"Wow! You Are So Beautiful Today!"

Luoqi Liu[1], Hui Xu[2], Junliang Xing[1], Si Liu[1], Xi Zhou[2], Shuicheng Yan[1]
[1]Department of Electrical and Computer Engineering, National University of Singapore
[2]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
{liuluoqi, elexing, dcslius, eleyans}@nus.edu.sg, {xuhui, zhouxi}@cigit.ac.cn

ABSTRACT

Beauty e-Experts, a fully automatic system for hairstyle and facial makeup recommendation and synthesis, is developed in this work. Given a user-provided frontal face image with short/bound hair and no/light makeup, the Beauty e-Experts system can not only recommend the most suitable hairdo and makeup, but also show the synthetic effects. To obtain enough knowledge for beauty modeling, we build the Beauty e-Experts Database, which contains 1,505 attractive female photos with a variety of beauty attributes and beauty-related attributes annotated. Based on this Beauty e-Experts Dataset, two problems are considered for the Beauty e-Experts system: what to recommend and how to wear, which describe a similar process of selecting hairstyle and cosmetics in our daily life. For the what-to-recommend problem, we propose a multiple tree-structured super-graphs model to explore the complex relationships among the high-level beauty attributes, mid-level beauty-related attributes and low-level image features, and then based on this model, the most compatible beauty attributes for a given facial image can be efficiently inferred. For the how-to-wear problem, an effective and efficient facial image synthesis module is designed to seamlessly synthesize the recommended hairstyle and makeup into the user facial image. Extensive experimental evaluations and analysis on testing images of various conditions well demonstrate the effectiveness of the proposed system.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Algorithms, Experimentation, Performance

Keywords

Beauty Recommendation, Beauty Synthesis, Multiple Tree-structured Super-graphs Model

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502126.

Figure 1: Overall illustration of the proposed Beauty e-Experts system. Based on the user's facial and clothing characteristics, our Beauty e-Experts system automatically recommends the suitable hairstyle and makeup products for the user, and then produces the synthesized visual effects. All the figures in this paper are best viewed in original color PDF file. Please resize $\times 2$ for better visual effects.

1. INTRODUCTION

Beauty is a language, which enables people to express their personalities, gain self-confidence and open up to others. Everybody wants to be beautiful. Hairstyle and makeup are two main factors that influence one's judgment about whether someone looks beautiful or not, especially for female. By choosing proper hair and makeup style, one can enhance the whole temperament and thus look more attractive. However, people often encounter problems when they make choices. First of all, the effects of different makeup products and hairstyles vary for different individuals, and are highly correlated with one's facial traits, *e.g.* face shape, skin color, etc. It is quite difficult, or even unlikely, for people to analyze their own facial features and make proper choices of hair and makeup care & dressing products. Second, nowadays cosmetics have developed into a large industry and there exist an unimaginable variety of products. Making choices in such a great variety can cost people a lot of time and money. Therefore, how to choose the proper hairstyle & makeup rapidly becomes a challenge.

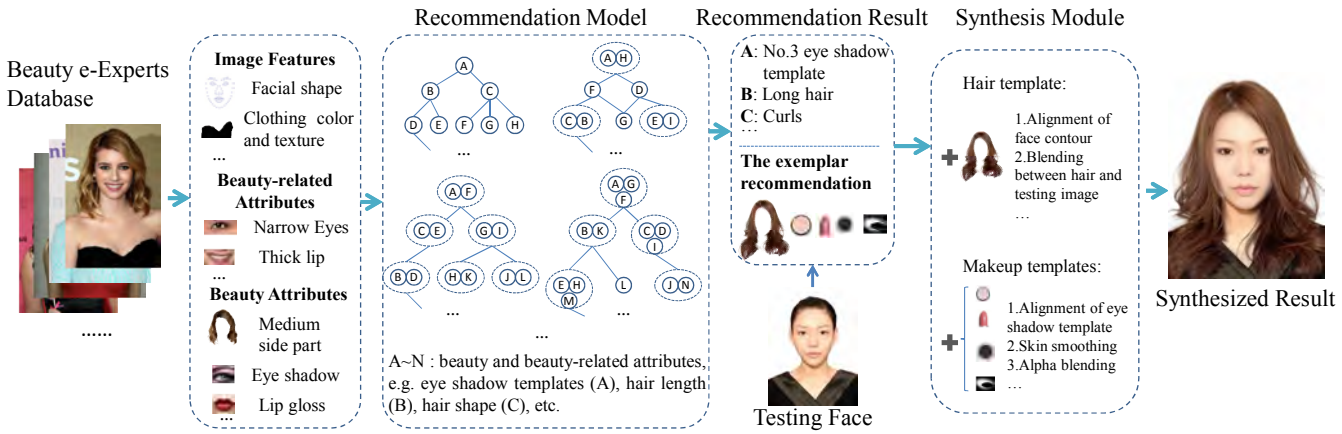

Figure 2: System processing flowchart. We firstly collect the Beauty e-Experts Database of $1,505$ facial images with different hairstyles and makeup effects. With the extracted facial and clothing features, we propose a multiple tree-structured super-graphs model to express the complex relationships among beauty and beauty-related attributes. Here, the results from multiple individual super-graphs are fused based on voting strategy. In the testing stage, the recommended hair and makeup templates for the testing face are then applied to synthesize the final visual effects.

In order to solve this problem, people have made some attempts. Some virtual hairstyle & makeup techniques have been developed. Softwares like Virtual Haircut & Makeover[1], which allow people to change hairstyle and perform virtual makeup on their photos, have been put into use. With software of this kind, users can input a facial image, and then choose any hairstyle or makeup they prefer from the options provided by the system. Users can see the effects of applying the chosen hairstyles and also makeup products on their faces, and make decisions on whether to choose these hairstyles or makeup products in reality. But there exists a problem of too much manual work for these softwares. For example, users have to mark out special regions, such as corners of eyes, nose, mouth or even pupil, etc., on their photos manually. Besides, these softwares do not have recommendation function. Users have to make choices on their own, and adjust the synthetic effects of these choices manually. It is too complicated for unprofessional people to accomplish such a process.

The research is still quite limited in this field, although some researchers have tried some approaches. Tong *et al.* [28] extracted makeup from before-and-after training image pairs, and transferred the makeup effect defined as ratios to a new testing image. Guo and Sim [15] considered makeup effect existing in two layers of the three-layer facial decomposition result, and the makeup effect of a reference image was transferred to the target image. Some patents also try to address the hair suggesting problem, *e.g.* [25], which searches for hairstyles in a database from the plurality of the hairstyle parameters based on manually selected hair attributes by the user. These works all fail to provide a recommendation function, and the synthetic effects may not be suitable for every part of the face. Besides, to our best knowledge, most of these works need a lot of user interactions and the final results are highly dependent on the efforts of users.

The aim of this work is to develop a novel Beauty e-Experts system, which helps users to select hairstyle and makeups automatically and produces the synthesized visual effects as shown in Figure 1. The main challenge in this problem is how to model the complex relationships among different beauty and beauty-related attributes for reliable recommendation and natural synthesis. To address this challenge, we build a large dataset, Beauty e-Experts Dataset, which contains $1,505$ images of beautiful female figures selected from professional fashion websites. Based on this Beauty e-Experts Dataset, we first annotate all the beauty attributes for each image in the whole dataset. These beauty attributes, including different hairstyles and makeup types, are all adjustable in the daily life. Their specific combination is considered as the recommendation objective of the Beauty e-Experts System. To narrow the gap between the high-level beauty attributes and the low-level image features, a set of mid-level beauty-related attributes, such as facial traits and clothing properties, are also annotated for the dataset.

Based on all these attributes, we propose to learn a multiple tree-structured super-graphs model to explore the complex relationships among these attributes. As a generalization of a graph, a super-graph can theoretically characterize any type of relationships among different attributes and thus provide powerful recommendation. We propose to use its multiple tree-structured approximations to reserve the most important relationships and make the inference procedure tractable. Based on the recommended results, an effective and efficient facial image synthesis module is designed to seamlessly synthesize the recommended results into the user facial image and show it back to the user. Experimental results on 100 testing images show that our system can obtain very reasonable recommendation results and appealing synthesis results. The whole system processing flowchart is illustrated in Figure 2.

The contributions of this work are summarized as follows:

- A comprehensive system considering both hairstyle and makeup recommendation and synthesis is explored for the first time.
- A large database called Beauty e-Experts Database is constructed, including $1,505$ facial images with various hairstyles and makeup effects, and fully annotated with different types of attributes.
- A multiple tree-structured super-graphs model is proposed for hairstyle and makeup recommendation.

[1]http://www.goodhousekeeping.com/beauty

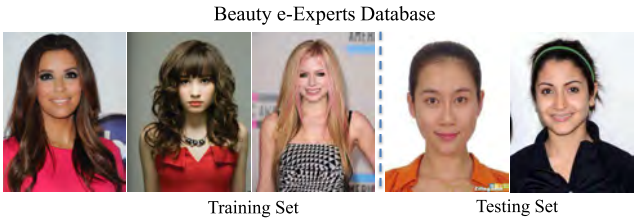

Figure 3: Some exemplar images from the Beauty e-Experts Dataset and the additional testing set. The left three images are from the Beauty e-Experts Dataset used for training, while the right two images are from the testing set.

2. DATASET, ATTRIBUTES AND FEATURES

Hairstyle & makeup products make a lucrative market among female customers, but no public datasets for academic research exist. Most previous research [27, 15, 28] only works for a few samples. Chen and Zhang [6] released a benchmark for facial beauty study, but their focus is geometric facial beauty, not facial makeup and hairstyle. Wang and Ai [29] sampled 1,021 images with regular hairstyles from Labeled Faces in the Wild (LFW) Database [18], which is not designed for hairstyle analysis. In addition, the sampled LFW database contains only a few hairstyles, since it is designed only for hair segmentation. In order to obtain enough knowledge to perform beauty modeling, we need a large dataset specifically designed for this task. In the following, we will describe the construction of the Beauty e-Experts Dataset, as well as its attribute annotation and feature extraction process.

2.1 The Beauty e-Experts Dataset

To build our Beauty e-Experts Dataset, we have downloaded $\sim 800K$ images of female figures from professional hairstyle and makeup websites (e.g. www.stylebistro.com). We search these photos with the key words like *makeup, cosmetics, hairstyle, haircut and celebrity*. The initial downloaded images are screened by a commercial face analyzer[2] to remove images with no face detected, and then 87 keypoints are located for each image. The images with high resolution and confident landmark detection results are retained. The retained images are further manually selected, and only those containing female figures who are considered as attractive and with complete hairstyle and obvious makeup effects are retained. The final 1,505 retained images contain female figures in distinct fashions and with clear frontal faces, and are thus very good representatives for beauty modeling. Besides, we also collect 100 face images with short/bound hair and no/light makeup, which better demonstrate the synthesized results, for experimental evaluation purpose. Figure 3 shows some exemplar images in the Dataset.

2.2 Attributes and Features

To obtain beauty knowledge from the built dataset, we comprehensively explore different beauty attributes on these images, including various kinds of hairstyles and facial makeups. All these beauty attributes can be easily adjusted and modified in the daily life and thus have practical meaning for our beauty recommendation and synthesis system. We

[2]OMRON OKAO Vision: `http://www.omron.com/r_d/coretech/vision/okao.html`

Table 1: A list of the high-level beauty attributes.

Name	Values
hair length	long, medium, short
hair shape	straight, curled, wavy
hair bangs	full, slanting, center part, side part
hair volume	dense, normal
hair color	20 classes
foundation	15 classes
lip gloss	15 classes
eye shadow color	15 classes
eye shadow template	20 classes

carefully organize these beauty attributes and set their attribute values based on some basic observations or preprocessing on the whole dataset. Table 1 lists the names and values of all the beauty attributes considered in the work. For the first four beauty attributes in Table 1, their values are set intuitively, and for the last five ones, their values are obtained by running the k-means clustering algorithm on the training dataset for the corresponding features (the cluster number is determined empirically according to each specific attribute). The pixel values within the specific facial regions on each training image are clustered by Gaussian mixture models (GMM) in RGB color space. The centers of the largest GMM components are used as the representative colors. Then the colors are clustered by k-means to obtain the color attributes of hair and makeup templates. Hair templates and eye shadows are extracted by image matting [22]. For eye shadows, only the alpha channel is retained and further clustered to learn the eye shadow template attribute. The left eye shadows are sufficient for clustering. We show the visual examples of specific attribute values for some beauty attributes in Figure 4.

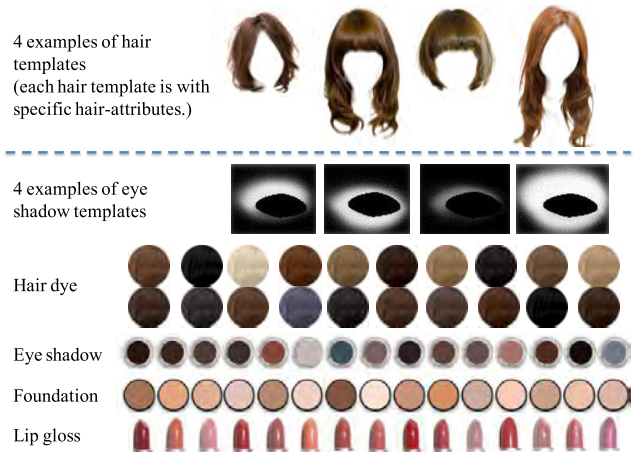

Figure 4: Visual examples of the specific values for some beauty attributes.

A straightforward way to predict the values of these high-level beauty attributes is using some low-level features to train some classifiers. However, since there is a relatively big gap between the high-level beauty attributes and the low-level image features, and the beauty attributes are intuitively related to some mid-level attributes like eye shape and mouth width, we further explore a set of mid-level beauty-related attributes to narrow the gap between the high-level beauty attributes and the low-level image features. Table 2 lists all the mid-level beauty-related attributes annotated for the dataset. These mid-level attributes mainly focus on the

Table 2: A list of mid-level beauty-related attributes considered in this work.

Names	Values
forehead	high, normal, low
eyebrow	thick, thin
eyebrow length	long, short
eye corner	upcurved, downcurved, normal
eye shape	narrow, normal
ocular distance	hypertelorism, normal, hypotelorism
cheek bone	high, normal
nose bridge	prominent, flat
nose tip	wide, narrow
mouth opened	yes, no
mouth width	wide, normal
smiling	smiling, neutral
lip thickness	thick, normal
fatness	fat, normal
jaw shape	round, flat, pointed
face shape	long, oval, round
collar shape	strapless, v-shape, one-shoulder, high-necked, round, shirt collar
clothing pattern	vertical, plaid, horizontal, drawing, plain, floral print
clothing material	cotton, chiffon, silk, woolen, denim, leather, lace
clothing color	red, orange, brown, purple, yellow, green, gray, black, blue, white, pink, multi-color
race	Asian, Western

facial shapes and clothing properties, which are kept fixed during the recommendation and the synthesis process.[3]

After the annotation of the high-level beauty attributes and mid-level beauty-related attributes, we further extract various types of the low-level image features on the clothing and facial regions for each image in the Beauty e-Experts Dataset to facilitate further beauty modeling. The clothing region of an image is automatically determined based on its geometrical relationship with the face region. Specifically, the following features are extracted for image representation:

- RGB color histogram and color moments on clothing region.
- Histograms of oriented gradients (HOG) [10] and local binary patterns (LBP) [1] features on clothing region.
- Active shape model [8] based shape parameters.
- Shape context [2] features extracted at facial points.

All the above features are concatenated to form a feature vector of $7,109$ dimensions, and then Principal Component Analysis (PCA) [20] is performed to reserve 90% of the energy. The compacted feature vector with 173 dimensions and the annotated attribute values are then fed into an SVM classifier to train for each attribute a classifier.

3. THE RECOMMENDATION MODEL

A training beauty image is denoted as a tuple $(\langle \mathbf{x}, \mathbf{a}^r \rangle, \mathbf{a}^b)$. Here $\mathbf{x}$ is the image features extracted from the raw image data; $\mathbf{a}^r$ and $\mathbf{a}^b$ denote the set of the beauty-related attributes and beauty attributes respectively. Each attribute may have multiple different values, *i.e.* $a_i \in \{1, \ldots, n_i\}$, where n_i is the number of attribute values for the i-th attribute. The beauty-related attributes $\mathbf{a}^r$ act as the mid-level cues to narrow the gap between the low-level image features $\mathbf{x}$ and the high-level beauty attributes $\mathbf{a}^b$. The recommendation model needs to uncover the complex relationships among the low-level image features, mid-level beauty-related attributes and high-level beauty attributes, and make the final recommendation for a given image.

[3]Although the clothes of a user can be changed to make one look more beautiful, they are kept fixed in our current Beauty e-Experts system.

3.1 Model Formulation

We model the relationships among the low-level image features, the mid-level beauty-related attributes, and the high-level beauty attributes from a probabilistic perspective. The aim of the recommendation system is to estimate the probability of beauty attributes, together with beauty-related attributes, given the image features, *i.e.* $p\left(\mathbf{a}^b, \mathbf{a}^r | \mathbf{x}\right)$, which can be modeled using the Gibbs distribution,

$$p\left(\mathbf{a}^b, \mathbf{a}^r | \mathbf{x}\right) = \frac{1}{Z\left(\mathbf{x}\right)} \exp\left(-E\left(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x}\right)\right), \qquad (1)$$

where $Z\left(\mathbf{x}\right) = \sum_{\mathbf{a}^b, \mathbf{a}^r} \exp\left(-E\left(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x}\right)\right)$, also known as the partition function, is a normalizing term dependent on the image features, and $E(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x})$ is an energy function measuring the compatibility among the beauty attributes, beauty-related attributes and image features. The beauty recommendation results can be obtained by finding the most likely joint beauty attribute state $\hat{\mathbf{a}}^b = \arg\max_{\mathbf{a}^b} \max_{\mathbf{a}^r} p\left(\mathbf{a}^b, \mathbf{a}^r | \mathbf{x}\right)$.

The capacity of this probabilistic model fully depends on the structure of the energy function $E(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x})$. Here we propose to learn a general super-graph structure to build the energy function which can theoretically be used to model any relationships among the low-level image features, mid-level beauty-related attributes, and high-level beauty attributes. To begin with, we give the definition of super-graph.

Definition 1. *Super-graph: a super-graph $\mathcal{G}$ is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is called super-vertexes, consisting a set of non-empty subsets of a basic node set, and $\mathcal{E}$ is called super-edges, consisting of a set of two-tuples, each of which contains two different elements in $\mathcal{V}$.*

It can be seen that a super-graph is actually a generalization of a graph in which a vertex can have multiple basic nodes and an edge can connect any number of basic nodes. When all the super-vertexes only contain one basic node, and each super-edge is forced to connect to only two basic nodes, the super-graph then becomes a traditional graph. A super-graph can be naturally used to model the complex relationships among multiple factors, where the factors are denoted by the vertexes and the relationships are represented by the super-edges.

Definition 2. *k-order super-graph: for a super-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, if the maximal number of vertexes involved by one super-edge in $\mathcal{E}$ is k, $\mathcal{G}$ is said to be a k-order super-graph.*

Based on the above definitions, we propose to use the super-graph to characterize the complex relationships among the low-level image features, mid-level beauty-related attributes, and high-level beauty attributes in our problem. For example, the aforementioned pairwise correlations can be sufficiently represented by a 2-order super-graph (traditional graph), while other more complex relationships, such as one-to-two and two-to-two relationships, can be represented by other higher order super-graphs. Denote the basic node set A as the union of the beauty attributes and beauty-related attributes, *i.e.* $A = \{a_i | a_i \in \mathbf{a}^r \cup \mathbf{a}^b\}$. Supposing the underlying relationships among all the attributes are repre-

sented by a super-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{\mathbf{a}_i | \mathbf{a}_i \subset A\}$[4] is a set of non-empty subsets of A, and $\mathcal{E}$ is the super-edge set that models their relationships, the energy function can then be defined as,

$$E\left(\mathbf{a}^b, \mathbf{a}^r, \mathbf{x}\right) = \sum_{\mathbf{a}_i \in \mathcal{V}} \phi_i\left(\mathbf{a}_i, \mathbf{x}\right) + \sum_{(\mathbf{a}_i, \mathbf{a}_j) \in \mathcal{E}} \phi_{ij}\left(\mathbf{a}_i, \mathbf{a}_j\right). \quad (2)$$

The first summation term is called FA (feature to attribute) potential which is used to model the relationships between the attributes and low-level image features, and the second one is called AA (attribute to attribute) potential and is used to model the complex relationships among different attributes represented by the super-edges. $\phi_i\left(\mathbf{a}_i, \mathbf{x}\right)$ and $\phi_{ij}\left(\mathbf{a}_i, \mathbf{a}_j\right)$ are the potential functions of the corresponding inputs, which can be learned in different ways. Generally, the FA potential $\phi_i\left(\mathbf{a}_i, \mathbf{x}\right)$ is usually modeled as a generalized linear function in the form like

$$\phi_i\left(\mathbf{a}_i = \mathbf{s}_i, \mathbf{x}\right) = \psi_{\mathbf{a}_i}\left(\mathbf{x}\right)^\top \mathbf{w}_i^{\mathbf{s}_i}, \quad (3)$$

where $\mathbf{s}_i$ is the values for attribute subset $\mathbf{a}_i$, $\psi_{\mathbf{a}_i}\left(\mathbf{x}\right)$ is a set of feature mapping functions for the attributes in $\mathbf{a}_i$ by using SVM on the extracted features (see Section 2.2), and $\mathbf{w}_i$ is the FA weight parameters to be learned for the model. And the AA potential function $\phi_i\left(\mathbf{a}_i, \mathbf{a}_j\right)$ is defined by a scalar parameter for each joint state of the corresponding super-edge,

$$\phi_{ij}\left(\mathbf{a}_i = \mathbf{s}_i, \mathbf{a}_j = \mathbf{s}_j\right) = w_{i,j}^{\mathbf{s}_i \mathbf{s}_j}, \quad (4)$$

where $w_{i,j}^{\mathbf{s}_i \mathbf{s}_j}$ is a scalar parameter for the corresponding joint state of $\mathbf{a}_i$ and $\mathbf{a}_j$ with the specific value $\mathbf{s}_i$ and $\mathbf{s}_j$.

3.2 Model Learning

The learning of the super-graph based energy function includes learning the structure of the underlying super-graph and the parameters in the potential functions.

3.2.1 Structure Learning

Learning a fully connected super-graph structure is generally an NP-complete problem, which makes finding the optimal solution technically intractable [21]. However, we can still find many good approximations which can model a very large proportion of all the possible relationships. Among all the possible approximations, tree structure provides a very good choice which can be learned efficiently using many algorithms [21]. Another merit of tree structure is that the inference on a tree can be efficiently performed using methods like dynamic programming. Based on these considerations, we therefore use the tree-structured super-graph to model the underlying relationships. To remedy the information loss during the approximation procedure, we further propose to simultaneously learn multiple different tree-structured super-graphs to collaboratively model the objective relationships. Learning multiple tree-structured super-graphs is also supposed to produce more useful recommendation results, since it is intuitively similar to daily recommendation scenario. These tree-structured super-graphs can be supposed to be different recommendation experts, each of which is good at modeling some kinds of relationships. The recommendation results generated by these experts are voted to form the final recommendation result.

For a 2-order super-graph, learning a tree-structured approximation can be efficiently solved using the maximum spanning tree algorithm [7]. The edge weights in the graph are given by the mutual information between the attributes, which can be estimated from the empirical distribution from the annotations in the training data. For higher order super-graph, however, learning its tree-structured approximation will not be a trivial task, since the choices of vertex subsets for each super-edge are combinational.

[4]Note that in this paper we use $\mathbf{a}_i$ to denote a non-empty attribute set and a_i to denote a single attribute.

Algorithm 1 Candidate set of subsets generation for super-graph structure learning.

Input: basic node set $A = \{a_1, \ldots, a_M\}$, adjacency matrix $W = \{w_{ij}\}_{1 \le i,j \le M}$, desired order of the super-graph k.
Output: candidate set of subsets $\mathcal{V} = \{\mathbf{a}_i | \mathbf{a}_i \in A\}$.

```
Initialization: set V with m unique subsets randomly collected from A, each of which has no more than ⌊(k+1)/2⌋ elements. Set w_max = f(V, W).
while not converged do
  for i = 1 → M do
    for j = 1 → m do
      w_j = f(V,W) if move a_i to a_j.
    end for
    w_l = argmax_j({w_j}).
    if l > w_max then
      Move a_i to a_l, w_max = w_l.
      if |a_l| > ⌊(k+1)/2⌋ then
        Split |a_l| into two subsets.
        m ← m + 1.
      end if
      if m > ⌈2 × M/(k−1)⌉ then
        Merge two smallest subsets.
        m ← m − 1.
      end if
    end if
  end for
end while
Generate candidate vertex subsets.
```

Suppose for a super-graph built on basic node set $A = \{a_1, \ldots, a_M\}$ with M elements, we need to find a k-order tree-structured super-graph for these vertexes. We first calculate the mutual information between each pair of vertexes, and denote the results in the adjacency matrix form, i.e. $W = \{w_{ij}\}_{1 \le i,j \le M}$. Then we propose a two-stage algorithm to find the k-order tree-structured super-graph.

In the first stage, we aim to find the candidate set of basic node subsets $\mathcal{V} = \{\mathbf{a}_i | \mathbf{a}_i \in \mathcal{V}\}$, which will be used to form the super-edges. The objective here is to find the set of subsets that has the largest amount of total mutual information in the result k-order super-graph. Here we first define a function that calculates the mutual information of a subset set with a specified mutual information matrix,

$$f(\mathcal{V}, W) = \sum_{|\mathbf{a}_i| \ge 2} \sum_{a_j, a_k \in \mathbf{a}_i} w_{jk}. \quad (5)$$

Based on this definition, we formulate the candidate set generation problem as the following optimization problem

$$\begin{aligned} \underset{\mathcal{V}}{\text{argmax}} \quad & f(\mathcal{V}, W), \\ \text{s.t. } & |\mathbf{a}_i| \le \lfloor \frac{k+1}{2} \rfloor, \forall i, \\ & |\mathcal{V}| \le m, \end{aligned} \quad (6)$$

where the first inequation is from the k-order constraint from the result super-graph, $\lfloor \cdot \rfloor$ is the floor operator, and the

Algorithm 2 Learning multiple tree-structured super-graphs.

Input: basic node set $A = \{a_1, \ldots, a_M\}$, adjacency matrix $W = \{w_{ij}\}_{1 \le i,j \le M}$, number of desired super-graphs T.
Output: T tree-structured super-graphs $\mathbf{G} = \{\mathcal{G}_t\}_{t=1}^{T}$.

1: **Initialization**: set $\mathbf{G} = \emptyset$, $K = 5$.
2: **for** $t = 1 \rightarrow T$ **do**
3: Generate a random variable $k \in \{2, \ldots K\}$.
4: Obtain a candidate vertex subsets $\mathcal{V}$ using Alg. 1.
5: Calculate the mutual information between the elements pair with no more than k vertexes in $\mathcal{V}$.
6: Make a graph using the calculated mutual information as adjacency matrix.
7: Find its maximal spanning tree using the algorithm in [7].
8: Form the k-order tree-structured super-graph $\mathcal{G}_t$.
9: $\mathbf{G} \leftarrow \mathbf{G} \cup \{\mathcal{G}_t\}$.
10: **end for**
11: Generate tree-structured super-graph set $\mathbf{G}$.

parameter m in the second inequation is used to ensure that the generated subsets have a reasonable size to cover all the vertexes and make the inference on the result super-graph more efficient. Specifically, its value can be set as

$$m = \begin{cases} M, & k = 2, \\ 2\lceil M/(k-1) \rceil, & \text{otherwise}, \end{cases} \tag{7}$$

where $\lceil \cdot \rceil$ is the ceil operator. To solve this optimization problem, we design a k-means like iterative optimization algorithm to find the solution. The algorithm first initializes some random vertex subsets and then re-assigns each vertex to the subsets that result in maximal mutual information increment; if one vertex subset has more than $\lfloor (k+1)/2 \rfloor$ elements, it will be split into two subsets; if the total cardinality of the vertex subset set is larger than $2\lceil M/(k-1) \rceil$, two subsets with the smallest cardinalities will be merged into one subset. This procedure is repeated until converge. Alg. 1 gives the pseudo-code description of this procedure.

Based on the candidate vertex subsets, the second stage of the two-stage algorithm first calculates the mutual information between the element pair that satisfies the order restrictions in the each vertex subset. Then it builds a graph by using the calculated mutual information as adjacency matrix, and the maximum spanning tree algorithm [7] is adopted to find its tree-structured approximation.

The above two-stage algorithm is run many times by setting different k values and initializations of subsets, which then generates multiple tree-structured super-graphs with different orders and structures. In order to make the parameters learning in the following tractable, the maximal k-value K is set to be equal to 5. The detailed description of this process is summarized in Alg. 2.

3.2.2 Parameter Learning And Inference

For each particular tree-structured super-graph, its parameter set, including the parameters in the FA potentials and the AA potentials, can be denoted in a whole as $\mathbf{\Theta} = \{\mathbf{w}_i^{\mathbf{s}_i}, w_{ij}^{\mathbf{s}_i\mathbf{s}_j}\}$. We adopt the maximal likelihood criterion to learn these parameters. Given N i.i.d. training samples $\mathbf{X} = \{\langle \mathbf{x}_n, \mathbf{a}_n^r \rangle, \mathbf{a}_n^b\}$, we need to minimize the loss function

$$\begin{aligned} \mathcal{L} &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n + \frac{1}{2}\lambda\sum_{i,\mathbf{s}_i}\|\mathbf{w}_i^{\mathbf{s}_i}\|_2^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}\left\{-\ln p\left(\mathbf{a}_n^b, \mathbf{a}_n^r|\mathbf{x}_n\right)\right\} + \frac{1}{2}\lambda\sum_{i,\mathbf{s}_i}\|\mathbf{w}_i^{\mathbf{s}_i}\|_2^2, \end{aligned} \tag{8}$$

where λ is the tradeoff parameter between the regularization term and log-likelihood and its value is chosen by k-fold validation on the training set. Since the energy function is linear with respect to the parameters, the log-likelihood function is concave and the parameters can be optimized using gradient based methods. The gradient of the parameters can be computed by calculating their marginal distributions [24]. Denoting the value of attribute $\mathbf{a}_i$ for training image n as $\hat{\mathbf{a}}_i$, we have

$$\frac{\partial \mathcal{L}_n}{\partial \mathbf{w}_i^{\mathbf{s}_i}} = \left([\hat{\mathbf{a}}_i = \mathbf{s}_i] - p\left(\mathbf{a}_i = \mathbf{s}_i|\mathbf{x}_n\right)\right)\psi_{\mathbf{a}_i}\left(\mathbf{x}_n\right), \tag{9}$$

$$\frac{\partial \mathcal{L}_n}{\partial w_{ij}^{\mathbf{s}_i\mathbf{s}_j}} = [\hat{\mathbf{a}}_i = \mathbf{s}_i, \hat{\mathbf{a}}_j = \mathbf{s}_j] - p\left(\mathbf{a}_i = \mathbf{s}_i, \mathbf{a}_j = \mathbf{s}_j|\mathbf{x}_n\right), \tag{10}$$

where $[\cdot]$ is the Iverson bracket notation, i.e., $[\cdot]$ equals 1 if the experssion is true, and 0 otherwise.

Based on the calculation of the gradients, the parameters can be learned from different gradient based optimization algorithms [21]. In the experiments, we use the implementation by Schmidt[5] to learn these parameters. The learned parameters, together with the corresponding super-graph structures, form the final recommendation model.

Here each learned tree-structured super-graph model can be seen as a beauty expert. Given an input testing image, the system first extracts the feature vector $\mathbf{x}$, and then each beauty expert makes its recommendation by performing inference on the tree structure to find the maximum posteriori probability of $p\left(\mathbf{a}^b, \mathbf{a}^r|\mathbf{x}\right)$. The recommendation results output by all the Beauty e-Experts are then fused by majority voting to make the final recommendation to the user.

3.3 Relations with Other Models

The proposed multiple tree-structured super-graphs model characterizes the complex relationships among different attributes from a probabilistic prospective. When the maximal order value K is set to 2, our model degenerates into the classical graphical model used by most previous works [4, 29, 24], where only the one-to-one pairwise correlations between two attributes are considered to model the complex relationships. Our model can generally model any order of the relationships. When the maximal order value K of the super-graph is set to 5, many other types of relationships, *e.g.*, one-to-two, two-to-two, and two-to-three, can be simultaneously modeled.

The pairwise correlations are also extensively modeled using the latent SVM model [30] from a deterministic perspective, which has been successfully applied into the problem like object detection [12], pose estimation [31], image classification [30], as well as clothing recommendation [23]. Compared with the latent SVM model, our tree-structured super-graph model not only can consider much more complex relationships among the attributes, but also is more efficient since tree structure makes both learning and inference process much faster. For a tree structured model with n nodes and k different values for each node, the time complexity of

[5] `http://www.di.ens.fr/~mschmidt/Software/UGM.html`

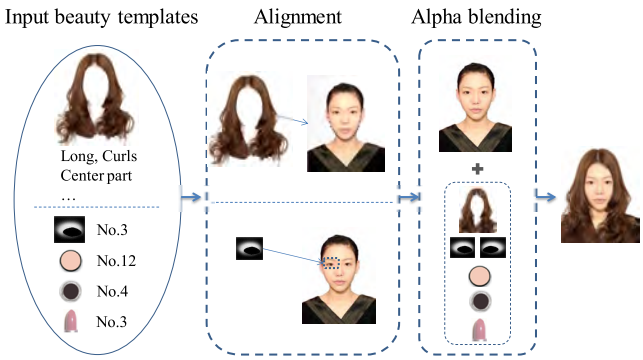

Figure 5: The flowchat of the synthesis module.

the inference process is only $O(k^2n)$, while a fully connected model (*e.g.* latent SVM) has the complexity of $O(k^n)$. Actually, during the training of the latent SVM model, some intuitively small correlations have to be removed manually to accelerate the training process. Our tree-structured super-graphs model, on the contrary, can automatically remove the small relationships during the structure learning process in a principled way. By extending to multiple tree-structured super-graphs, our model can produce much more reliable and useful recommendations as verified in experimental part, since it can well simulate the common recommendation scenario in our daily life, where one usually asks many people for recommendations and takes the majority or the most suitable one as the final choice.

4. THE SYNTHESIS MODULE

With the beauty attributes recommended by the multiple tree-structured super-graphs model, we further synthesize the final visual effect of hairstyle and makeup for the testing image. To this end, the system first uses beauty attributes to search for hair and makeup templates. A hair template is a combination of hairstyle attributes, such as long curls with bangs. We use the recommended hairstyle attributes to search the Beauty Expert Database for suitable hair templates. As mentioned in Section 2.2, each hair template is extracted from a training image. If more than one template is obtained, we randomly select one from them. If we cannot find the hair template with exactly the same hairstyle attribute values, we use the one which has the values most approximating to the recommended hairstyle attribute values. Each makeup attribute forms a template which can be directly obtained from the dataset. These obtained hair and makeup templates are then fed into the synthesis process, which mainly has two steps: alignment and alpha blending, as shown in Figure 5.

In the alignment step, both of the hairstyle and the makeup templates need to be aligned with the testing image. For hair template alignment, a dual linear transformation procedure is proposed to put the hair template on the target face in the testing image. The dual linear transformation process first uses a linear affine transformation to perform rough alignment and then adopts a piecewise-linear affine transformation [14] to perform precise alignment. Figure 6 gives an illustration of this process. In the linear affine transformation, the 21 face contour points generated by the face analyzer are adopted to calculate affine transformation matrix between the hair template and the testing face. The hair template then can be roughly aligned to the testing

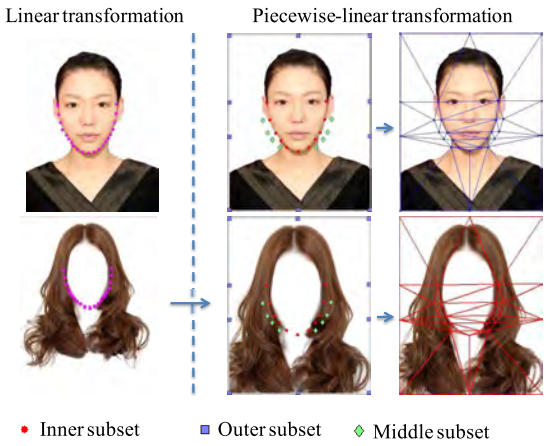

Figure 6: Hair template alignment process.

face using transformation matrix. In piecewise linear affine transformation, three subsets of keypoints, namely, the inner subset, the middle subset, and the outer subset, are sampled based on the result of rough alignment. In Figure 6, these keypoints in the three subsets are drawn in red, green and blue respectively. 11 points are sampled interlacedly from 21 face contour points to consist the inner subset. The points in the inner subset are extended on the horizontal direction with half of the distance between two eye centers. They form the middle subset of 8 points. 10 points in the outer subset are fixed on the edge of the image. Their coordinates are determined by the image corners or the horizontal lines of eye centers and mouth corners. Note that points of the middle and outer subsets are at the same position in both the hair template and the testing image, which aims to keep the overall shape of the hairstyle. The total 29 points in the three subsets are then used to construct a Delaunay triangulation [11] to obtain 46 triangles. Then affine transformations are applied within the corresponding triangles between the testing face and the hair template. After that, these points on the hair template are precisely aligned with the testing face.

For the makeup templates alignment, only the eye shadow template need to be aligned to the eye region in the testing image. Other makeup templates can be directly applied to the corresponding regions based on the face keypoint detection results. To align the eye shadow template to eye contour on the face, we use the thin plate spline method [3] to warp the eye shadow template by using the eye contour points. Because the eye shadow template attributes are learned by clustering from the left eye, the left template is mirrored to the right to obtain the right eye shadow template.

In the alpha blending step, the final result R is synthesized with hair template, makeup and the testing face I. The synthesis process is performed in CIELAB color space. L^* channel is considered as lightness because of its similarity to human visual perception. a^* and b^* are the color channels. We firstly use the edge-preserving operator on image lightness channel L^* to imitate the smoothing effect of foundation. We choose the guided filter [17], which is more efficient and has better performance near the edges among all the edge-preserving filters. It is applied to the L^* channel of facial region determined by facial contour points. Note that since we do not have contour points on the forehead, the forehead region is segmented out by GrabCut [26]. The final synthesis result is generated by alpha blending of the

testing image I and hair and makeup template T in the L^*, a^* and b^* channels, respectively,

$$R = \alpha I + (1 - \alpha)T, \tag{11}$$

where α is a weight to balance I and T. For hair and eye shadow templates, the value of α is obtained from the templates themselves. For foundation and lip gloss, the α value is set to 0.5 for L^* channel, and 0.6 for a^* and b^* channels.

5. EXPERIMENTS

In this section, we design experiments to evaluate the performance of the proposed Beauty e-Experts System from different aspects. We first visualize and analyze the intermediate result of model learning processing. Then the recommendation result is evaluated by comparison with several baselines, such as latent SVM [30], multi-class SVM [5] and neural network [16]. The synthesis effects are finally presented and compared with some commercial systems related to hairstyle and makeup recommendation and synthesis.

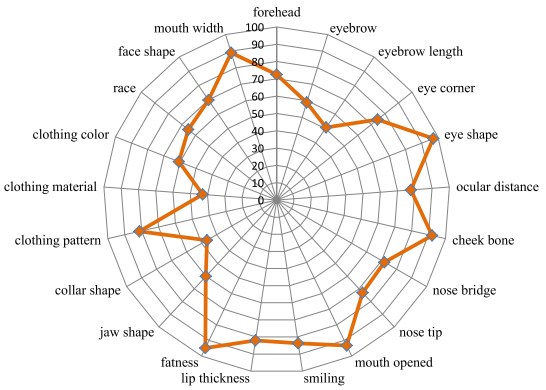

Figure 7: Accuracies of the predicted beauty attributes from the SVM classifier.

5.1 Model Learning and Analysis

We look into some intermediate aspects for a deep insight of the recommendation model structure and learning process. In Figure 7, we present the accuracy of predicted beauty-related attributes from the SVM classifier, which is used to build the FA potential function in the energy function to model the recommendation distribution (see Eqn. (2)). Note that we do not present the accuracy of beauty attributes, for we cannot obtain the ground truth label in this stage. It can be seen that most classifiers have the accuracies of more than 70%. It is sufficient to provide enough information to predict beauty attributes. Clothing-related attributes have the accuracy of 40% $\sim$ 50%, which is a little bit low. This is mainly caused by the large number of categories of clothing-related attributes.

We also visualize one example of the learned tree structure in the recommendation model in Figure 8. This tree structure is of order 4 and each super-vertex can only include two attributes at most (see Alg. 1). The weight of super-edge represents the mutual information between two related super-vertexes. From the results, we make some observation. Firstly, meaningful relationships are learned as shown in the tree structure. The super-edges between super-vertex "hair shape, hair length" and other 5 super-vertex are retained, while the super-vertex "eye corner" only remains one super-edge with other super-vertex. It means "hair shape, hair length" is more important and has broader relationship with other nodes than "eye corner" in this structure. Secondly, some highly correlated attributes are clustered into one super-vertex, such as "hair shape" with "hair length" and "eye shadow template" with "face shape". It well fit to the intuitive perception of humans. Long hair may match well with curled hair, and certain shape of eye shadow template may also fit to some face shapes. Thirdly, the correlation between super-vertexs is represented on the super-edges. The super-vertex "eye shadow template, eye shape" has the weight 0.5606 with 'eye shadow color, face shape" , which is higher than the weight 0.0501 with "eye corner". It means "eye shadow template, eye shape" has stronger correlation with "eye shadow color, face shape".

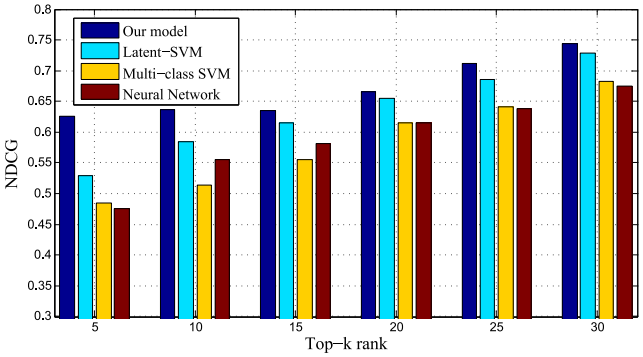

Figure 9: NDCG values of multiple tree-structured super-graphs model and three baselines. The horizontal axis is the rank of top-k results, while the vertical axis is the corresponding NDCG value. Our proposed method achieve better performance than the latent SVM model and other baselines.

5.2 Recommendation Results Evaluation

For the recommendation model in the Beauty e-Experts system, we also implement some alternatives using multi-class SVM, neutral network, and latent SVM. The first two baselines only use the low-level image features to train classifiers for high-level beauty attributes. The latent SVM baseline considers not only the low-level image features but also the pair-wise correlations between the beauty and the beauty-related attributes. We use the 100 testing images to evaluate the performance of the three baseline methods and our algorithm. To evaluate the recommendation result of the Beauty e-Experts system quantitatively, the human perception of suitable beauty makeups is considered as the ground truth measured on 50 random combinations of the attributes for all the 100 testing images. We asked 20 participants (staffs and students in our group) to label the ground truth of ranking results of the 50 types beauty makeup effects for each testing image. Instead of labeling absolute ranks from 1 to 50, we use a k-wise strategy similar to [23]: labelers are shown k images as a group each time, where k is set to 10. They only need to rank satisfying levels within each group. $C(k, 2)$ pairwise preferences can be obtained from the k ranks, and then the final rank is calculated across groups by ranking SVM [19].

In Figure 9, we plot the comparison results of our model and other baselines. The performance is measured by Normalized Discounted Cumulative Gain (NDCG) [13], which is widely used to evaluate ranking systems. From the results, we can observe that our model and latent SVM significantly

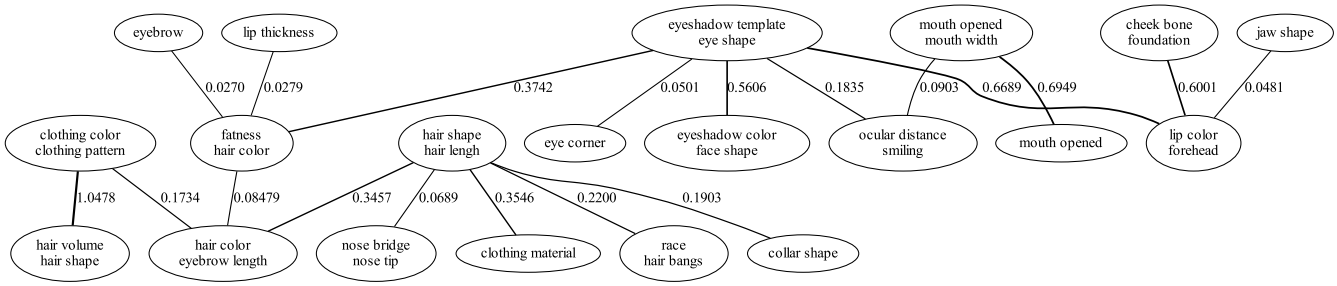

Figure 8: Visualization of one learned tree-structured super-graphs model.

outperform multi-class SVM and neural network. This is mainly because that our model and the latent SVM method are equipped with mid-level beauty-related attributes to narrow the semantic gap between low-level image features and the high-level beauty attributes. These two models are both able to characterize the co-occurrence information to mine the pairwise correlations between every two factors. From Figure 9 we can further find that our model has overall better performance than the latent SVM method, especially in the top 15 recommendations. With higher order relationships embedded, our model can express more complex relationship among different attributes. What is more, by employing multiple tree-structured super-graphs, our model tend to obtain more robust recommendation results.

5.3 Synthesis Results Evaluation

We compare our Beauty e-Experts system with some commercial virtual hairstyle and makeup systems, including Virtual Hairstyle (VH)[6], Instant Hair Makeover (IHM)[7], Daily Makeover (DM)[8], Virtual Makeup Tool (VMT)[9], and the virtual try-on website TAAZ[10]. They are all very popular in female customers on the Internet.

Table 3: Comparisons of several popular hairstyle and makeup synthesis systems.

	VH	IHM	DM	VMT	TAAZ	Ours
hairstyle	✓	✓	✓	×	✓	✓
makeup	×	✓	✓	✓	✓	✓
face detection	×	✓	✓	×	✓	✓
easy of use	×	×	✓	×	×	✓
500+ templates	×	×	×	×	✓	✓
composition freedom	×	✓	×	✓	✓	✓
recommendation	×	×	×	×	×	✓

We first compare these systems in an overview manner, which means that we focus on the comparison of the main functionalities among these systems. The comparison results are summarized in Table 3. It can be seen that IHM, DM and TAAZ systems can provide both hairstyle and makeup synthesis functions. They also provide face detection, which can largely reduce the manual workload. IHM, VMT and TAAZ ask users to choose makeup and hair products to perform composition, while VH and DM cannot support this, since their methods are mainly based on holistic transformation between the testing face and the example template. However, all these systems cannot support large data set with more than 500 templates and do not provide hairstyle and makeup recommendation functions. In the contrast, our Beauty e-Experts system can support all the functions mentioned above. What is more, it is fully automatic and can work in more general cases. The recommendation function of our system is really useful to help female users choose suitable hairstyle and makeup products.

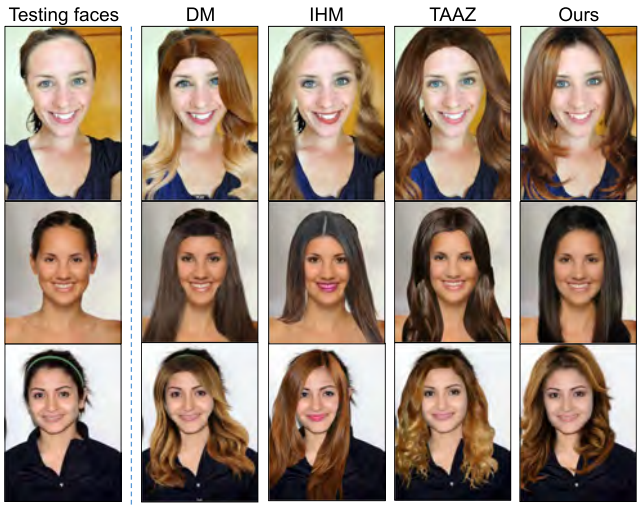

Figure 10: Contrast results of synthesized effect among websites and our paper.

We then compare the hairstyle and makeup synthesis results with these commercial systems. As shown in Figure 10, the first column is the testing images, and other four columns are the results generated by DM, IHM, TAAZ, and our system, respectively. The reason why we select these three systems is that only these three can synthesize both the hairstyle and makeup effects. The makeup and hairstyle templates used in the synthesis process are selected with some user interactions to ensure that all the four methods share similar makeups and hairstyles. It can be seen that, even after some extra user interactions, the results generated from these three websites have obvious artifacts. The selected hair templates cannot cover the original hair area. IHM even cannot handle the mouth opened cases. We further present several testing faces with different hairstyle and makeup effects in Figure 11. These results still look natural with a variety of style changing, which demonstrates the robustness of our system.

6. CONCLUSIONS AND FUTURE WORK

In this work, we have developed the Beauty e-Experts system for automatic facial hairstyle and makeup recommendation and synthesis. To the best of our knowledge, it is the first study to investigate into a fully automatic hairstyle

[6]http://www.hairstyles.knowage.info
[7]http://www.realbeauty.com/hair/virtual/hairstyles
[8]http://www.dailymakeover.com/games-apps/games
[9]http://www.hairstyles.knowage.info
[10]http://www.taaz.com

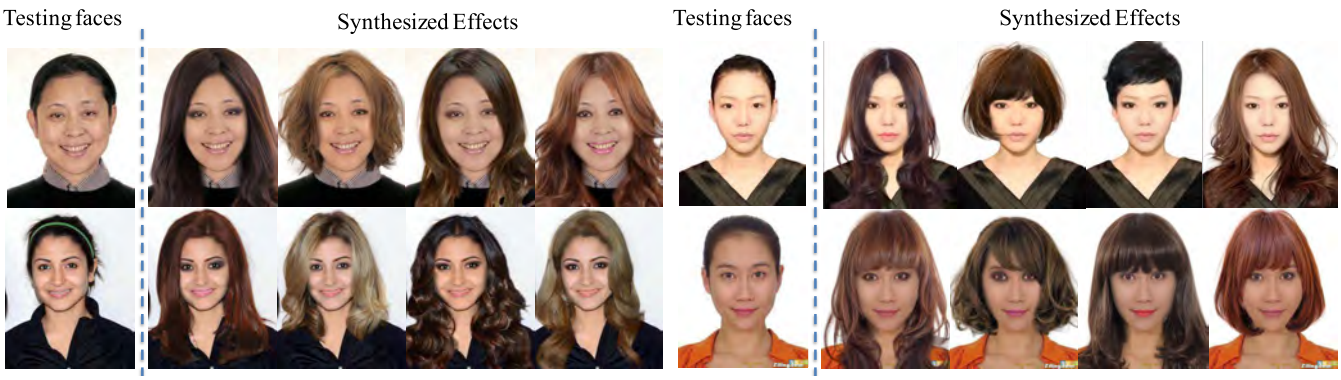

Figure 11: More synthesized results of the proposed Beauty e-Experts system.

and makeup system that simultaneously deals with hairstyle and makeup recommendation and synthesis. Based on the proposed multiple tree-structured super-graphs model, our system can capture the complex relationships among the different attributes, and produce reliable and explainable recommendations results. The synthesis model in our system also produces nature and appealing results. Extensive experiments on a newly built dataset have verified the effectiveness of our recommendation and synthesis models.

Current system is definitely not perfect yet. We are planning to further improve the system in the following directions. First of all, we may further consider to segment the hair region, and fill in the uncovered region with example-based inpainting techniques [9]. Secondly, we may extend current system for male users by constructing another male beauty dataset. At last, we also plan to extend current system to perform occasion-aware and personalized recommendation by introducing more occasion-related attributes [23] and employing information collected from personal photo album in a user's social network.

7. ACKNOWLEDGMENTS

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. Also it is partially supported by Singapore Ministry of Education under research Grant MOE2010-T2-1-087 and the "Extreme Facial Analysis" project from Ministry of Home Affairs, Singapore.

8. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: application to face recognition. *TPAMI*, 2006.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *TPAMI*, 2002.

[3] F. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *TPAMI*, 1989.

[4] Y. Boykov and O. Veksler. Fast approximate energy minimization via graph cuts. *TPAMI*, 2001.

[5] C. Chang and C. Lin. Libsvm: A library for support vector machines. In *TIST*, 2011.

[6] F. Chen and D. Zhang. A benchmark for geometric facial beauty study. In *Int. Conf. Medical Biometrics*, 2010.

[7] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *TIT*, 1968.

[8] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995.

[9] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *TIP*, 2004.

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

[11] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, third edition, 2008.

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

[13] R. Feris and L. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

[14] A. Goshtasby. Piecewise linear mapping functions for image registration. *PR*, 1986.

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

[16] S. Haykin. *Neural Networks*. Prentice Hall, 1999.

[17] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, 2010.

[18] G. Huang, M. Ramesh, T. Berg, and E. Learned. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, 2007.

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

[20] I. Jolliffe. Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*, 2002.

[21] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[22] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *TPAMI*, 2008.

[23] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. "Hi, magic closet, tell me what to wear". In *ACM MM*, 2012.

[24] T. Mensink, J. Verbeek, and G. Csurka. Tree-structured crf models for interactive image labeling. *TPAMI*, 2013.

[25] Y. Nagai, K. Ushiro, Y. Matsunami, T. Hashimoto, and Y. Kojima. Hairstyle suggesting system, hairstyle suggesting method, and computer program product. US Patent US20050251463 A1, 2005.

[26] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *TOG*, 2004.

[27] K. Scherbaum, T. Ritschel, M. Hullin, T. Thormählen, V. Blanz, and H. Seidel. Computer-suggested facial makeup. *CGF*, 2011.

[28] W. Tong, C. Tang, M. Brown, and Y. Xu. Example-based cosmetic transfer. In *FG*, 2007.

[29] N. Wang, H. Ai, and F. Tang. What are good parts for hair shape modeling? In *CVPR*, 2012.

[30] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

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