



(b) Western

Fig. 12. The importance of different attributes (with latent attributes) with respect to different tasks for (a) Eastern and (b) Western. For better viewing, please see original color pdf file.

Figure 12 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 on the female attractiveness. At a closer look, the attribute *age* is extremely sensitive in Eastern group compared with Western counterpart. Meanwhile, *glasses* is well-responded in Western group. For the dressing modality, the responses of attributes are different in two ethnic groups. The skirt or pants "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*. *Color* has the small impact on the dressing attractiveness. For the voice modality, *smoothness* is the most important attribute to Eastern people. In the meantime, the loudness of voice plays the main factor for Western people to decide the voice attractiveness. Additionally, there is an "overridden" 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 altogether.

5.7 The Failure Cases

Although the algorithm performed well on the collected challenging M^2B dataset, it can fail in various circumstances as illustrated in Figure 13. If the input face is in a non-frontal pose or heavily occluded by microphone, glasses or hat, the erroneous results of the corresponding face task are achieved. Another typical failure arises when the human detector returns the inaccurate bounding boxes. Such a situation becomes more severe when the background is cluttered. Thus, it motivates us to focus on how to refine the human detection and parsing results in the future.

Fig. 13. The examples illustrate typical failure cases. (a-b) The detected faces are at non-frontal pose or heavily occluded by microphone, glasses or hat. (c-d) The wrong dressing boxes are returned by the detection part. Note that all faces are blurred to hide identities.

6. CONCLUSION AND FUTURE WORK

We have investigated the decryption of human beauty sense via multi-modality cues. To the best of our knowledge, we are the first to build female attractiveness dataset of multiple modalities: facial, dressing and voice. Its multicultural property may also be helpful for the further researches on cultures. We also proposed two trilayer learning frameworks, namely DFAT and LDFAT, to learn attributes and attractiveness simultaneously. Extensive experimental evaluations on the M^2B dataset well demonstrate the effectiveness of the proposed DFAT and LDFAT frameworks for female attractiveness prediction. The latent attribute based approach, LDFAT, outperforms the previous approach. We also show that the facial shape feature helps to improve the performance. In short, the best performance is achieved when we use the latent attributes and the facial shape features.

For future work, we would like to investigate more suitable features of different modalities for attractiveness prediction. We recognize that the cross-gender/age/ethnicity sense of attractiveness may invite future research. For example, the attractiveness of women sensed by men is different from that of women especially when we consider dressing style. We also plan to build workable real system for practical applications.

REFERENCES

AARABI, P., HUGHES, D., MOHAJER, K., AND EMAMI, M. 2001. The automatic measurement of facial beauty. In *Proceedings of the International Conference on Systems, Man and Cybernetics*, 2644–2647.

ALLEY, T. AND CUNNINGHAM, M. 1991. Average faces are attractive, but very attractive faces are not average. *Psych. Sci. 2*, 123–125.

BEAUPRE, M. 2006. An ingroup advantage for confidence in emotion recognition judgments: The moderating effect of familiarity with the expressions of outgroup members. *Personality Soc. Psych. Bull. 32*, 16–26.

BELONGIE, S., MALIK, J., AND PUZICHA, J. 2002. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 4, 509–522.

BERG, T. L., BERG, A. C., AND SHIH, J. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the European Conference on Computer Vision*. 663–676.

BOURDEV, L., MAJI, S., AND MALIK, J. 2011. Describing people: A poselet-based approach to attribute classification. In *Proceedings of the International Conference on Computer Vision*. 1543–1550.

BRINTON, D. 1890. *Races and Peoples: Lectures on the Science of Ethnography*. N.D.C. Hodges.

COOTES, T. F., TAYLOR, C. J., COOPER, D. H., AND GRAHAM, J. 1995. Active shape models-their training and application. *Computer Vision Image Understand. 61*, 1, 38–59.

DALAL, N. AND TRIGGS, B. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 886–893.

DAUGMAN, J. 1985. Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. 2*, 1160–1169.

DION, K., BERSCHEID, E., AND WALSTER, E. 1972. What is beautiful is good. *J. Appl. Soc. Psych. 24*, 90.

EISENTHAL, Y., DROR, G., AND RUPPIN, E. 2006. Facial attractiveness: Beauty and the machine. *Neural Comput. 18*, 1, 119–142.

GLASSENBERG, A., FEINBERG, D., JONES, B., LITTLE, A., AND DEBRUINE, L. 2009. Sex-dimorphic face shape preference in heterosexual and homosexual men and women. *Arch. Sexual Behav. 39*, 6, 1289–1296.

GRAY, D., YU, K., XU, W., AND GONG, Y. 2010. Predicting facial beauty without landmarks. In *Proceedings of the European Conference on Computer Vision*. 434–447.

GREEN, C. 1995. All that glitters: A review of psychological research on the aesthetics of the golden section. *Perception 24*, 937–968.

GUO, D. AND SIM, T. 2009. Digital face makeup by example. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 73–79.

HAYKIN, S. 1999. *Neural Networks*. Prentice Hall.

HUGHES, S., DISPENZA, F., AND GALLUP, G. 2004. Ratings of voice attractiveness predict sexual behavior and body configuration. *Evolution Human Behav. 25*, 5, 295–304.

JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*. 133–142.

KAGIAN, A., DROR, G., LEYVAND, T., COHEN-OR, D., AND RUPPIN, E. 2005. A humanlike predictor of facial attractiveness. *Adv. Neural Inf. Process. Sys.* 649–656.

KUMAR, N., BELHUMEUR, P. N., AND NAYAR, S. K. 2008. Facetracer: A search engine for large collections of images with faces. In *Proceedings of the European Conference on Computer Vision*. 340–353.

LARTILLOT, O. AND TOIVIAINEN, P. 2007. MIR in Matlab: A toolbox for musical feature extraction from audio. In *Proceedings of the International Society for Music Information Retrieval Conference*. 127–130.

LENNON, S. 1990. Effects of clothing attractiveness on perceptions. *Home Economics Res. J. 18*, 303–310.

LIKERT, R. 1932. A technique for the measurement of attitudes. *Arch. Psych. 22*, 140, 1–55.

LIU, S., NGUYEN, T., FENG, J., WANG, M., AND YAN, S. 2012a. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM International Conference on Multimedia*. 1333–1334.

LIU, S., SONG, Z., LIU, G., XU, C., LU, H., AND YAN, S. 2012b. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3330–3337.

MITTAL, A., ZISSERMAN, A., AND TORR, P. H. S. 2011. Hand detection using multiple proposals. In *Proceedings of the British Machine Vision Conference*. 1–11.

NGUYEN, T., LIU, S., NI, B., TAN, J., RUI, Y., AND YAN, S. 2012. Sense beauty via face, dressing and/or voice. In *Proceedings of the ACM International Conference on Multimedia*. 239–248.

OJALA, T., PIETIKÄINEN, M., AND MÄENPÄÄ, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 7, 971–987.

OLIVA, A. AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision 42*, 3, 145–175.

PALLETT, P., LINK, S., AND LEE, K. 2009. New golden ratios for facial beauty. *Vision Res. 50*, 149–154.

PARIKH, D. AND GRAUMAN, K. 2011. Relative attributes. In *Proceedings of the International Conference on Computer Vision*. 503–510.

RIESENHUBER, M. AND POGGIO, T. 1999. Hierarchical models of object recognition in cortex. *Nature Neurosci. 2*, 1019–1025.

SONG, Z., WANG, M., HUA, X.-S., AND YAN, S. 2011. Predicting occupation via human clothing and contexts. In *Proceedings of the IEEE International Conference on Computer Vision*. 1084–1091.

TANAKA, J., KIEFER, M., AND BUKACH, C. 2004. A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study. *Cognition 93*, 1–9.

VIOLA, P. AND JONES, M. 2004. Robust real-time face detection. *Int. J. Comput. Vision 57*, 2, 137–154.

YANG, Y. AND RAMANAN, D. 2011. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1385–1392.

ZUCKERMAN, M. AND MIYAKE, K. 1993. The attractive voice: What makes it so? *J. Nonverbal Behavior 17*, 119–135.

Received September 2012; revised December 2012, February 2013; accepted February 2013